Extensive and deep sequencing of the Venter/HuRef genome for developing and benchmarking genome analysis tools

Bo Zhou[1,3], Joseph G. Arthur[2,3], Steve S. Ho[1], Reenal Pattni[1], Wing H. Wong[2], Alexander E. Urban[1]

[1]Department of Psychiatry and Behavioral Sciences, Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA
[2]Department of Statistics, Department of Biomedical Data Science, Bio-X Program, Stanford University, Stanford, California 94305, USA
[3]These authors contributed equally to this work.

corresponding author:

Alexander E. Urban (aeurban@stanford.edu)

ABSTRACT

We produced an extensive collection of deep re-sequencing datasets for the Venter/HuRef genome using the Illumina massively-parallel DNA sequencing platform. The original Venter genome sequence is a very-high quality phased assembly based on Sanger sequencing. Therefore, researchers developing novel computational tools for the analysis of human genome sequence variation for the dominant Illumina sequencing technology can test and hone their algorithms by making variant calls from these Venter/HuRef datasets and then immediately confirm the detected variants in the Sanger assembly, freeing them of the need for further experimental validation. This process also applies to implementing and benchmarking existing genome analysis pipelines. We prepared and sequenced 200 bp and 350 bp short-insert whole-genome sequencing libraries (sequenced to 100x and 40x genomic coverages respectively) as well as 2 kb, 5 kb, and 12 kb mate-pair libraries (43x, 97x, and 122x physical coverages respectively). Lastly, we produced a linked-read library (133x physical coverage) from which we also performed haplotype phasing.

BACKGROUND & SUMMARY

Almost two decades ago the extensive efforts of the Human Genome Project, backed up by work from Celera, resulted in the release of a draft of the first complete sequence of the human genome [1,2]. This catalyzed a new era of human whole-genome analysis where the now-available human genome sequence has been studied intensely to understand the functions of its parts and their interactions with each other and where a concurrent genome technology revolution has produced ever more powerful platforms to carry out such functional studies [3]. Since then, increasingly large numbers of human genomes have been sequenced, yielding insights into population-level genetic variation [4–6], structural genome variation [7–9], and mutational mechanisms [10]. Technological advances have progressively improved the information content and reduced the noise profile of sequencing data [11]. A large variety of methodologies for the routine analysis of sequencing data is now available [12]. “Whole-genome sequencing” is now a standing term that refers to the re-sequencing of a given sample of human genomic DNA using, typically, the dominant Illumina DNA sequencing platforms which can quickly produce several hundred million short sequencing reads at affordable costs. These reads are then aligned to the human reference genome and analyzed using various approaches [12–14], such as mismatch analysis, read-depth analysis, split-read analysis and discordant read-pairs analysis, producing an extensive catalog of sequence variants that are present in the DNA sample in question relative to the human reference sequence. The promise of human genome research is nothing short of a complete transformation of basic life science research, translational research, and eventually the way we diagnose, treat, and find cures for human disease.

It is clear, however, that current standard whole-genome sequence analysis leaves a rather large room for improvement. The standard genome analysis practices of today perform rather poorly in certain contexts, such as in repetitive regions (i.e. in around half the human genome), in the detection and resolution of complex structural variation, or in placing detected variants in their proper haplotypes. Although more advanced and novel computational algorithms that address these limitations are continuously being developed, one essential requirement during this process is that the detected variants are to be experimentally validated in order to establish false-positive rates and to make it possible to further tune and optimize the new algorithms. Experimental validation, especially of complex variants, during the tool development and testing phases is a very laborious and time-consuming process, but it can be circumvented by using a genome for which sufficiently large numbers of variants are already known, i.e. prevalidated. Several studies have been conducted with the goal of extensively characterizing the variants in a small number of human genomes using multiple sequencing technologies [15,16]. In some human genomes, variants have been carefully and extensively documented, providing a benchmark for other studies [9,17–20].

The Venter (HuRef) Genome, however, is especially distinguished for quality among the publicly-available human genome sequences as it is the only one for which its complete diploid assembly was generated from high-quality Sanger reads [17] and for which extensive catalogs of SNPs, indels, and structural variation are available [18,20]. To date, no extensive Illumina sequencing datasets have been available for the Venter/HuRef genome in contrast to other genomes that have been characterized for benchmarking purposes [15,16].

To unlock the potential of the Venter/HuRef genome as the outstanding benchmark genome, we have conducted deep whole-genome sequencing (WGS) using a variety of sequencing strategies for the Illumina platform (**Table 1**). Specifically, we produced short-insert paired-end WGS datasets at a combined sequence coverage of 140x, linked-read data at 42x de-duplicated sequencing coverage (133x physical coverage), and three long-insert (2 kb, 5 kb, and 12 kb) paired-end (i.e. mate-pair) WGS datasets with physical coverages of 43x, 97x, and 122x, respectively (**Figure 1**]. These datasets are of very high quality (**Figures 2-4**] and are complemented by the existing Venter/HuRef assembly-quality Sanger reads [17] and long-read sequencing data, which was produced using the Pacific Biosciences platform [21].

Researchers developing novel computational tools for analyzing whole-genome sequencing data can now test their algorithms by processing the appropriate Venter/HuRef Illumina datasets described here and then turn to the already-available catalogs of sequence variants, or to the original Sanger reads [17], to confirm the characterization of variants detected by their algorithms . Likewise, whenever a laboratory implements a new computational pipeline for human genome analysis, it can now use these Illumina Venter/HuRef datasets to confirm proper implementation and to optimize proper settings for the pipeline.

METHODS

Venter/HuRef DNA Sample

The Venter/HuRef DNA sample as obtained as a 50 µg aliquot of LCL-extracted DNA (NS12911) from the Coriell Institute for Medical Research where the iPSC (GM25430) of the same subject is also available (https://catalog.coriell.org/1/HuRef).

Illumina paired-end WGS

Library Preparation

The library preparation was previously described in detail in Mu *et al* [20]. Briefly, 1 µg of genomic DNA was fragmented using 2µL of NEBNext dsDNA fragmentase (New England Biolabs, Ipswich, MA) in 1x fragmentation buffer and 1x BSA. Reaction was kept on ice for 5 minutes before adding the fragmentase and was incubated at 37□°C for 20 minutes. The reaction was stopped by addition of 5 µL of 0.5 M EDTA. DNA was purified from the reaction mixture using 0.9x by volume AMPure XP beads (Beckman Coulter, Cat# A63880) and eluted in 50 µL of 10mM Tris-Acetate (pH 8.0) buffer. Six independent fragmentation reaction replicates were performed, and the sizes of the DNA were analyzed using Agilent 2100 Bioanalyzer before library preparation.

Library preparation was performed using the KAPA Library Preparation kit (KAPA Biosystems, Wilington, MA) where 200 ng of fragmented DNA was used as input. Library was constructed according to manufacture's protocol where the DNA was end-repaired and A-tailed before adapter ligation with Illumina TruSeq Adapter (Index 1). DNA was then purified using 0.8x by volume AMPure XP beads and quantified using the Qubit ds DNA High Sensitivity Assay Kit (Life Technologies, Cat# Q32851). For PCR amplification, 50 ng of DNA was amplified using the KAPA HiFi DNA Polymerase with the following thermocycling conditions: 98□°C/45□s, 5 cycles of (98□°C/15□s, 60□°C/30□s, 72□°C/45□s), 72□°C/1□min, and 4□°C /hold. Primers from the KAPA Library Preparation kit was used for PCR amplification. Afterwards, DNA was purified from the PCR reaction using AMPure XP beads and eluted in 30 µL of 10mM Tris-Acetate (pH 8.0) buffer. Six independent experimental replicates were performed, and the purified PCR amplified DNA fragments from each replicate was pooled for size selection and gel-purified from 2% agarose gel. Two size selections were made at 200 bp and 350 bp.

Sequencing

Sequencing of the 200 bp and 350 bp insert-size libraries was described previously in Mu et al [20]. The libraries were sequenced separately (2x100 bp) on an Illumina HiSeq 2000 instrument in rapid run mode. For the 200 bp insert-size library, a total of 3,214,626,588 reads generated from 5 sequencing runs was pooled together to obtain 100x genomic coverage. For the 350 bp insert-size library, a total of 1,280,576,580 reads generated from two sequencing runs was pooled together to obtain 40x genomic coverage.

Analysis

Reads were trimmed at the 3' end to a uniform length of 100 bp using FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/; version 0.0.13). The trimmed reads were aligned by BWA-MEM (Li and Durbin 2009; version 0.7.17-r1188) using the hg38 reference with ALT alleles removed (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/), and the resulting alignment records were sorted with Samtools (http://www.htslib.org/; version 1.7). Marking of PCR duplicates and calculations of insert-size and coverage information was performed using Picard (http://picard.sourceforge.net; version 2.17.10).

Illumina mate-pair WGS

Library Preparation

Mate Pair libraries at insert sizes 2 kb, 5 kb, and 12 kb were generated from Venter/HuRef DNA using the Nextera Mate Pair Sample Preparation Kit (Illumina, Cat# FC-132–1001) following standard manufacturer's instructions with the exception of the shearing step (see below). The Venter/HuRef DNA sample was first verified as high molecular weight (>15 kb) by running 60 ng, quantified by using the Qubit dsDNA HS Assay Kit (Life Technologies, Cat# Q32851), on 0.8% 1X TAE agarose gel next to the 1 kb Plus DNA Ladder (ThermoFisher Cat# 10787018). Afterwards, for each insert size, 4□µg of the high molecular weight genomic DNA was tagmented with biotinylated junction adapters and fragmented to about 7-8□kb on average in a 400 µL tagmentation reaction containing 12□µL of Tagmentase at 55□°C for 30□min. The tagmented DNA fragments were purified by adding 2X the volume of DNA Binding Buffer with Zymo Genomic DNA Clean & Concentrator Kit (Zymo Research, Cat# D4010) and eluted in 30 uL of Elution Buffer after two washes with the provided Wash Buffer. To fill in the gaps in the DNA adjacent to the junction adapters as a by-product of Tagmentation, single-strand displacement reaction was performed in a 200 µL reaction by adding 132 µL of water, 20 µL of 10x Strand Displacement Buffer, 8 µL of dNTPs, and 10 µL of Strand Displacement Polymerase to the 30 µL elution and at 20□°C for 30□min. DNA purification was then performed in 30 µL elution with 0.5x volume of AMPure XP Beads (Beckman Coulter, Cat# A63880) and size-selected by using BluePippin (Sage Science). The 0.75% DF 3-10kb Marker S1 – Improved Recovery and the 0.75% DF 10-18kb Marker U1 protocols were used for size selection on the BluePippin for insert sizes 5 kb and 12 kb respectively, and 0.75% DF 1-6kb Marker S1 protocol was used for insert size 2 kb. The “Tight Selection” option was used instead of “Range” for all size selections. The size selected DNA was then circularized overnight (12-16 hours) at 30□°C with Circularization Ligase in a 300□µL reaction.

After overnight circularization, the uncirculated linear DNA was digested by adding 9 µL of Exonuclease and incubated at 30□°C for 30 minutes and heat inactivated at 70□°C for 30 minutes. Afterwards, 12 µL of Stop Ligation Buffer was added. Circularized DNA was then transferred to T6 (6×32□mm) glass tube (Covaris, Part# 520031 and 520042) and sheared *twice* on the Covaris S2 machine (Intensity of 8, Duty Cycle of 20%, Cycles Per Burst of 200, Time of 40□s, Temperature of 2–6□°C). We find that shearing *twice* creates a tighter final library size distribution which leads to a higher fraction of pass-filter clusters during the Illumina sequencing step. The mate pair fragments within the sheared DNA fragments contain the biotinylated junction adapter and were selected by binding to Dynabeads M-280 Streptavidin Magnetic Beads (Invitrogen, Part# 112-05D) by adding an equal volume of the Bead Bind Buffer (incubated at 20□°C for 15 minutes on shaking heat block at highest rpm setting). The non-biotinylated molecules in solution were washed away using the Wash Buffer. All downstream reactions were carried out on streptavidin beads with magnetic immobilization and washes with the Wash Buffer between successive reactions (e.g. End Repair, A-Tailing, and Adapter Ligation. The sheared DNA was first End-repaired followed by A-Tailing and TruSeq indexed adapter ligation. The adapter-ligated DNA was resuspended in 20 µL of Resuspension Buffer and then PCR amplified in a 50 µL reaction with 25 µL of PCR 2X Master Mix and 5 µL of Primers both provided in the Nextera Mate Pair Sample Preparation Kit (Illumina, Cat# FC-132–1001) to generate the final library. The thermocycling conditions are 98□°C/1□min, 15 cycles of (98□°C/10□s, 60□°C/30□s, 72□°C/30□s), 72□°C/5□min, and 4□°C /hold. The amplified library (supernatant) was purified using a 0.66x volume of AMPure XP Beads (0.67x vol) and eluted in 20 µL of Resuspension Buffer. The size distribution of the library was determined by Agilent Technologies 2100 Bioanalyzer (High Sensitivity Assay), and the indexed library concentration was measured by the Qubit dsDNA HS Assay Kit (Life Technologies, Cat# Q32851).

Sequencing

The Mate-Pair libraries were sequenced on the Illumina NextSeq 500 using the NextSeq 500/550 Mid Output v2 kit (300 cycles) (Illumina, Cat# FC-404-2003) to generate 2×151□bp paired-end reads. The libraries were loaded onto the flowcell at a final concentration of 1.8pM and 1% PhiX Control v3 (Illumina, Cat# FC-110-3001). Additional rounds of sequencing also used a final library concentration of 1.8pM and 1% PhiX Control v3.

Analysis

Illumina Nextera Mate Pair junction adapter sequences were first trimmed using NxTrim (O'Connell et al. 2015; version 0.4.3) with the "--aggressive --preserve-mp" settings in order to maximize the number of long-insert pairs. Nxtrim outputs four sets of reads, designated "Mate Pair", "Paired-End", "Singleton", and "Unknown." "Mate Pair" reads have junction adapter sequence trimmed off from the 3' end of Read 1 and/or Read 2; "Paired-End" (short-insert) reads have junction adapter sequence trimmed from the 5' end of Read 1 and/or Read 2; "Singleton" reads have junction adapter sequence trimmed from the middle of either Read 1 or Read 2 rendering one of the reads useless. "Unknown" reads have no junction adapter sequences detected. This is most likely because the junction adapter sequence sits in the un-sequenced portion of the template, thus whether reads are "Mate Pair" or "Paired-End" cannot be discerned. Nonetheless, mate-pair reads are present in the "Unknown" fractions as well as paired-end reads. The "Unknown" reads can be used for alignment and analysis if more long-insert information is desired??? (O'Connell et al. 2015). Here, the reads designated as "Mate Pair" and "Unknown" were combined, aligned with BWA-MEM (Li and Durbin 2009) against the hg38 reference without ALT alleles (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/), and sorted using samtools (http://www.htslib.org/; version 1.7). Marking of PCR duplicates and calculations of insert-size and coverage information was performed using Picard (http://picard.sourceforge.net; version 2.17.10).

10X Genomics Chromium library for Illumina sequencing

Input genomic DNA preparation

The Venter/HuRef DNA sample (obtained from the Coriell Institute for Medical Research) was first verified as high molecular weight (>15 kb) by running 60 ng, quantified by using the Qubit dsDNA HS Assay Kit (Life Technologies, Cat# Q32851), on 0.8% 1X TAE agarose gel next to the 1 kb Plus DNA Ladder (ThermoFisher Cat#

10787018). Afterwards, 4□µg of the high molecular weight genomic DNA was loaded on a BluePippin (Sage Science) instrument to select for DNA fragments 30 kb to 80 kb using the “0.75%DF Marker U1 high-pass 30- 40 kb vs3” protocol. The concentration of the selected DNA fragments was then quantified by using the Qubit dsDNA HS Assay Kit (Life Technologies, Cat# Q32851) and diluted to 1 ng/ µL. The final dilution concentration of 1 ng/ µL was verified again by performing three technical replicates of Qubit dsDNA HS Assay with 5 µL of the DNA dilution as input.

Chromium whole-genome linked-read library preparation and sequencing

The linked-read whole-genome library was prepared using the Chromium Genome kit and reagent delivery system (10X Genomics, Pleasanton, CA). The linked-read library was made following standard manufacturer’s protocol with 10 cycles of PCR amplification. Briefly 1□ng of DNA (~300 genome equivalents) of size-selected high molecular DNA was partitioned into ~1.5 million oil droplets in emulsion, tagged with a unique 16 bp barcode within each droplet, and subjected isothermal amplification (30□°C for 3 hours; 65□°C for 30□minutes) by random priming within each droplet. Amplified (isothermal) DNA was then purified from the droplet emulsion following the manufacturer’s protocol using SPRI beads. The purified DNA was then End-Repaired and A-tailed followed by adapter ligation of adapter in the same reaction mixture. DNA was purified from the was the reaction mixture using SPRI beads and eluted in 40 uL. Sample Index PCR amplification (primers and 2x master mix provided in the Chromium Genome kit) was then performed on the eluted DNA in a toal volume of 100 uL with the following thermocycling conditions: 98□°C/45□s, 10 cycles of (98□°C/20□s, 54□°C/30□s, 72□°C/20□s), 72□°C/1□min, and 4□°C /hold. Primer index SI-GA-A6 was used. DNA (final linked-read library) was purified from the PCR reaction with SPRI bead size selection following manufacturer’s protocol. The final purified library was quantified by qPCR (KAPA Library Quantification Kit for Illumina platforms, Kapa Biosystems, Wilmington, MA) using the following thermocycling conditions: 95□°C/3 min, 30 cycles of (95□°C/5□s, 67□°C/30□s). The library concentration was calculated in nanomolar (nM) concentration and then diluted to 5 nM. Sequencing (2x151bp, 8 cycles of single indexing) on two lanes of Illumina HiSeq X was performed at Macrogen (Rockville, MD).

Sequencing

The final purified library was quantified by qPCR (KAPA Library Quantification Kit for Illumina platforms, Kapa Biosystems, Wilmington, MA) using the following thermocycling conditions: 95□°C/3 min, 30 cycles of (95□°C/5□s, 67□°C/30□s). The library concentration was calculated in nanomolar (nM) concentration and then diluted to 5 nM. Sequencing (2x151bp, 8 cycles of single indexing) on two lanes of Illumina HiSeq X (flowcell ID: H3MHGALXX, lanes #4 and #5) was performed at Macrogen (Rockville, MD) resulting in a total of 789,239,544 paired reads (**Table 1**).

Analysis

FASTQ files were generated raw BCL files using “*mkfastq*” mode in the Long Ranger software (version 2.1.3) from 10X Genomics (Pleasanton, CA). 10X Genomics

Chromium library index “SI-GA-A6” was specified in the required sample sheet file for “*mkfastq*”. Before alignment, the hg38 genome files were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz and indexed using the “*mkref*” mode in Long Ranger. Sequencing alignment and haplotype phasing were performed using the “*wgs*” mode in Long Ranger, and the options *“--sex=male*” and *“--vcmode=freebayes*” were specified. Only “PASS” SNPs and Indels 50 bp or smaller were included in the final phased variant vcf.

DATA RECORDS

The Venter/HuRef genome sequenced is publicly available through The Coriell Institute for Medical Research (Camden, NJ, USA) both as genomic DNA (catalog ID: NS12911) extracted from lymphoblastoid cell line (LCL) or as retroviral reprogrammed induced pluriplotent stem cell culture (catalog ID: GM25430). As described in the Methods, Venter/HuRef LCL DNA (NCBI SRA biosample accession SAMN03491120) was used for sequencing library preparation in this work.

Illumina short-insert WGS

Approximately 100x sequencing coverage 2x100bp Illumina short-insert (200 bp) WGS data generated from the Illumina HiSeq 2000 is available through NCBI SRA accession SRR7097858 [Data Citation 1: NCBI SRA SRR7097858]. Approximately 40x sequencing coverage 2x100bp Illumina short-insert (350 bp) WGS data generated from the Illumina HiSeq 2000 platform is available through NCBI SRA accession SRR7097859 [Data Citation 2: NCBI SRA SRR7097859].

Illumina mate-pair WGS

Illumina mate-pair data sequenced (2x150 bp) on the Illumina NextSeq 500 are available through NCBI SRA accessions SRR6951312 [Data Citation 3: NCBI SRA SRR6951312], SRR6951313 [Data Citation 4: NCBI SRA SRR6951313], and SRR6951310 [Data Citation 5: NCBI SRA SRR6951310] for insert sizes 2 kb, 5 kb, and 12 kb respectively.

10X Genomics Chromium linked-read Library

10X Genomics Chromium linked-read data sequenced (2x150 bp) on two lanes of the Illumina HiSeq X Ten is available through NCBI SRA accession SRR6951311 [Data Citation 6: NCBI SRA SRR6951311]. The phased variants of the Venter/HuRef genome obtained through the analysis linked reads is available through dbSNP NCBI_ss# 2137543904 to 3651364986 (For phasing information, request for original submitted vcf file through NCBI dbSNP.) [Data Citation 7: NCBI dbSNP NCBI_ss# 2137543904-3651364986].

TECHNICAL VALIDATION

Illumina short-insert WGS

Sequencing quality of the WGS mate-pair libraries were assessed using FastQC (Supplementary Information). Insert-size, coverage, GC-bias, alignment, and duplication metrics were analyzed using Picard tools (http://broadinstitute.github.io/picard/). These statistics are summarized in Table 1, Table 2 and Figure 2A.

Illumina mate-pair WGS

Sequencing quality of the WGS mate-pair libraries was assessed using FastQC (Supplementary Information). Insert-size, coverage, GC-bias, alignment, and duplication metrics were analyzed using Picard tools (http://broadinstitute.github.io/picard/). These statistics are summarized in **Table 1, Table 2 and Figure 2C-J**. Read fractions that were designated by NxTrim [23] as “Mate Pair”, “Paired-End”, “Singletons”, and “Unknown” are summarized in **Table 3**. The “Mate Pair” fraction for all libraries fall within the expected range (~40-60%). The relatively high rates of PCR duplication (expected for mate-pair libraries) result in significant decreases in sequence coverage (3x to 7x) (**Table 1, Table 2, Figure 2, Supplementary Information**). However, the more useful metric for mate-pair sequencing is high physical coverage [15]. The mean insert sizes for the mate pair libraries are 1.8 kb, 4.8 kb, and 12.2 kb (**Table 2, Figure 2**), which results in physical genomic coverage values of 62x, 136x, and 162x respectively.

10X Genomics Chromium Library

Sequencing quality of the WGS mate-pair libraries were assessed using FastQC (Supplementary Information). Input molecule length, coverage, alignment, duplication, droplet barcode, and phasing metrics were analyzed using Long Ranger software version 2.1.5 from 10X Genomics (Pleasanton, CA, USA). These statistics are summarized in **Table 1, Table 2, Table 4 and Figure 3**. Overall, 2.4 million and 1.5 million, 0.42 million and 0.29 million heterozygous and homozygous SNVs and indels respectively were called (**Table 4**). Of which, 96.7% and 93.85% of heterozygous SNVs and Indels respectively were successfully phased in the Venter/HuRef Genome in a total of 8882 haplotype blocks (N50 ~ 0.9 Mbp, longest phase block ~ 6.5 Mbp) (**Table 4**). Phase blocks for each chromosome are shown in **Figure 3**. Similar to mate-pair libraries, the physical coverage of the linked read library is calculated to be 133x from the mean input DNA molecule length of 32kb.

USAGE NOTES (optional)

The Venter/HuRef genome sequenced in this work is publicly available as both cell line and DNA from Coriell Institute for Medical Research. The mate-pair and linked-read sequencing data used the same DNA sample/extraction as input. It is possible that small differences may exist when compared to the short-insert datasets since the input DNA came from different cell passages and extractions. Researchers are especially encouraged to use the sequencing data in this work in combination with diploid Sanger sequencing data available for the Venter/HuRef genome published in Levy et al.

ACKNOWLEDGEMENTS

A.E.U. was supported by NIH grant P50-HG007735 and the Stanford Medicine Faculty Innovation Program, and B.Z. was additionally supported by NIH training grant T32-HL110952-04. W.H.W. was supported by NIH grants P50-HG007735 and R01-HG007834. J.G.A. received funding from NIH training grant T32-GM096982 and NSF Graduate Fellowship DGE-114747.

AUTHOR CONTRIBUTIONS

BZ and RP performed experiments. JGA, BZ, SSH performed data analysis. BZ, JGA, and AEU wrote the manuscript.

COMPETING INTERESTS

The authors declare no conflict of interest.

REFERENCES

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science (80-.).* **291,** 1304–51 (2001).
3. Reuter, J. A., Spacek, D. V. & Snyder, M. P. High-Throughput Sequencing Technologies. *Mol. Cell* **58,** 586–597 (2015).
4. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).
5. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
6. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
7. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–6 (2007).
8. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).
9. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2017). doi:10.1101/193144
10. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in

1,548 trios from Iceland. *Nature* **549,** 519–522 (2017).

11. Kumar, V. *et al.* Uniform, optimal signal processing of mapped deep-sequencing data. *Nat. Biotechnol.* **31,** 615–22 (2013).
12. Pabinger, S. *et al.* A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **15,** 256–278 (2014).
13. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12,** 363–376 (2011).
14. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43,** 491–8 (2011).
15. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. data* **3,** 160025 (2016).
16. Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27,** 157–164 (2017).
17. Levy, S. *et al.* The Diploid Genome Sequence of an Individual Human. *PLoS Biol.* **5,** e254 (2007).
18. Pang, A. W. *et al.* Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* **11,** R52 (2010).
19. Parikh, H. *et al.* svclassify: a method to establish benchmark structural variant calls. *BMC Genomics* **17,** 64 (2016).
20. Mu, J. C. *et al.* Leveraging long read sequencing from a single individual to provide a comprehensive resource for benchmarking variant calling methods. *Sci. Rep.* **5,** 14493 (2015).
21. Lin, M. Comparing de novo assemblies of J. Craig Venter's genome. (2015). doi:10.6084/m9.figshare.1319564.v1
22. Arthur, J. G., Chen, X., Zhou, B. & Urban, A. E. Detection of complex structural variation from paired-end sequencing data. *bioRxiv* 1–32 (2017). doi:10.1101/200170
23. O'Connell, J. *et al.* NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31,** 2035–2037 (2015).

DATA CITATIONS

1. **Arthur, J. G. NCBI SRA SRR7097858 (2015)**
2. **Arthur, J. G. NCBI SRA SRR7097859 (2015)**
3. **Zhou, B. NCBI SRA SRR6951312 (2018)**
4. **Zhou, B. NCBI SRA SRR6951313 (2018)**
5. **Zhou, B. NCBI SRA SRR6951310 (2018)**
6. **Zhou, B. NCBI SRA SRR6951311 (2018)**
7. **Zhou, B. NCBI dbSNP NCBI_ss# 2137543904-3651364986 (2018)***

*For phasing information, request the originally submitted VCF file through NCBI dbSNP.

FIGURE & TABLE LEGENDS

Figure 1. (A) Schematic diagram of the study. Venter/HuRef genomic DNA was used to generate short-insert (200 bp and 350 bp), mate-pair (2 kb, 5 kb, and 12 kb), and linked-read libraries. (**B**) Detailed overview of data generation including bio-sample used, types of Illumina WGS libraries constructed, sequencing instrument platforms, types of sequencing runs, and subsequent analysis of data.

Figure 2. Normalized coverage, GC (%) content windows, base quality at GC (%), and corresponding insert-size histograms for all WGS libraries: 200 bp short-insert (**A,B**), 350 bp short-insert (**C,D**), 2kb-mate-pair (**E,F**), 5kb-mate-pair (**G,H**), 12kb-mate-pair (**I,J**).

Figure 3. Coverage (deduplicated) histograms of (**A,B**) short-insert, (**C,D,E**) 2 kb, 5 kb, and 12 kb mate-pair, and (**F**) linked-read libraries. Only reads with mapping score > 20 were used.

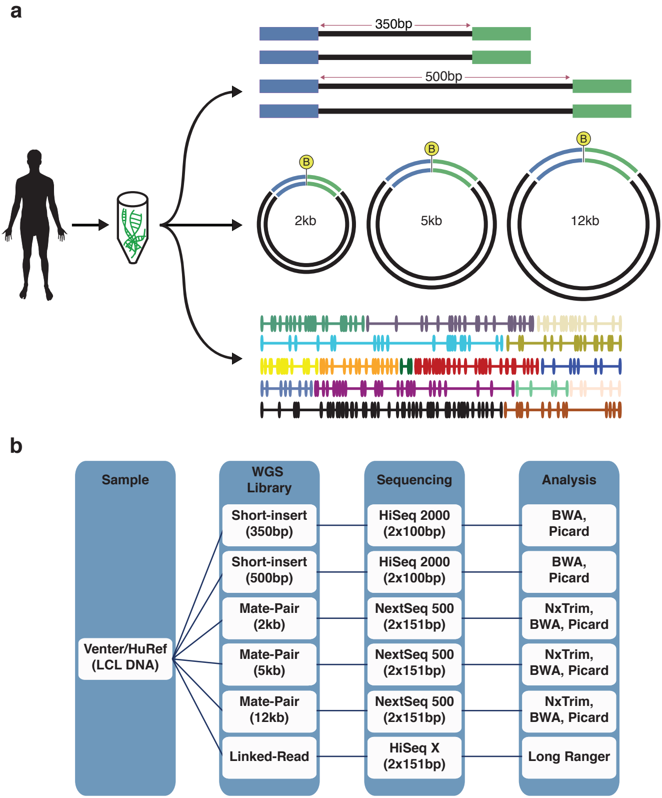
Figure 4. Violin plot of sizes of haplotype blocks constructed using linked-read sequencing (133x physical coverage) for HuRef/Venter Genome for all chromosomes.

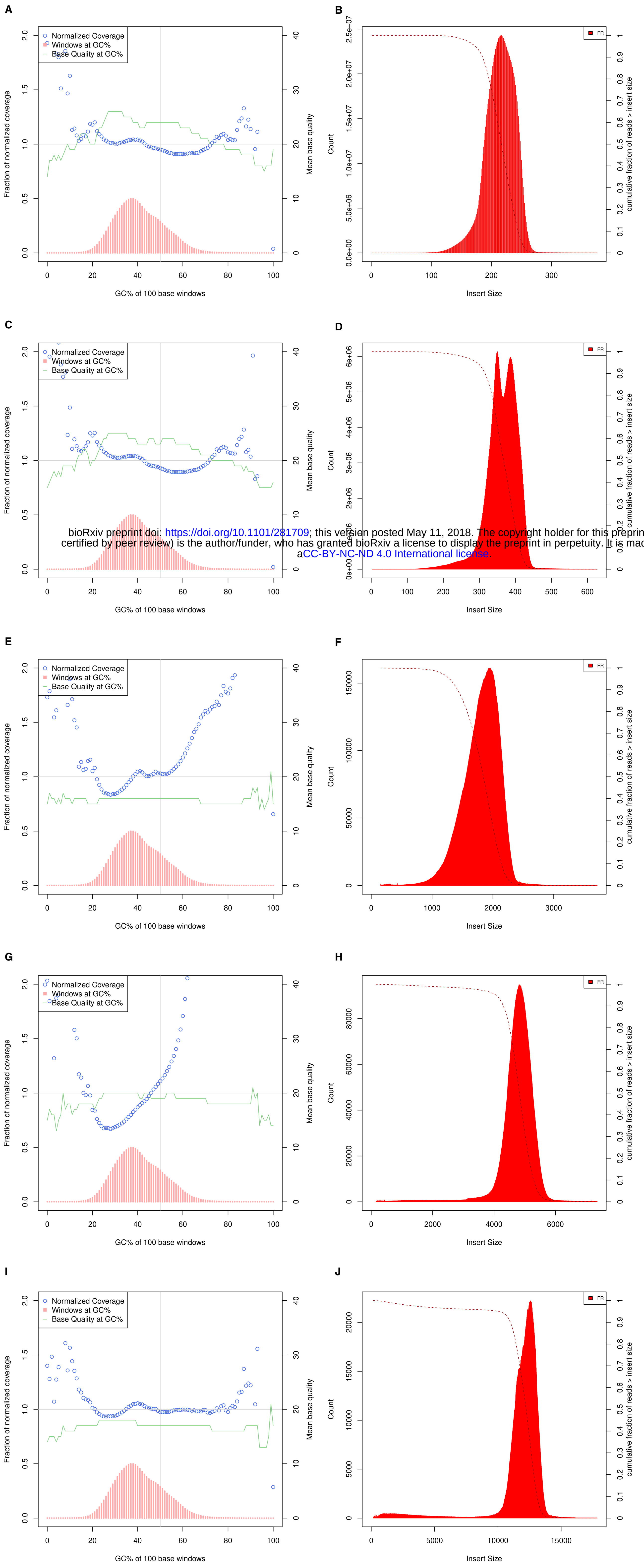
Table 1. Summary of library construction and sequencing for short-insert, mate-pair, and linked-read HuRef/Venter WGS libraries.

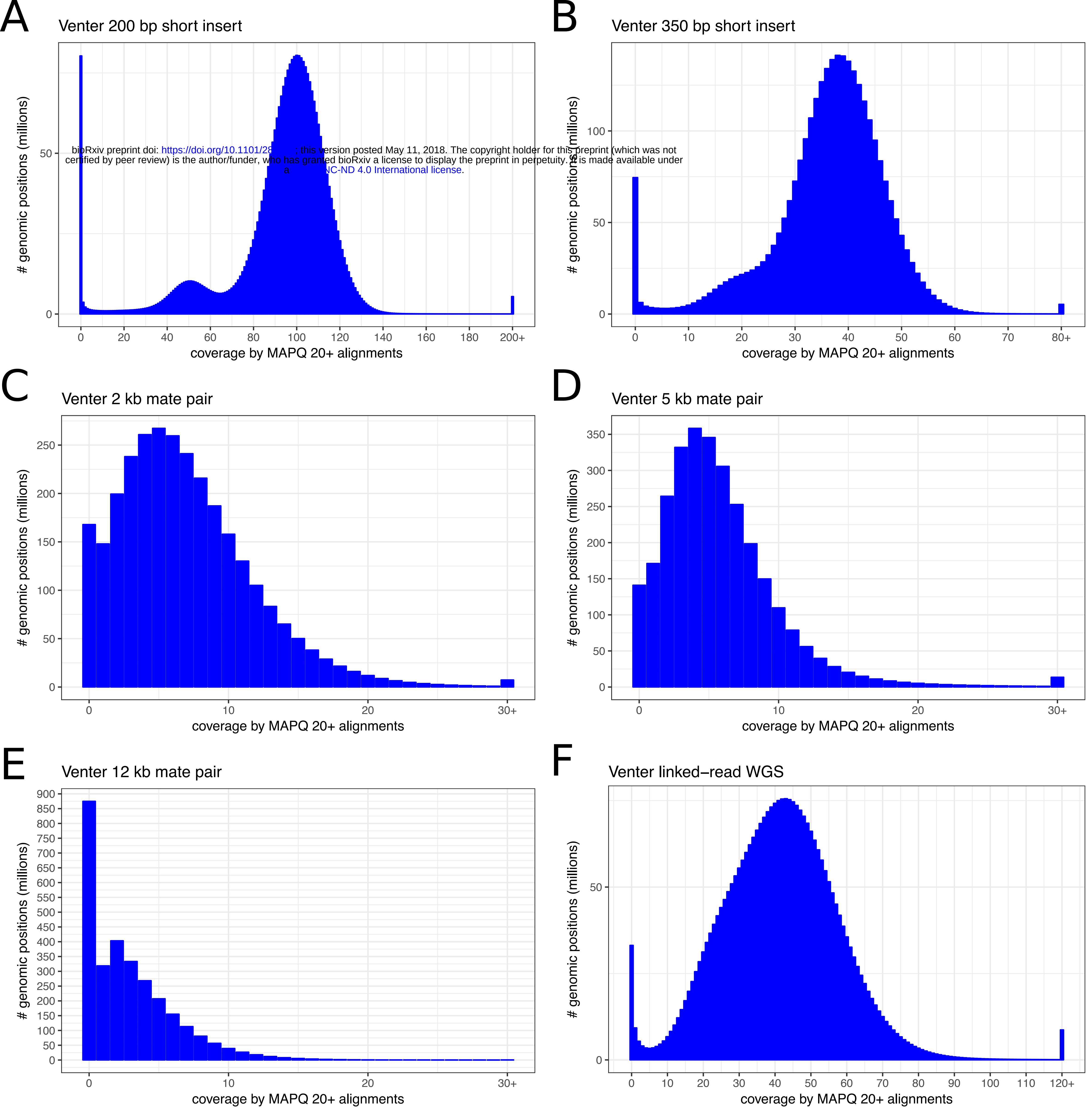
Table 2. Summary of post sequencing QC, alignment, duplication, coverage and insert-size analysis for all libraries.

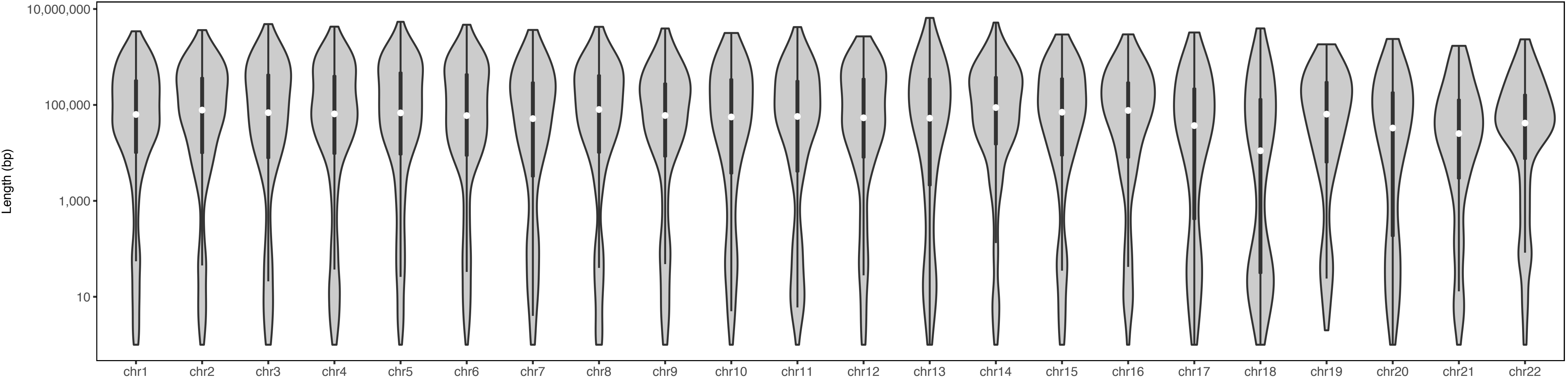
Table 3. Statistics for trimming of Nexera junction adapter sequence using NxTrim [23] for all mate-pair libraries.

Table 4. Summary of metrics for linked-read sequencing and phasing of the HuRef/Venter genome.


a
350bp
500bp
B
2kb
B
5kb
B
12kb
b
Sample
Venter/HuRef
(LCL DNA)
WGS
Library
Short-insert
(350bp)
Short-insert
(500bp)
Mate-Pair
(2kb)
Mate-Pair
(5kb)
Mate-Pair
(12kb)
Linked-Read
Sequencing
HiSeq 2000
(2x100bp)
HiSeq 2000
(2x100bp)
NextSeq 500
(2x151bp)
NextSeq 500
(2x151bp)
NextSeq 500
(2x151bp)
HiSeq X
(2x151bp)
Analysis
BWA,
Picard
BWA,
Picard
NxTrim,
BWA, Picard
NxTrim,
BWA, Picard
NxTrim,
BWA, Picard
Long Ranger


A
B
C
D
E
F
G
H
I
J
Normalized Coverage
Windows at GC%
Base Quality at GC%
Fraction of normalized coverage
Mean base quality
GC% of 100 base windows
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
0
20
40
60
80
100
FR
Count
Insert Size
cumulative fraction of reads > insert size
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
0
100
200
300
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
0
100
200
300
400
500
600
0
50000
100000
150000
0
1000
2000
3000
0
20000
40000
60000
80000
0
2000
4000
6000
0
5000
10000
15000
20000
0
5000
10000
15000


A
Venter 200 bp short insert
# genomic positions (millions)
50
0
0
20
40
60
80
100
120
140
160
180
200+
coverage by MAPQ 20+ alignments
B
Venter 350 bp short insert
# genomic positions (millions)
100
50
0
0
10
20
30
40
50
60
70
80+
coverage by MAPQ 20+ alignments
C
Venter 2 kb mate pair
# genomic positions (millions)
250
200
150
100
50
0
0
10
20
30+
coverage by MAPQ 20+ alignments
D
Venter 5 kb mate pair
# genomic positions (millions)
350
300
250
200
150
100
50
0
0
10
20
30+
coverage by MAPQ 20+ alignments
E
Venter 12 kb mate pair
# genomic positions (millions)
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50
0
0
10
20
30+
coverage by MAPQ 20+ alignments
F
Venter linked−read WGS
# genomic positions (millions)
50
0
0
10
20
30
40
50
60
70
80
90
100
110
120+
coverage by MAPQ 20+ alignments


Length (bp)
10,000,000
100,000
1,000
10
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22