Title

Highly-accurate long-read sequencing improves variant detection and assembly of a human genome

Authors

Aaron M. Wenger[1*], Paul Peluso[1*], William J. Rowell[1], Pi-Chuan Chang[2], Richard J. Hall[1], Gregory T. Concepcion[1], Jana Ebler[3,4,5], Arkarachai Fungtammasan[6], Alexey Kolesnikov[2], Nathan D. Olson[7], Armin Töpfer[1], Chen-Shan Chin[6], Michael Alonge[8], Medhat Mahmoud[9], Yufeng Qian[1], Adam M. Phillippy[10], Michael C. Schatz[8], Gene Myers[11], Mark A. DePristo[2], Jue Ruan[12], Tobias Marschall[3,4], Fritz J. Sedlazeck[9], Justin M. Zook[7], Heng Li[13], Sergey Koren[10], Andrew Carroll[2], David R. Rank[1]†, Michael W. Hunkapiller[1]†

* These authors contributed equally to this work.
† Address correspondence to M.W.H. (mhunkapiller@pacb.com) or D.R.R. (drank@pacb.com).

1. Pacific Biosciences, Menlo Park, CA, USA
2. Google Inc., Mountain View, CA, USA
3. Center for Bioinformatics, Saarland University, Saarbrücken, Germany
4. Max Planck Institute for Informatics, Saarland Informatics Campus E1.4, Saarbrücken, Germany
5. Graduate School of Computer Science, Saarland University, Saarland Informatics Campus E1.3, Saarbrücken, Germany
6. DNAnexus, Mountain View, CA, USA
7. National Institute of Standards and Technology, Gaithersburg, MD, USA
8. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
9. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
10. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA
11. Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
12. Agricultural Genomics Institute, Chinese Academy of Agriculture Sciences, Shenzen, China
13. Dana-Farber Cancer Institute, Boston, MA, USA

Abstract

The major DNA sequencing technologies in use today produce either highly-accurate short reads or noisy long reads. We developed a protocol based on single-molecule, circular consensus sequencing (CCS) to generate highly-accurate (99.8%) long reads averaging 13.5 kb and applied it to sequence the well-characterized human HG002/NA24385. We optimized existing tools to comprehensively detect variants, achieving precision and recall above 99.91% for SNVs, 95.98% for indels, and 95.99% for structural variants. We estimate that 2,434 discordances are correctable mistakes in the high-quality Genome in a Bottle benchmark. Nearly all (99.64%) variants are phased into haplotypes, which further improves variant detection. *De novo* assembly produces a highly contiguous and accurate genome with contig N50 above 15 Mb and concordance of 99.998%. CCS reads match short reads for small variant detection, while enabling structural variant detection and *de novo* assembly at similar contiguity and markedly higher concordance than noisy long reads.

Introduction

DNA sequencing technologies have improved at rates eclipsing Moore's law[1] revolutionizing biological sciences. Beginning in the 1970s, Sanger sequencing[2], and subsequent automation[3] facilitated large scale DNA sequencing projects and paved the way for modern genomic research[4–7]. The first reference genomes were followed by the advent of several high-throughput sequencing technologies (next-generation sequencing or NGS) including 454™, Solexa/Illumina®, ABI® Solid™, Complete Genomics™, and Ion Torrent™. These methods employed a range of chemistries and detection strategies[8–13]. All produce relatively accurate reads but are limited in read length, typically to less than 300 basepairs. These accurate short reads are well-suited for calling single-nucleotide variants (SNVs) and small insertions and deletions (indels), but are lacking for long-range applications such as *de novo* assembly, haplotype phasing, and structural variant detection.

For these applications, vastly superior results[14–16] are obtained with technologies like PacBio® SMRT Sequencing[17] and Oxford Nanopore sequencing[18] that produce long reads (>10 kb). These technologies rely on single-molecule detection and are characterized by reduced read accuracy (75-90%)[17,18]. High consensus accuracy has been demonstrated through read-to-read error correction, but the process is computationally intensive, and errors remain from mis-mapping reads and mixing haplotypes during correction[15,19]. As a result of the error rate, long-read technologies are rarely used to detect SNVs and indels.

Today, human genomes are sequenced at population scales, but it remains necessary to combine sequencing technologies to cover all types of genetic variation, which increases cost and adds complexity to projects. A sequencing technology with long

read length and high accuracy would enable a single experiment for comprehensive variant discovery.

Recent gains in read length for SMRT Sequencing and optimized DNA template preparation suggested an opportunity to unify high accuracy with long read lengths using circular consensus sequencing (CCS)[20,21]. CCS derives a consensus sequence from multiple passes of a single template molecule, producing accurate reads from noisy individual subreads.

Here, we highlight the performance of these highly-accurate, long CCS™ reads by sequencing and analyzing the well-characterized human male HG002/NA24385[22,23]. The HG002 sample is one of the benchmark samples available from the Genome in a Bottle (GIAB) Consortium. GIAB provides physical reference materials along with detailed characterization of the sample genome, defining "high-confidence regions" at which the sequence of the sample is known and "high-confidence variants" within those regions at which the sample differs from the human reference genome. Thus, it is an ideal sample for study of sequencing accuracy and variant detection. We apply and extend standard analysis tools to identify variation in HG002, demonstrating performance that rivals or surpasses existing technologies for short and large variation detection as well as genome assembly and haplotype phasing.

Results

CCS Library Preparation and Sequencing

A SMRTbell library tightly distributed at 15 kb was chosen for circular consensus sequencing (**Figure 1a**, **Supplementary Figure 1**) based on estimates of 150 kb polymerase read length and a requirement of 10 passes to achieve Q30 read accuracy (**Figure 1b**). CCS reads with a predicted accuracy of at least Q20 (99%) were retained (**Supplementary Figure 2**). The total CCS read yield was 89 Gb, an average of 2.3 Gb per SMRT Cell, with an average read length of 13.5 kb ± 1.2 kb (**Figure 1c**). The predicted accuracy of the CCS reads has a median of Q30 (99.9%) and a mean of Q27 (99.8%) (**Figure 1c**). Predicted accuracy matches well with concordance to the GRCh37 human reference genome (average [$Q_{predicted} - Q_{concordance}$] = -0.9), which indicates that the predicted accuracy is well calibrated (**Supplementary Figure 3**). Average mapped coverage of the genome is 28-fold, with minimal difference across [GC] content (**Supplementary Figure 4**).

Quality Evaluation of CCS Reads

To characterize the few residual errors in CCS reads, discordances between the reads and the reference genome at high-confidence, non-variant positions in HG002 were tallied. The large majority of discordances are indels in homopolymer contexts: 3.4% are mismatches, 4.6% are indels in non-homopolymer contexts, and 92.0% are indels in homopolymers. This equates to a mismatch every 13,048 bp in CCS reads, a non-

homopolymer indel every 9,669 bp, and a homopolymer indel every 477 bp (**Supplementary Table 1**).

To confirm independently the high quality of CCS reads, error rates were measured through read-to-read alignments[24]. Consistent with the reference-based methods, the average read accuracy is estimated at 99.8%. A putative large artifact is detected in 0.6% of reads: 0.5% are molecular chimeras, likely due to ligation of DNA fragments during library construction, 0.1% contain a "low quality" run of bases, anecdotally in microsatellites, and 0.03% have a missing SMRTbell adapter on one end. Overall, the read-to-read comparison supports the predicted quality of the CCS reads.

Increased Mappability of CCS Reads

To evaluate increases in mappability with long reads, the 13.5 kb CCS reads and a coverage-matched number of 2×250 bp NGS short reads were mapped to GRCh37. A genomic position was considered to be mappable if it is covered by least ten reads. At the highest reported mapping quality (60), 97.5% of the non-gap GRCh37 is mappable with 13.5 kb CCS reads, while 94.8% is mappable with NGS short reads (**Figure 2a**).

The additional regions that are now accessible with longer CCS reads include numerous medically-relevant genes which have been previously reported as recalcitrant to NGS sequencing[25]. Of the 193 reported medically-relevant genes with at least one NGS problem exon, 152 are fully mappable with the CCS data, including *CYP2D6*, *GBA*, *PMS2*, and *STRC* (**Figure 2b,c**).

The 13.5 kb CCS reads also resolve complex regions, like the HLA class 1 and 2 genes, which are fully phased and typed to four-field resolution[26] (**Supplementary Figure 5**).

Small Variant Detection with CCS Reads

GATK[27] was used to call SNVs and small indels with CCS reads. Evaluated against the GIAB benchmark[23], precision for SNVs is 99.468% and recall is 99.559%. For indels, precision is 78.977% and recall is 81.248%. While GATK performance with CCS reads is comparable to NGS for SNVs, it is lower for indels (**Table 1**). Unlike NGS read errors, which are mostly mismatches, CCS read errors are mostly indels (**Supplementary Table 1**), contributing to the low indel precision and recall of GATK for CCS reads.

Variant callers based on deep learning have an inherent ability to adapt to the error profiles of new data types. To evaluate variant calling with a deep learning framework, Google DeepVariant[28] was used to call SNVs and indels from CCS reads. Using a model trained on Illumina reads, precision is 99.533% and recall is 99.793% for SNVs, and precision is 23.991% and recall is 81.692% for indels (**Supplementary Table 2**). Training a model on CCS reads provides a large boost in precision and recall for both SNVs and indels. For SNVs, DeepVariant achieves precision of 99.914% and recall of

99.959%. For indels, DeepVariant achieves 96.901% precision and 95.980% recall (**Figure 3a**, **Table 1**). The plurality of discordant indels occur in homopolymer runs, matching the most common discordance in CCS reads (**Supplementary Figure 6**).

Phasing Small Variants with CCS Reads

To determine whether CCS reads could provide both highly-accurate variant calls and long-range information needed to generate haplotypes, we used WhatsHap[29] to phase the DeepVariant variant calls. Nearly all (99.64%) autosomal heterozygous variants were phased into 19,215 blocks with an N50 of 206 kb (**Supplementary Table 3**). The phase block size distribution closely matches the theoretical limit estimated by creating breaks between variants that are separated by more than the average CCS read length of 13.5 kb. This suggests that the phase block size is limited by read length and the amount of variation in HG002, not by coverage or the quality of the variant calls (**Figure 3b**). Evaluated against the GIAB benchmark phase set, the switch error rate is 0.37% and the Hamming error rate is 1.91% (**Supplementary Table 3**).

Improving Small Variant Detection with Haplotype Phasing

GATK and DeepVariant do not directly incorporate long-range haplotype phase information when calling variants. To evaluate whether phase information improves results, CCS reads were haplotype-tagged based on trio-phased variants from GIAB and a DeepVariant model was then trained on reads passed in haplotype-sorted order. The haplotype-sorted model performs similarly to the original DeepVariant CCS model for SNVs but provides a large improvement for indels achieving precision of 97.835% and recall of 97.141% (**Table 1**).

Structural Variant Detection with CCS Reads

Insertion and deletion structural variants ≥50 bp were called using two read mapping-based tools, pbsv (https://github.com/PacificBiosciences/pbsv) and Sniffles[30]. The callsets show similar precision (>94%) and recall (>91%) against the GIAB benchmark (**Supplementary Table 4**). Precision is consistent across variant size, but recall is lower for variants ≥3kb (**Supplementary Figure 7**). To increase recall for larger variants, haplotype-resolved FALCON and Canu *de novo* assemblies were analyzed with paftools[31] (see "*De Novo* Assembly of CCS Reads"), with precision >93% and recall >89% (**Supplementary Table 4**).

An integrated callset produced with SURVIVOR[32] includes 8,432 deletions and 12,091 insertions. Precision is 96.13% and recall is 95.99% (**Figure 3c**, **Supplementary Table 4**), with similar performance for insertions as deletions and for variants <1kb as ≥1kb (**Figure 3d**), indicating the complementarity of mapping- and assembly-based structural variant calling.

For comparison, structural variants were called in Illumina 2×250 bp short reads (with Manta[33] and Delly[34]) and 10X Genomics linked reads (with LongRanger[35]) available from GIAB[22]. The Manta callset has precision of 85.34% and recall of 55.88%, with

much worse recall for insertions (39.65%) than deletions (76.90%). The LongRanger callset has precision of 83.79% and recall of 39.83%, again with worse recall for insertions (16.41%) than deletions (70.18%). A callset from paftools run on a linked-read SuperNova assembly has precision of 64.52% and recall of 52.74% (**Supplementary Table 4**, **Supplementary Figure 8**). All considered short- and linked-read callsets have worse performance than all CCS callsets in both precision and recall.

De Novo Assembly of CCS Reads

Three different algorithms – FALCON[36], Canu[37], and wtdbg2 (https://github.com/ruanjue/wtdbg2) – were used to assemble the full CCS read set, which is a mix of paternal and maternal reads. By skipping the initial read-to-read error correction step, the algorithms completed 10-100× faster than is typical for long-read assemblies[19] (**Supplementary Table 5**). All assemblies have high contiguity with a contig N50 from 15.43 to 28.95 Mb. The total assembly size is near the expected human genome size for FALCON and wtdbg2. The Canu assembly has a total genome size of 3.42 Gb, larger than the expected haploid human genome, because it resolves some heterozygous alleles into separate contigs (**Table 2**, **Supplementary Figure 9**).

Short reads from the parents of HG002 were used to identify k-mers unique to one parent and then partition ("trio bin") the CCS reads by haplotype[38]. Three different k-mer sizes were evaluated: 21 bp (previously reported for trio binning) and longer k-mers of 51 bp and 91 bp enabled by the accuracy of CCS reads. The 21-mer binning assigns 35.3% of reads to the mother and 33.6% to the father (68.9% binned). The 51-mer binning is more complete at 78.5% binned; using longer 91-mers provides only a small additional gain to 79.2% binned. The 51-mer binning was selected for assembly (**Supplementary Table 6**).

FALCON, Canu, and wtdbg2 were run separately on the paternal and maternal reads, with the unassigned reads included in both sets. All algorithms produce highly contiguous and nearly complete assemblies for the parental genomes, with N50 from 12.10 to 19.99 Mb and genome size from 2.67 to 3.04 Gb (**Table 2**). From 95.3% to 98.2% of human genes are identified as single-copy in each parental assembly (**Table 2**). Assembly-based structural variant calls have high precision and recall, suggesting few large-scale mis-assemblies (**Supplementary Table 4**). Furthermore, analysis of the phase-consistency[39] of maternal and paternal haplotigs shows the assemblies are phased properly (**Supplementary Figure 10**).

All mixed and parental assemblies are high quality with concordance to the HG002 benchmark ranging from Q44-Q48 for polished[40] and Q26-Q45 for unpolished assemblies (**Table 2**, **Supplementary Table 7**). This greatly exceeds that of previously published and accessioned assemblies at Q40 (6× worse) for PacBio noisy long reads and Q29 (77× worse) for Oxford Nanopore reads with Illumina polishing (**Figure 4a**, **Supplementary Table 7**).

Large segmental duplications often result in contig breaks in *de novo* assemblies, and assemblies of noisy long reads typically span less than 50 of 175 Mb of segmental duplications in the human genome[15,41,42]. The most contiguous assemblies of CCS reads span over 60 Mb of segmental duplications, a 20% improvement (**Supplementary Table 8**). A model of assembly contiguity based on large repeat resolution suggests that the current assemblies of CCS reads resolve 15 kb repeats of 99 to 99.5% identity (**Figure 4b**).

Coverage Requirements for Variant Calling and *De Novo* Assembly

To evaluate the depth of CCS read coverage required for variant calling and assembly, we randomly subsampled from the full dataset. For SNVs, precision and recall with DeepVariant remain above 99.5% for coverage down to 15-fold; performance decays steeply below 10-fold (**Supplementary Figure 11a**). For indels, DeepVariant remains comparable to typical NGS performance (>90%) down to 17-fold coverage (**Supplementary Figure 11b**). For structural variants, precision with pbsv is above 95% for all evaluated coverage levels. Recall is above 90% down to 15-fold coverage, and decays steeply below 10-fold (**Supplementary Figure 11c**). For phasing with WhatsHap, the phase block N50 remains above 150 kb down to 10-fold coverage (**Supplementary Figure 11d**). Mixed-haplotype wtdbg2 assemblies have consistent size above 2.7 Gb, contig N50 around 15 Mb, and concordance above Q42 until coverage falls below 15-fold (**Supplementary Figure 12**).

Revising and Expanding Genome in a Bottle Benchmarks

High-quality callsets from CCS reads provide an opportunity to identify mistakes in the GIAB benchmarks, particularly for structural variants where the benchmark is still in draft form. Sixty small variant and 40 structural variant discrepancies between the GIAB benchmark (small variant v3.3.2, structural variant v0.6) and the CCS callsets (DeepVariant haplotype-sorted, structural variant integrated) were selected for manual curation. Selected variants were spread across variant types, discrepancy types, and both inside and outside homopolymers and tandem repeats.

For small variants, 29 of 31 discrepancies in homopolymers were classified as correct in the benchmark. Outside of homopolymers, 19 of 29 were classified as errors in the benchmark. Most of these benchmark errors (13 of 19) are true variants in L1 elements called homozygous reference in GIAB (**Supplementary Figure 13**, **Supplementary Table 9**). The identified benchmark errors overlap with putative errors in a DeepVariant Illumina whole genome case study (https://github.com/google/deepvariant). Of 745 putative false positive SNVs in the case study, 344 agree with the CCS callset, with 282 (82.0%) falling within large interspersed repeats. Fewer of the false negative SNVs (8%), false negative indels (25%), and false positive indels (19%) from the case study agree with the CCS callset. Extrapolating from manual curation, we estimate that 2,434 (1,313-2,611; 95% confidence interval) errors in the current GIAB benchmark could be corrected using the CCS reads.

For structural variants, curator classification was unclear for 11 of 40 discrepancies, typically because of tandem repeat structure that permits multiple representations of a variant. For the remainder, 15 of 16 false negative discrepancies were classified as correct in the benchmark. However, for false positive discrepancies, 11 of 13 were classified as errors in the benchmark (**Supplementary Figure 13**, **Supplementary Table 10**). This suggests that the GIAB structural variant benchmark set is precise but incomplete.

The high-quality CCS callsets also provide an opportunity to expand the benchmarks into repetitive and highly polymorphic regions that have been difficult to characterize with confidence using short reads. Adding the CCS DeepVariant callset to the existing GIAB small variant integration pipeline would expand the benchmark regions by up to 1.3% and 418,875 variants (210,184 SNVs and 208,691 indels). For structural variants, only 9,232 of 18,832 autosomal variant calls overlap benchmark regions, which means that the number of variants in the benchmark would more than double if all CCS variants calls were incorporated.

Discussion

We present a protocol for producing highly-accurate long reads using circular consensus sequencing (CCS) on the PacBio Sequel System. We apply the protocol to sequence the human HG002 to 28-fold coverage with average read length of 13.5 kb and an average read accuracy of 99.8%. We analyze the CCS reads to call SNVs, indels, and structural variants; to phase variants into haplotype blocks; and to *de novo* assemble the HG002 genome.

The CCS performance for SNV and indel calling rivals that of the commonly-used pairing of BWA and GATK on 30-fold short-read coverage. Interestingly, though the overall accuracy of CCS reads is similar to short reads, direct application of the GATK pipeline to CCS reads produces inferior results, especially for indels. The major residual error in CCS reads – indels in homopolymers – is not as frequent in short reads. We suspect that the current GATK, which was designed for short reads, does not properly model the CCS error profile, and thus performance lags for indels. This is supported by results with DeepVariant. When a DeepVariant model trained on Illumina reads is run on CCS reads, the performance is poor for indels. When DeepVariant is trained on CCS reads, performance improves dramatically. As more CCS datasets are made available, both model-based callers like GATK and learning-based callers like DeepVariant will have the opportunity to improve on the performance reported here, including by incorporating haplotype phase information and evaluating and training against updated GIAB benchmarks that correct errors with CCS reads. Further, advances in sequencing chemistry or consensus base calling (such as the application of deep learning) that reduce the residual indel errors in CCS reads also could improve variant calling performance.

Structural variant calling and *de novo* genome assembly with CCS reads match or exceed that reported for noisy long reads. The CCS reads have an advantage of high accuracy, which eliminates the need for read correction, allows more stringent criteria to be used in variant calling or read overlapping, and ultimately produces more accurate assemblies and variant calls. Noisy long reads have an advantage of longer maximum read length, but increased accuracy of CCS reads compensates for the length required for highly contiguous assembly. Modeling (**Figure 4b**) suggests modest advances in accuracy (to 99.9%) at 15 kb read length would double the current contiguity, which already matches the best published *de novo* assemblies[43].

The CCS read approach alleviates some other challenges of long-read sequencing. First, aiming for fragments in the 10-20 kb size range relaxes the need to isolate ultra-long genomic DNA. Second, increased accuracy allows for more stringent alignment and overlap comparisons, greatly reducing the compute time and cost while improving assembly results by recognizing fine-grained repeat and haplotype phase information. Third, familiar tools like GATK that were developed for accurate short reads are readily applied to CCS reads.

The major current limitation of CCS reads compared to noisy long reads is reduced throughput per run, and thus higher cost per basepair. A lower coverage requirement partially compensates; variant calling and assembly with CCS reads perform well down to 15-fold coverage. This shortcoming ultimately will be alleviated with more unique reads per run (5-10× is expected) and increased read length to allow larger fragment sizes. Such advances should facilitate rapid, population-scale analysis of full genomes to improve human health.

Methods

CCS Library Preparation

PacBio library preparation and sequencing was performed on the human reference genome sample HG002 obtained from NIST. Genomic DNA was sheared using the Megaruptor® from Diagenode with a long hydropore cartridge and a 20 kb shearing protocol. Prior to library preparation, the size distribution of the sheared DNA was characterized on the Agilent 2100 BioAnalyzer System using the DNA 12000 kit. A sequencing library was constructed from this sheared genomic DNA using the SMRTbell™ Template Prep Kit v 1.0 (Pacific Biosciences Ref. No. 100-259-100). In order to tighten the size distribution of the SMRTbell™ library, the sample was separated into 3 kb fractions using the SageELF System from Sage Science. Fractions having the desired size distribution ranges were identified on the Agilent 2100 BioAnalyzer using the DNA 12000 kit (**Supplementary Figure 1**). The fraction centered at 15 kb was used for sequencing.

Sequencing

Sequencing reactions were performed on the PacBio Sequel System with the Sequel Sequencing Kit 3.0 chemistry. The samples were pre-extended without exposure to illumination for 12 hours to enable the polymerase enzymes to transition into the highly processive strand-displacing state and sequencing data was collected for 24 hours to ensure suitably long read lengths.

Consensus reads ("CCS reads") were generated using the ccs software version 3.0.0 (https://github.com/pacificbiosciences/unanimity/) with `--minPasses 3 --minPredictedAccuracy 0.99`. Average run time is 3,035 CPU core hours per SMRT Cell (118,365 total).

Read Mapping

CCS reads were mapped to the GRCh37 human reference genome using pbmm2 version 0.10.0 (https://github.com/PacificBiosciences/pbmm2) with `--preset CCS --sample HG002 --sort`. For read mapping and subsequent analysis, GRCh37 refers to the hs37d5 build from the 1000 Genome Project[44].

Measuring HG002 Concordance

To measure concordance to HG002, alignments to GRCh37 were evaluated at positions within GIAB v3.3.2 high-confidence regions that have no high-confidence variant call[23]. `Concordance = 1-(M+D+I)/(X+M+D)` where M is the number of matches, X is the number of mismatches, D is the number of deletion basepairs, and I is the number of insertion basepairs. $\text{Phred} = -10*\log_{10}(1-\text{concordance})$. Reads with perfect concordance are assigned a Phred score of 60.

A deleted basepair is considered a homopolymer deletion when it matches the preceding or following basepair in the reference genome. An insertion is considered a homopolymer insertion when the basepairs of the insertion are identical and match either the preceding or following basepair in the reference genome.

Coverage by [GC] content

To measure coverage by local [GC] content, bedtools[45] version 2.27.1 was used to divide the GRCh37 reference genome into 500 bp windows (`bedtools makewindows -w 500`) and then to calculate the [GC] content (`bedtools nuc`) and average coverage (`bedtools coverage -mean`) of each window.

Reference-independent Quality Evaluation

The Dazzler suite (https://dazzlerblog.wordpress.com/) was used to evaluate the accuracy of the CCS reads without relying on a reference genome. Briefly, daligner[24] was used to produce all local alignments longer than 1 kb between pairs of CCS reads. Each CCS subject read was partitioned into 100 bp panels, within which its coverage by and concordance to aligned target reads was calculated. Panels with a concordance in

the worst 0.1% were considered low quality. Abrupt breaks in read-to-read alignments were used to estimate library artifacts like chimeric molecules and missing adapters.

Mappability of CCS and NGS Reads

To compare with the mappability of 13.5 kb CCS reads, a coverage-matched (89 Gb) set of 2×250bp Illumina HiSeq 2500 reads for HG002 were obtained from GIAB[22] and mapped to GRCh37 with minimap2[31] version 2.14-r883 with `-x sr`.

A genome position is considered mappable if it is covered by alignments for at least ten reads at a specified mapping quality or higher, which was evaluated using `bedtools bamtobed` and `bedtools genomecov -bga`. Gaps ("N" basepairs in the reference) were excluded.

Previously-reported NGS problem exons[25] were considered mappable if every basepair in the exon is covered by a read at mapping quality of 60.

HLA Typing

The *HLA-A* and *HLA-DPA1* genes were typed by comparing the sequence of CCS reads that span the genes to entries in the IMGT database[46] version 3.19.0.

Small Variant Detection and Benchmarking

To develop a workflow for calling variants in CCS reads with GATK[27] HaplotypeCaller v4.0.6.0, different values of the HaplotypeCaller parameter `--pcr-indel-model` and VariantFiltration parameter `--filter-expression` were considered to maximize SNV and indel F1 without excessive complication, starting from the GATK best practices for hard filtering. In the end, HaplotypeCaller was run on reads with a minimum mapping quality of 60 using allele-specific annotations (`--annotation-group AS_StandardAnnotation`) and `--pcr_indel_model AGGRESSIVE`. Autosomes and the pseudo-autosomal regions (PARs) on chromosome X were called with `--ploidy 2`; chromosome Y and the non-PAR regions of chromosome X were called with `--ploidy 1`. Multi-allelic variant sites were split into separate entries for filtration with a custom script (https://gist.github.com/williamrowell/16cd89fcb23ab9f11a7bd387c308d29d). SNVs were filtered using GATK VariantFiltration with `--filter_expression` of `AS_QD < 2.0` for SNVs and indels longer than 1bp, and `AS_QD < 5.0` for 1 bp indels. A similar pipeline was used to call variants in coverage-matched 2×151 bp Illumina NovaSeq reads with a few differences: a minimum mapping quality of 20, `--pcr-indel-model NONE`, `--standard_min_confidence_threshold_for_calling 2.0`, and no variant filtration.

A Google DeepVariant model for CCS reads was generated as previously reported[28] using DeepVariant version 0.7.1. Briefly, models were trained using CCS reads for chromosomes 1-19 and the HG002 GIAB v3.3.2 high-confidence regions and variants.

A single model was selected based on performance in chromosomes 21 and 22 to avoid overfitting. Neither training nor model selection considers chromosome 20, which is available for accuracy evaluations. To support long reads, local reassembly is disabled for DeepVariant with CCS reads. The wgs_standard model version 0.7.1 was used to call variants in NovaSeq reads and to apply a model trained on Illumina reads to CCS reads.

To incorporate long-range haplotype information, DeepVariant was modified to produce pileups with reads sorted by the BAM haplotype ("HP") tag. Haplotype information was added to the pbmm2 CCS alignments using WhatsHap v0.17 (`whatshap haplotag`) with the trio-phased variant calls from GIAB (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_MPI_whatshap_08232018/RTG.hg19.10x.trio-whatshap.vcf.gz). A new DeepVariant model then was trained as described above.

Small variant callsets were benchmarked against the GIAB v3.3.2 HG002 set[23] by vcfeval[47] (https://github.com/RealTimeGenomics/rtg-tools) with no partial credit run through hap.py version 0.3.10 (https://github.com/Illumina/hap.py). Only PASS calls were considered.

Phasing Small Variants

Small variant calls were phased using WhatsHap v0.17 (`whatshap phase`). The number of switch and Hamming errors was computed against trio-phased variant calls from GIAB using `whatshap compare`.

To model the phase blocks achievable with a given read length, cuts were introduced between heterozygous variants in the GIAB trio-phased variant callset that are separated by more than the read length, which effectively assumes that adjacent heterozygous variants separated by less than the read length can be phased.

Structural Variant Detection

pbsv version 2.1.0 (https://github.com/PacificBiosciences/pbsv) was run on pbmm2 CCS read alignments. The `pbsv discover` stage was run separately per chromosome with tandem repeat annotations (https://github.com/PacificBiosciences/pbsv/tree/master/annotations) passed with `--tandem-repeats`. The `pbsv call` stage was run on the full genome.

Sniffles version 1.0.10 was run on pbmm2 CCS reads alignments with `-s 3 --skip_parameter_estimation` and with the variant sequence obtained from reads.

Structural variants in the maternal and paternal Canu and FALCON assemblies from CCS reads (see "*De novo* Assembly") were called using a previously described workflow[48]. Briefly, contigs were mapped to GRCh37 using `minimap2 --paf-no-`

`hit --cxasm5 --cs -r2k`; variants were called with `paftools.js call`[31]; maternal and paternal variants were concatenated; and indel calls at least 30 bp were retained.

An integrated callset was produced from the pbsv, Sniffles, and paftools/Canu callsets using SURVIVOR[32] and custom scripts. Two calls were considered supporting if the calls had the same structural variation type, a start position with 1 kb, and a difference in length less than 5%. One call from each matching set was retained with precedence given to pbsv, then Sniffles, and then paftools. Because pbsv and Sniffles have poor sensitivity for calls larger than 1 kb, all non-matched calls from paftools that are larger than 1 kb were retained.

NovoAlign (http://www.novocraft.com) alignments to GRCh37 of 300-fold coverage of HG002 with 2×250bp Illumina HiSeq 2500 reads were obtained from GIAB. Structural variants were called with Manta[33] version 1.4.0 with all coverage, and Delly[34] version 0.7.6 with coverage subsampled to 30-fold using `samtools view -b -s 0.1`.

Structural variant callsets on 10X Genomics reads from LongRanger version 2.2 were obtained from GIAB (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/10XGenomics_ChromiumGenome_LongRanger2.2_Supernova2.0.1_04122018/). Insertion and deletion variants at least 30 bp were combined from the sequence-resolved indels and large deletion calls (NA24385_LongRanger_snpindel.vcf.gz, NA24385_LongRanger_sv_deletions.vcf.gz). Another callset was produced using paftools on the diploid Supernova 2.0.1 assembly as described above.

Structural variant callsets were benchmarked against the GIAB v0.6 HG002 structural variant set (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6) using Truvari (https://github.com/spiralgenetics/truvari) commit 600b4ed7 modified to allow a single variant in the truth set to support multiple variants in the callset. Truvari was run with `-r 1000 -p 0.01 --multimatch --includebed HG002_SVs_Tier1_v0.6.bed -c HG002_SVs_Tier1_v0.6.vcf.gz`. The `-p 0` option was used to disable sequence checks for callsets that report symbolic alleles instead of sequence-resolved calls (LongRanger, Delly).

De novo Assembly

Mixed haplotype assemblies were produced using all CCS reads. Canu[37] version 1.7.1 was run with `-p asm genomeSize=3.1g correctedErrorRate=0.015 ovlMerThreshold=75 batOptions="-eg 0.01 -eM 0.01 -dg 6 -db 6 -dr 1 -ca 50 -cp 5" -pacbio-corrected`. FALCON[36] kit version 1.2.0 was run with `ovlp_HPCdaligner_option = -v -B128 -M24 -k24 -h1024 -e.97 -l2500 -s100`, `ovlp_DBsplit_option = -s400`, and `overlap_filtering_setting = --max-diff 90 --max-cov 120 --min-cov 2`. Wtdbg2

(https://github.com/ruanjue/wtdbg2) version 2.2 was run with `-k 0 -p 21 -AS 4 -s 0.5 -e 2 -K 0.05` and followed by wtdbg2-cns.

CCS reads from HG002 were "trio binned" as maternal, paternal, or unassigned as previously described[38]. Briefly, 2×250 bp Illumina HiSeq 2500 reads for the father (HG003/NA24149) and mother (HG004/NA24143) of HG002 were obtained from GIAB. Sequence k-mers unique to the mother or father were identified and used to categorize CCS reads (https://github.com/skoren/triobinningScripts), using k-mer size of 21, 51, and 91 and excluding k-mers that occur 25 times or fewer. The maternal and unassigned reads were used for the "maternal" assemblies; paternal and unassigned reads were used for the "paternal" assemblies.

The maternal and paternal assemblies were generated with Canu and wtdbg2 using the same software version and same options as for the mixed haplotype assembly. For the maternal and paternal assemblies, FALCON version 0.7 was run with `length_cutoff_pr = 2000`, `ovlp_HPCdaligner_option = -k24 -e.95 -s100 -l1000 -h600 -mdust -mrep8 -mtan -M21`, `ovlp_DBsplit_option = -x2000 -s400`, `falcon_sense_option = --min_idt 0.70 --min_cov 4 --max_n_read 200`, and `overlap_filtering_setting = --max_diff 40 --max_cov 80 --min_cov 2 --min_len 500`.

The maternal and paternal Canu assemblies were polished with Arrow version 2.2.2 run through ArrowGrid (https://github.com/skoren/ArrowGrid) using subreads that correspond to the CCS reads used for each assembly. The maternal and paternal FALCON assemblies were polished with Arrow version 2.2.2 using all subreads.

Assembly Evaluation

For each assembly, contigs were broken into 100 kb chunks with remainders shorter than 100 kb ignored. The chunks were aligned to GRCh37 using `minimap2 --eqx -x asm5`, and primary alignments that span at least 50 kb in the reference at higher than 50% identity were retained. The concordance of each chunk was evaluated just as for CCS reads (see "Measuring HG002 Concordance"). The overall assembly concordance was calculated as the average concordance of the 100 kb chunks.

Gene completeness was measured using BUSCO[49] version 3.0.2 using the Mammalia ODB9 gene set. The single plus duplicated gene count in the BUSCO summary is reported. For a human-specific measure of completeness, we calculated the fraction of single-copy human genes that remain single-copy in each assembly. The human transcript sequences from ENSEMBL[50] build r94 were mapped to each assembly with `minimap2 -cx splice -B 4 -O 4,34 -C9 -uf --cs` and evaluated with `paftools.js asmgene -i 0.98`, which retains the longest of overlapping transcripts, and counts a transcript hit if 99% of the transcript sequence maps at 98% identity or higher. A single-copy transcript has exactly one hit. Counts are normalized

to the number of transcripts that are considered single-copy by these criteria in GRCh38 (GCA_000001405.15).

To measure the number of segmental duplications spanned by each assembly, the assemblies were processed with segDupPlots[42] (https://github.com/mvollger/segDupPlots), which maps contigs to GRCh38 and considers a segmental duplication to be spanned by the assembly if a contig alignment extends through the segmental duplication with at least 50 kb on each flank.

Model of Assembly Contiguity

To predict assembly contiguity at different read lengths and read accuracies, a previously described model[19] was updated with improvements for high-accuracy reads. Briefly, all repeat annotations for GRCh38 were downloaded from the UCSC Genome Browser. Repeat identity was defined as by each track except for: the nested repeat track where identity was 50+50*score/1000, RepeatMasker where identity was 1-((mismatches + deleted + inserted)/1000), and microsat and windowmasker/sdust which does not define identity and thus was treated as 100%. Gaps were included as 100% identity repeats. Additional repeats were added from self-matches using MashMap[51] (https://github.com/marbl/MashMap).

The assembly contiguity was predicted based on the ability to resolve repeats. At a given percent identity, repeats below that identity were excluded and remaining repeats separate by 15 bp or fewer were merged. Then, cuts were introduced at each at repeats of each given length, and assembly NG50 was calculated assuming that contigs end at each cut.

Coverage Titration

To evaluate the performance of variant calling and assembly at different coverage levels, CCS reads were downsampled from the 28-fold dataset and processed. For small variant calling, alignments were subsampled in DeepVariant version 0.7.1 from 4% to 100% in steps of 3%. Variants were called on each subsample using the DeepVariant CCS model. Precision and recall for SNVs and indels were evaluated with hap.py as described above (see "Small Variant Detection and Benchmarking"). For phasing, alignments were subsampled (`samtools view -s`) at rates from 10% to 100% in steps of 10%. The DeepVariant callset from the full 28-fold coverage data was phased using WhatsHap v0.17 (`whatshap phase`) with the subsampled alignments. For structural variants, alignments were subsampled (`samtools view -s`) at rates from 10% to 100% in steps of 10%. Variants were called on the subsampled alignments with pbsv version 2.1.0 and benchmarked with Truvari as described above (see "Structural Variant Detection"). For assembly, reads were subsampled at rates from 10% to 100% in steps of 10%. Sampling was performed based on read name (10% sample is reads that end in 0, 20% is reads that end in 0-1, and so on). Assembly of subsamples reads was performed with wtdbg2 version 2.2 and benchmarked as described above (see "*De novo* Assembly" and "Assembly Evaluation").

Revising and Expanding Genome in a Bottle Benchmarks

Discrepancies between the GIAB v3.3.2 small variant benchmark and the DeepVariant callset from haplotype-sorted CCS reads were identified with vcfeval and hap.py. Discrepancies between the GIAB v0.6 structural variant benchmark and the integrated structural variant callset from CCS reads were identified with Truvari. A sample of 60 small variant and 40 structural variant discrepancies were selected for manual curation by random sampling across discrepancy types (false positive, false negative, genotype difference), variant types (SNV, indel, insertion structural variant, and deletion structural variant), both inside and outside homopolymers and tandem repeats. Curators evaluated variants in IGV along with alignments of CCS reads, 10X Genomics reads, Illumina short reads, and Illumina reads from a 6 kb mate pair library, all obtained from GIAB. The benchmark error rate was estimated by variant type and discrepancy type and used to extrapolate from the sample to the number of errors in the full GIAB benchmark. Confidence intervals were calculated assuming a binomial distribution.

Figures

Figure 1

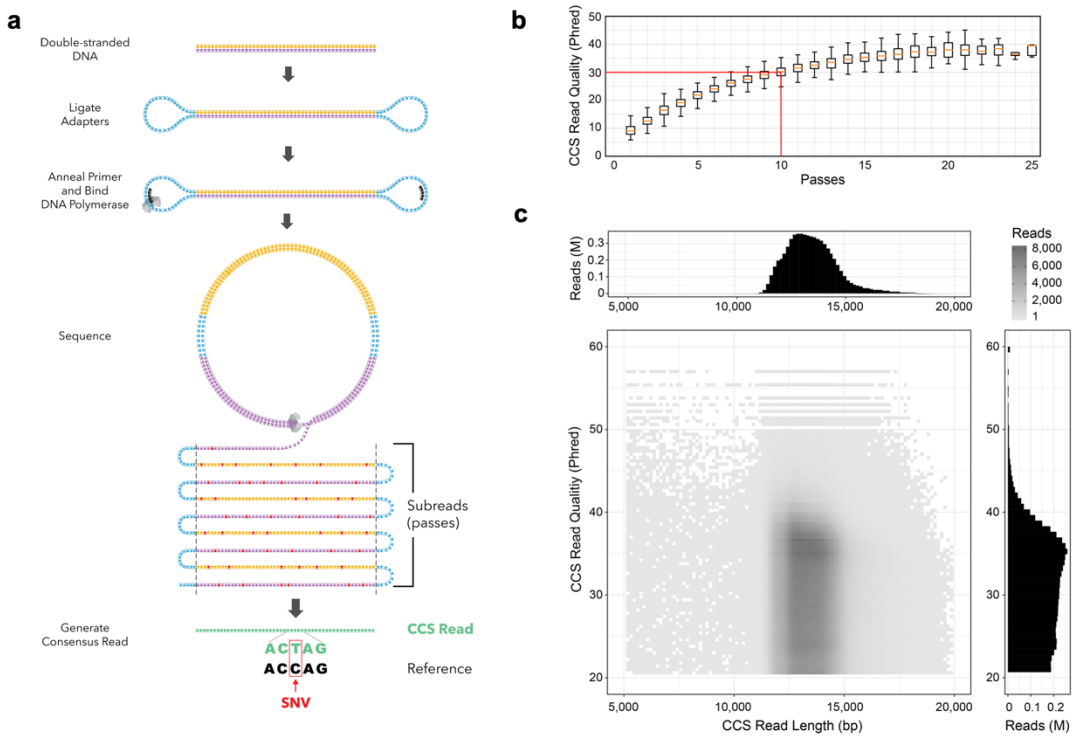
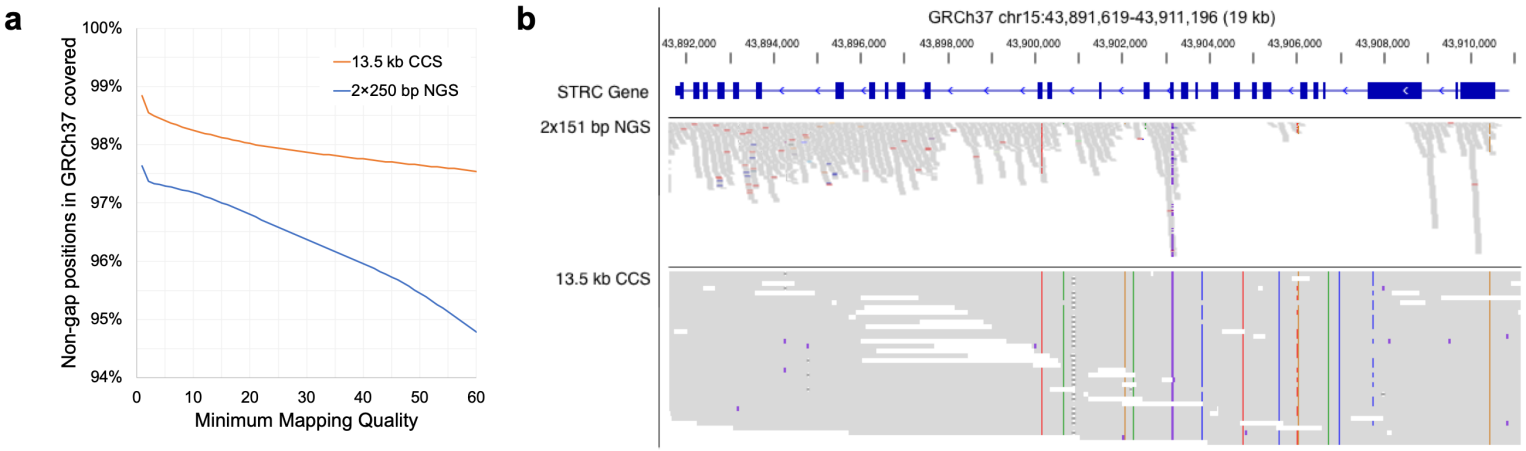


Figure 1. Sequencing HG002 with highly-accurate, long reads. (a) Circular consensus sequencing (CCS) derives a consensus read from multiple passes of a single template molecule, producing accurate reads from noisy individual subreads (passes). (b) Predicted quality for CCS reads with different numbers of passes, for sequencing of the human male HG002. At 10 passes, the median read achieves Q30 predicted quality. (c) Length and predicted quality of CCS reads.

Figure 2



c

Improved mappability of "NGS problem exons" in 193 medically-relevant genes with 13.5 kb CCS reads

% problem exons mappable	Genes	Genes
100%	*ABCC6, ABCD1, ACAN, ACSM2B, AKR1C2, ALG1, ANKRD11, BCR, CATSPER2, CD177, CEL, CES1, CFH, CFHR1, CFHR3, CFHR4, CGB, CHEK2, CISD2, CLCNKA, CLCNKB, CORO1A, COX10, CRYBB2, CSH1, CYP11B1, CYP11B2, CYP21A2, CYP2A6, CYP2D6, CYP2F1, CYP4A22, DDX11, DHRS4L1, DIS3L2, DND1, DPY19L2, DUOX2, ESRRA, F8, FAM120A, FAM205A, FANCD2, FCGR1A, FCGR2A, FCGR3A, FCGR3B, FLG, FLNC, FOXD4, FOXO3, FUT3, GBA, GFRA2, GON4L, GRM5, GSTM1, GYPA, GYPB, GYPE, HBA1, HBA2, HBG1, HBG2, HP, HS6ST1, IDS, IFT122, IKBKG, IL9R, KIR2DL1, KIR2DL3, KMT2C, KRT17, KRT6A, KRT6B, KRT6C, KRT81, KRT86, LEFTY2, LPA, MST1, MUC5B, MYH6, MYH7, NEB, NLGN4X, NLGN4Y, NOS2, NOTCH2, NXF5, OPN1LW, OR2T5, OR51A2, PCDH11X, PCDHB4, PGAM1, PHC1, PIK3CA, PKD1, PLA2G10, PLEKHM1, PLG, PMS2, PRB1, PRDM9, PROS1, RAB40AL, RALGAPA1, RANBP2, RHCE, RHD, RHPN2, ROCK1, SAA1, SDHA, SDHC, SFTPA1, SFTPA2, SIGLEC14, SLC6A8, SMG1, SPATA31C1, SPTLC1, SRGAP2, SSX7, STAT5B, STK19, STRC, SULT1A1, SUZ12, TBX20, TCEB3C, TLR1, TLR6, TMEM231, TNXB, TRIOBP, TRPA1, TTN, TUBA1A, TUBB2B, UGT1A5, UGT2B15, UGT2B17, UNC93B1, VCY, VWF, WDR72, ZNF419, ZNF592, ZNF674*	152
[75%, 100%)	*ANAPC1, C4A, C4B, CHRNA7, CR1, DUX4, FCGR2B, HYDIN, OTOA, PDPK1, TMLHE*	11
[50%, 75%)	*ADAMTSL2, CDY2A, DAZ1, GTF2I, NAIP, OCLN, RPS17*	7
[25%, 50%)	*DAZ2, DAZ3, KIR3DL1, OPN1MW, PPIP5K1*	5
(0%, 25%)	*NCF1, RBMY1A1*	2
0%	*BPY2, CCL3L1, CCL4L1, CDY1, CFC1, CFC1B, GTF2IRD2, HSFY1, MRC1, OR4F5, PRY, PRY2, SMN1, SMN2, TSPY1, XKRY*	16

Figure 2. Mappability of the human genome with CCS reads. (a) Percentage of the non-gap GRCh37 human genome covered by at least 10 reads from 28-fold coverage NGS (2×250 bp) and CCS (13.5 kb) datasets at different mapping quality thresholds. (b) Coverage of the congenital deafness gene *STRC* in HG002 with 2×151 bp NGS (NovaSeq) reads and 13.5 kb CCS reads at a mapping quality threshold of 10. (c) Improvement in mappability with 13.5 kb CCS reads for 193 human genes previously reported as medically-relevant and problematic to map with NGS reads[25].

Figure 3

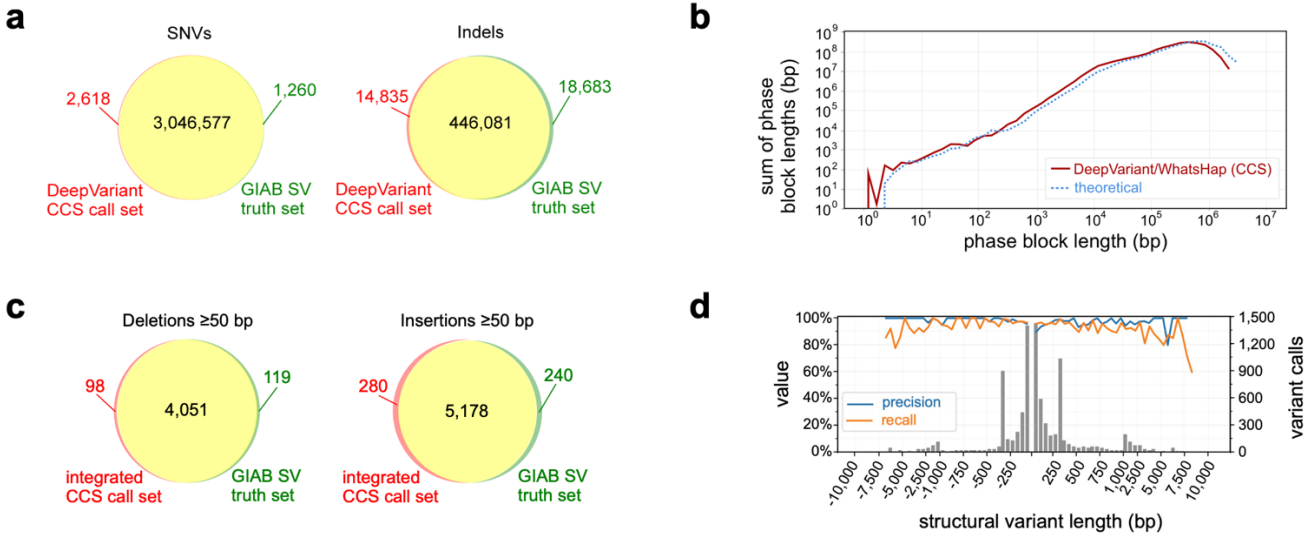


Figure 3. Variant calling and phasing with CCS reads. (a) Agreement of DeepVariant SNV and indel calls with Genome in a Bottle v3.3.2 benchmark measured with hap.py. (b) Phasing of heterozygous DeepVariant variant calls with WhatsHap, compared to theoretical phasing of HG002 with 13.5 kb reads. (c) Agreement of integrated CCS structural variant (SV) calls with the Genome in a Bottle v0.6 structural variant benchmark measured with Truvari, (d) by variant size. Negative length indicates a deletion; positive length indicates an insertion. The histogram bin size is 50 bp for variants shorter than 1 kb, and 500 bp for variants >1 kb.

Figure 4

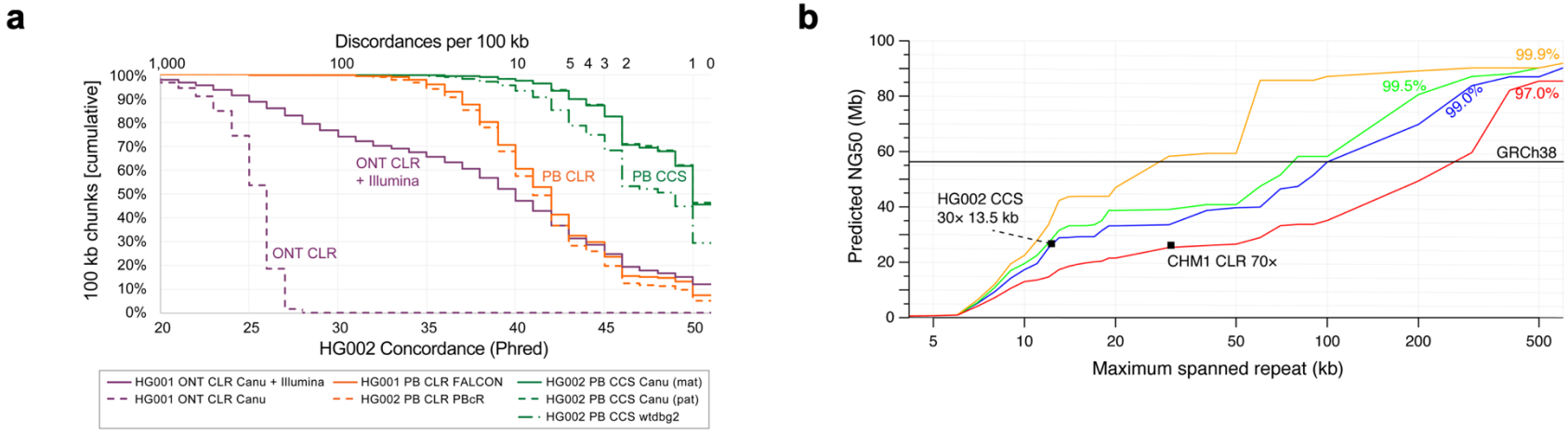


Figure 4. Impact of read accuracy on *de novo* assembly. (a) The concordance of seven assemblies to the reference genome at non-variant positions in Genome in a Bottle high-confidence regions (**Supplementary Table 7**). Contigs longer than 100 kb were segmented into 100 kb chunks and aligned to GRCh37. Concordance was measured per chunk, and chunks with no discordances were assigned concordance of Q51. PB=PacBio, ONT=Oxford Nanopore, CLR=continuous ("noisy") long reads. (b) Predicted contiguity of a human assembly based on ability to resolve repeats of different lengths (x-axis) and percent identities (colored lines)[19]. The solid line indicates the contiguity of GRCh38. The 97.0% identity line is representative of CLR assemblies using standard read-to-read error correction. The points show example CCS and CLR[43] assemblies using Canu. Repeat identity and length are proxies for read accuracy and length.

Tables

Table 1

			SNVs			Indels		
Platform	**Coverage**	**Variant Caller (training model)**	**Precision**	**Recall**	**F1 ^**	**Precision**	**Recall**	**F1**
Illumina (NovaSeq)	30-fold	DeepVariant (Illumina model)	**99.960%**	99.940%	**99.950%**	**99.633%**	**99.413%**	**99.523%**
PacBio (CCS)	28-fold	DeepVariant (CCS model)	99.914%	99.959%	99.936%	96.901%	95.980%	96.438%
PacBio (CCS)	28-fold	DeepVariant (haplotype-sorted CCS model)	99.904%	**99.963%**	99.934%	97.835%	97.141%	97.486%
Illumina (NovaSeq)	30-fold	GATK HaplotypeCaller (no filter)	99.852%	99.910%	99.881%	99.371%	99.156%	99.264%
PacBio (CCS)	28-fold	GATK HaplotypeCaller (hard filter)	99.468%	99.559%	99.513%	78.977%	81.248%	80.097%

Table 1. Performance of small variant calling with CCS reads. Precision, recall, and F1 of small variant calling measured against the Genome in a Bottle v3.3.2 benchmark using hap.py. **Bold** indicates the highest value in each column. Underline indicates a value higher than the GATK HaplotypeCaller run on 30-fold Illumina NovaSeq reads. Rows are sorted ("^") based on F1 for SNVs.

Table 2

Haplotype	Assembler	Total Size (Gb)	Contigs	N50 (Mb)	NG50 (Mb)	Max (Mb)	E-size[52] (Mb)	HG002 Concordance (Phred)	BUSCO Genes	RefSeq Genes
Mixed	Canu	3.42	18,006	22.78	25.02	108.46	30.16	31.1	92.3%	93.2%
Mixed	FALCON	2.91	2,541	28.95	24.51	110.21	38.04	25.8	87.6%	97.6%
Mixed	wtdbg2	2.79	1,554	15.43	12.62	84.67	22.61	44.6	94.2%	96.1%
Maternal	Canu*	3.04	5,854	18.02	17.04	48.81	19.78	47.2	94.1%	98.1%
Maternal	FALCON*	2.80	924	19.99	15.54	74.33	24.07	43.5	95.1%	97.8%
Maternal	wtdbg2	2.75	2,637	12.10	9.29	66.34	16.55	43.5	93.8%	95.6%
Paternal	Canu*	2.96	6,868	16.14	14.90	64.83	20.19	47.7	93.4%	98.2%
Paternal	FALCON*	2.70	1,489	16.40	14.06	95.34	25.61	43.5	93.6%	97.7%
Paternal	wtdbg2	2.67	1,444	13.96	10.86	50.51	15.36	42.1	92.6%	95.3%

Table 2. Statistics for *de novo* assembly of CCS reads. The “mixed” haplotype assemblies use all reads. The “maternal” and “paternal” assemblies use parent-specific reads from trio binning plus unassigned reads. HG002 concordance is measured at non-variant positions in Genome in a Bottle high-confidence regions. BUSCO gene completeness uses the Mammalia ODB9 gene set. RefSeq genes is the percentage of genes from Ensembl R94 that are full-length, single-copy in the assembly relative to the full-length, single-copy count for GRCh38. Contigs shorter than 13 kb were excluded from genome size and contiguity measurements; contigs shorter than 100 kb were excluded from the concordance measurement. “*” indicates polishing with Arrow.

Author Contributions

AMW, DRR, MWH, and PP conceived the study. DRR and PP developed the sample preparation protocol and performed sample preparation. DRR, PP, and YQ performed sequencing. AC, AF, AK, AMP, AMW, AT, CSC, DRR, FJS, GM, GTC, HL, JE, JMZ, JR, MA, MAD, MCS, MM, NDO, PC, PP, RJH, SK, TM, and WJR performed analysis. AC, AMP, CSC, DRR, FJS, JMZ, MAD, MCS, and MWH supervised analysis. AMW, DRR, GM, JMZ, PP, RH, SK, and WJR wrote the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We would like to thank John R. Harting for assistance with HLA typing and Kristin A. Robertshaw for figure generation.

SK and AMP were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

This work was supported by NIH grant 1R01HG010040 to HL and NSFC grants 31571353 and 31822029 to JR. MCS is funded by the National Science Foundation (DBI-1350041) and National Institutes of Health (R01-HG006677).

Certain commercial equipment, instruments, or materials are identified to specify adequate experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Competing Financial Interests

AMW, AT, DRR, GTC, MWH, PP, RJH, WJR, and YQ are employees and shareholders of Pacific Biosciences. CSC and AF are employees of DNAnexus. AC, AK, MAD, and PC are employees and shareholders of Google.

Data Availability

CCS reads and alignments to GRCh37 are available at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/.

References

1. DNA Sequencing Costs: Data. *National Human Genome Research Institute (NHGRI)* Available at: https://www.genome.gov/27541954/dna-sequencing-costs-data/. (Accessed: 7th December 2018)

2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnol. Read. Mass* **24**, 104–108 (1992).

3. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).

4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

5. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).

6. Mouse Genome Sequencing Consortium *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

7. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796–815 (2000).

8. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).

9. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

10. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).

11. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).

12. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

13. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).

14. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).

15. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).

16. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).

17. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).

18. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).

19. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

20. 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 (2010).

21. Loomis, E. W. *et al.* Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).

22. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).

23. Zook, J. *et al.* Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials. *bioRxiv* 281006 (2018). doi:10.1101/281006

24. Myers, G. Efficient Local Alignment Discovery amongst Noisy Long Reads. in *Algorithms in Bioinformatics* (eds. Brown, D. & Morgenstern, B.) 52–67 (Springer Berlin Heidelberg, 2014).

25. Mandelker, D. *et al.* Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18**, 1282–1289 (2016).

26. Ambardar, S. & Gowda, M. High-Resolution Full-Length HLA Typing Method Using Third Generation (Pac-Bio SMRT) Sequencing Technology. *Methods Mol. Biol. Clifton NJ* **1802**, 135–153 (2018).

27. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

28. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

29. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **22**, 498–509 (2015).

30. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

31. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

32. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).

33. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

34. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinforma. Oxf. Engl.* **28**, i333–i339 (2012).

35. Garcia, S. *et al.* Linked-Read sequencing resolves complex structural variants. *bioRxiv* 231662 (2017). doi:10.1101/231662

36. Chin, C.-S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).

37. 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).

38. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4277

39. Fungtammasan, A. & Hannigan, B. How well can we create phased, diploid, human genomes?: An assessment of FALCON-Unzip phasing using a human trio. *bioRxiv* 262196 (2018). doi:10.1101/262196

40. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

41. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* 193144 (2018). doi:10.1101/193144

42. Vollger, M. R. *et al.* Long-read sequence and assembly of segmental duplications. *Nat. Methods* (2018). doi:10.1038/s41592-018-0236-3

43. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

44. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

45. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47**, 11.12.1-34 (2014).

46. Robinson, J., Soormally, A. R., Hayhurst, J. D. & Marsh, S. G. E. The IPD-IMGT/HLA Database - New developments in reporting HLA variation. *Hum. Immunol.* **77**, 233–237 (2016).

47. Cleary, J. G. *et al.* Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* 023754 (2015). doi:10.1101/023754

48. Li, H. *et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat. Methods* **15**, 595–597 (2018).

49. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma. Oxf. Engl.* **31**, 3210–3212 (2015).

50. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

51. Jain, C., Koren, S., Dilthey, A., Phillippy, A. M. & Aluru, S. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinforma. Oxf. Engl.* **34**, i748–i756 (2018).

52. Salzberg, S. L. *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).