Nanopore sequencing and assembly of a human genome with ultra-long reads

M Jain[1,§], S Koren[2,§], KH Miga[1,§], J Quick[3,§], AC Rand[1,§], TA Sasani[4,5,§], JR Tyson[7,§], AD Beggs[8], AT Dilthey[2], IT Fiddes[1], S Malla[9], H Marriott[9], T Nieto[8], J O'Grady[10], HE Olsen[1], BS Pedersen[4,5], A Rhie[2], H Richardson[10], AR Quinlan[4,5,6], TP Snutch[7], L Tee[8], B Paten[1], AM Phillippy[2], JT Simpson[11,12], NJ Loman[3,*], M Loose[9,*]

Affiliations:

1. UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA
2. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA
3. Institute of Microbiology and Infection, University of Birmingham, Birmingham B15 2TT, UK
4. Department of Human Genetics, University of Utah, Salt Lake City, UT, USA
5. USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, UT, USA
6. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
7. Michael Smith Laboratories and Djavad Mowafaghian Centre for Brain Health, University of British Columbia, Vancouver, Canada
8. Surgical Research Laboratory, Institute of Cancer & Genomic Science, University of Birmingham, UK
9. DeepSeq, School of Life Sciences, University of Nottingham, UK
10. Norwich Medical School, University of East Anglia, Norwich, UK
11. Ontario Institute for Cancer Research, Toronto M5G 0A3, Canada
12. Department of Computer Science, University of Toronto, Toronto M5S 3G4, Canada

§ These authors contributed equally to this work.
* Authors for correspondence n.j.loman@bham.ac.uk, matt.loose@nottingham.ac.uk

Editors summary

A human genome is sequenced and assembled *de novo* using a pocket-sized nanopore device.

Abstract

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-based alignment enabled detection of large structural variants and epigenetic modifications. *De novo* assembly of nanopore reads alone yielded a contiguous assembly (NG50 ~3 Mb). Next, we developed a protocol to generate ultra-long reads (N50 > 100kb, up to 882 kb). Incorporating an additional 5×-coverage of these data more than doubled the assembly contiguity (NG50 ~6.4 Mb). The final assembled genome was 2,867 million bases in size, covering 85.8% of the reference. Assembly accuracy, after incorporating complementary short-read sequencing data, exceeded 99.8%. Ultra-long reads enabled assembly and phasing of the 4 Mb major histocompatibility complex (MHC) locus in its entirety, measurement of telomere repeat length and closure of gaps in the reference human genome assembly GRCh38.

The human genome is a yardstick to assess performance of DNA sequencing instruments [1–5]. Despite improvements in sequencing technology, assembling human genomes with high accuracy and completeness remains challenging. This is due to size (~3.1 Gb), heterozygosity, regions of GC% bias, diverse families repeat families and segmental duplications (up to 1.7 Mbp in size) that make up at least 50% of the genome [6]. Even more challenging are the pericentromeric, centromeric and acrocentric short arms of chromosomes, which contain satellite DNA and tandem repeats of 3-10 Mb in length [7,8]. Repetitive structures pose challenges for *de novo* assembly using "short read" sequencing technologies e.g. Illumina. Such data, whilst enabling highly accurate genotyping in non-repetitive regions, do not provide contiguous *de novo* assemblies. This limits the ability to reconstruct repetitive sequences, detect complex structural variation, and fully characterize the human genome.

Single-molecule sequencing, e.g. Pacific Biosciences (PacBio), can produce read lengths of 10 kb or more, which makes *de novo* human genome assembly more tractable [9]. However, single molecule sequencing reads have significantly higher error rates compared with Illumina sequencing. This has necessitated development of *de novo* assembly algorithms and the use of long noisy data in conjunction with accurate short reads to produce high quality reference genomes [10]. In May 2014, the MinION nanopore sequencer was made available to early access users [11]. Initially, the MinION nanopore sequencer was used to sequence and assemble microbial genomes or PCR products [12,13,14] because the output was limited to 500Mb

- 2 Gb of sequence bases. More recently, assemblies of eukaryotic genomes including yeasts, fungi and *C. elegans* have been reported [15–17].

Recent improvements to the protein pore (a laboratory-evolved *E. coli* CsgG mutant named R9.4), library preparation techniques (1D ligation and 1D rapid), sequencing speed (450 bases/s), and control software have increased throughput, so we hypothesized that WGS of a human genome might be feasible using only a MinION nanopore sequencer[17–19].

We report sequencing and assembly of a reference human genome for GM12878 from the Utah/CEPH pedigree, using MinION R9.4 1D chemistry, including ultra-long reads up to 882 kb in length. GM12878 has been sequenced on a wide variety of platforms, and has well-validated variation call sets, which enabled us to benchmark our results[20].

RESULTS

Sequencing dataset

Five laboratories collaborated to sequence DNA from the GM12878 human cell line. DNA was sequenced directly (avoiding PCR) thus preserving epigenetic modifications such as DNA methylation. 39 MinION flowcells generated 14,183,584 basecalled reads containing 91,240,120,433 bases with a read N50 of 10,589 bp (Supplementary Tables 1-4). Ultra-long reads were produced using 14 additional flowcells. Read lengths were longer when the input DNA was freshly extracted from cells compared with using Coriell supplied DNA (Figure 1A). Average yield per flow cell (2.3 Gb) was unrelated to DNA preparation methods (Figure 1B). 94.15% of reads had at least one alignment to the human reference (GRCh38) and 74.49% had a single alignment over 90% of their length. Median coverage depth was 26-fold and 96.95% (3.01/3.10 Gbp) bases of the reference were covered by at least one read (Figure 1C). The median identity of reads was 84.06% (82.73% mean, 5.37% standard deviation). No length-bias was observed in the error rate with the MinION (Figure 1D).

Base-caller evaluation

The base-calling algorithm used to decode raw ionic current signal can affect sequence calls. To analyze this effect we used reads mapping to chromosome 20 and compared base-calling with Metrichor (an LSTM-RNN base-caller) and Scrappie, an open-source transducer neural network (Online Methods). Of note, we observed that a fraction of the Scrappie output (4.7% reads, 14% bases) was composed of low-complexity sequence (Supplementary Figure 1), which we removed before downstream analysis.

To assess read accuracy we realigned reads from each base-caller using a trained alignment model [21]. Alignments generated by BWA-MEM were chained such that each read has at most one maximal alignment to the reference sequence (scored by length). The chained alignments were used to derive the maximum likelihood estimate of alignment model parameters [22], and the trained model used to realign the reads. The median identity after re-alignment for Metrichor was 82.43%

and for Scrappie 86.05%. We observed a purine-to-purine substitution bias in chained alignments where the model was not used (Supplementary Figure 2). The alignments produced by the trained model showed an improved substitution error rate, decreasing the overall transversion rate, but transition errors remained dominant.

To measure potential bias at the *k*-mer level, we compared counts of 5-mers in reads derived from chromosome 20. In Metrichor reads, the most underrepresented 5-mers were A/T-rich homopolymers. The most over-represented *k*-mers were G/C-rich and non-homopolymeric (Supplementary Table 5). By contrast, Scrappie showed no underrepresentation of homopolymeric 5-mers and had a slight over representation of A/T homopolymers. Overall, Scrappie showed the lowest *k*-mer representation bias (Figure 1E). The improved homopolymer resolution of Scrappie was confirmed by inspection of chromosome 20 homopolymer calls versus the human reference (Figure 1F, Supplementary Figure 3, Online Methods) [23]. Despite this reduced bias, whole-genome assembly and analyses proceeded with Metrichor reads, since Scrappie was still in early development at the time of writing.

De novo assembly of nanopore reads

We carried out a *de novo* assembly of the 30× dataset with Canu [24] (Table 1). This assembly comprised 2,886 contigs with an NG50 contig size of 3 Mbp (NG50, the longest contig such that contigs of this length or greater sum to at least half the haploid genome size). The identity to GRCh38 was estimated as 95.20%. Canu was fourfold slower on the Nanopore data compared to a random subset of equivalent coverage of PacBio requiring ~62K CPU hours (Online Methods). The time taken by Canu increased when the input was nanopore sequence reads because of systematic error in the raw sequencing data leading to reduced accuracy of the Canu-corrected reads, an intermediate output of the assembler. Corrected PacBio reads are typically >99% identical to the reference, our reads averaged 92% identity to the reference after correction (Supplementary Figure 1B).

We aligned assembled contigs to the GRCh38 reference and found that our assembly was in agreement with previous GM12878 assemblies (Supplementary Figure 4) [25]. The number of structural differences (899) that we identified between GM12878 and GRCh38 was similar to a previously published PacBio assembly of GM12878 (692) and comparable to other human genome assemblies [5,24], but with a higher than expected number of deletions, due to consistent truncation of homopolymer and low-complexity regions (Supplementary Figure 5, Supplementary Table 6). Consensus identity of our assembly with GRCh38 was estimated to be 95.20% (Table 1). However, GRCh38 is a composite of multiple human haplotypes, so this is a lower bound on accuracy. Comparisons with independent Illumina data from GM12878 yielded a higher accuracy estimate of 95.74%.

Despite the low consensus accuracy, contiguity was good. For example, the assembly included a single ~3 Mbp contig that includes all class I HLA genes from the major histocompatibility complex (MHC) region on chromosome 6, a region notoriously difficult to assemble using short reads. The more repetitive class II HLA gene locus was fragmented but most genes were present in a single contig.

Genome polishing

To improve the accuracy of our assembly we mapped previously generated whole-genome Illumina data (SRA:ERP001229) to each contig using BWA-MEM and corrected errors using Pilon. This improved the estimated accuracy of our assembly to 99.29% vs GRCh8 and 99.88% vs independent GM12878 sequencing (Table 1, Supplementary Figure 6) [26]. This estimate is a lower bound as true heterozygous variants and erroneously mapped sequences decrease identity. Recent PacBio assemblies of mammalian genomes that were assembled de novo and polished with Illumina data exceed 99.95% [9,27]. Pilon cannot polish regions that have ambiguous short-read mappings i.e in repeats. We also compared the accuracy of our polished assembly in regions with expected coverage versus those that had low-quality mappings (either lower coverage or higher than expected coverage with low mapping quality**)** versus GRCh38. When compared to GRCh38, accuracy in well-covered regions increased to 99.32% from the overall accuracy of 99.29% while the poorly covered region accuracy drops to 98.65%.

For further evaluation of our assembly, comparative annotation was carried out before and after polishing (Online Methods, Supplementary Table 7). 58,338 genes (19,436 coding / 96.4% of genes in GENCODE V24 / 98.2% of coding genes) were identified representing 179,038 transcripts in the polished assembly. Reflecting the assembly's high contiguity, only 857 (0.1%) of genes were found on two or more contigs.

Alternative approaches to improve assembly accuracy using different base-callers and exploiting the ionic current signal were attempted on a subset of reads from chromosome 20. Assembly consensus improvement using raw output is commonly used when assembling single-molecule data. To quantify the effect of base-calling on the assembly, the read sets from Metrichor and Scrappie were re-assembled with the same Canu parameters used for the whole-genome dataset. Whilst all assemblies had similar contiguity, using Scrappie reads improved accuracy from 95.74% to 97.80%. Signal-level polishing of Scrappie assembled reads using nanopolish increased accuracy to 99.44%, and polishing with Illumina data brought the accuracy up to 99.96% (Table 1).

Analysis of sequences not in the assembly

To investigate sequences omitted from the primary genome analysis we assessed 1,425 contigs filtered from Canu due to low coverage, or contigs that were single reads with many shorter reads within (26 Mbp), or corrected reads not incorporated into contigs (10.4 Gbp). Most sequences represented repeat classes e.g. LINEs, SINEs etc (Supplementary Figure 7), observed in similar proportion in the primary assembly, with the exception of satellite DNAs known to be enriched in human centromeric regions. These satellites were enriched 2.93× in the unassembled data and 7.9× in the Canu-filtered contigs. We identified 56 assembled contigs containing centromere repeat sequences specific to each of the 22 autosomes and X chromosome. The largest assembled satellite in these contigs is a 94 kbp tandem repeat specific to centromere 15 (D15Z1, tig00007244).

SNP and SV genotyping

Using SVTyper [28] and Platinum Illumina WGS alignments, we genotyped 2,414 GM12878 SVs, which were previously identified using LUMPY and validated with PacBio and/or Moleculo reads [29]. We looked for SVs using alignments of our nanopore reads from the 30x-coverage dataset and a modified version of SVTyper (Online Methods). We measured the concordance of genotypes at each site in the Illumina and nanopore-derived data deducing the sensitivity of SV genotyping as a function of nanopore sequencing depth (Figure 2A). Using all 39 flowcells, nanopore data recovered 91% of high-confidence SVs with a false-positive rate of 6% (Online Methods). Illumina and nanopore genotypes agreed at 81% of heterozygous sites and 90% of homozygous alternate sites. Genotyping heterozygous SVs using nanopore alignments is limited when homopolymer stretches occur at the breakpoints of these variants (Supplementary Figure 8A). We determined Illumina, nanopore, and PacBio genotype concordance at a set of 2,192 deletions common to our high-confidence set and a genotyped SV call set derived from PacBio sequencing of NA12878 [5,30] (Online Methods). PacBio and Illumina genotypes agree at 94% of heterozygous and 79% of homozygous alternate deletions; nanopore and

Illumina genotypes agree at 90% of heterozygous and 90% of homozygous alternate sites; nanopore and PacBio genotypes agree at 91% of heterozygous and 76% of homozygous alternate sites (Online Methods). Nearly a quarter (44) of the homozygous alternate sites at which PacBio and Illumina genotypes disagree overlap SINEs or LINEs. By manual inspection in IGV [31], sequencing reads are spuriously aligned at these loci and likely drive the discrepancy in predicted genotypes (Supplementary Figure 8B).

We evaluated nanopore data for calling genotypes at known single nucleotide polymorphisms (SNPs) using the ionic current by calling genotypes at non-singleton SNPs on chromosome 20 from phase 3 of the 1000 Genomes [32] (Online Methods) and comparing these calls to Illumina Platinum Genome calls (Figure 2B). 99.16% of genotype calls are correct (778,412 out of 784,998 sites). This result is dominated by the large number of homozygous reference sites. If we assess accuracy by the fraction of correctly called variant sites (heterozygous or homozygous non-reference), the accuracy of our caller is 91.40% (50,814 out of 55,595), with the predominant error mis-calling sites labelled homozygous in the reference as heterozygous (3,217 errors). Genotype accuracy when only considering sites annotated as variants in the platinum call set, is 94.83% (50,814 correct out of 53,582).

Detection of epigenetic 5-methyl cytosine modification

Changes in the ionic current when modified and unmodified bases pass through the minION nanopores enable detection of epigenetic marks [33,34]. We used nanopolish and SignalAlign to map 5-methyl cytosine at CpG dinucleotides as detected in our sequencing reads against chromosome 20 of the GRCh38 reference [35,36]. Nanopolish outputs a frequency of reads calling a methylated cytosine and SignalAlign outputs a marginal probability of methylation summed over reads. We compared the output of both methods to published bisulfite sequencing data from the same DNA region (ENCFF835NTC). Good concordance of our data with the published bisulfite sequencing was observed; the r-values for nanopolish and SignalAlign were 0.895 and 0.779 respectively (Figure 3, Supplementary Figure 9,10).

Ultra-long reads improve phasing and assembly contiguity

We modelled the contribution of read length to assembly quality (Online Methods) predicting ultra-long read datasets (N50 >100 kb) would substantially improve assembly contiguity (Figure 4A). We developed a method to produce ultra-long reads by saturating the Oxford Nanopore Rapid Kit with high molecular weight DNA. In so doing we generated an additional 5× coverage (Supplementary Figure 11). Two additional standard protocol flowcells generated a further 2× coverage and were used as controls for software and base-caller versions. The N50 read length of the ultra-long dataset was 99.7 kb (Figure 4B). Reads were impossible to align efficiently at first, because aligners algorithms are optimized for short reads. Further, CIGAR strings generated by ultra-long reads do not fit in the BAM specification, necessitating the use of SAM or CRAM only (https://github.com/samtools/hts-specs/issues/40). Instead, we used GraphMap [37] to align ultra-long reads to GRCh38, which took >25K CPU hours (Supplementary Table 8). Software optimized for long reads including NGM-LR [38] and Minimap2 [39] were faster: Minimap2 took 60 CPU hours. More than 80% of bases were in sequences aligned over 90% of their length with GraphMap and more than 60% with minimap2. Median alignment identity was 81% (83 with minimap2), slightly lower than observed for the control flowcells (83.46%/84.64%) and the original dataset (83.11%/84.32%). The longest full-length mapped read in the dataset (aligned with GraphMap) was 882 kb, corresponding to a reference span of 993 kb.

The addition of 5x coverage ultra-long reads more than doubled the previous assembly NG50 to 6.4 Mbp and resolved the MHC locus into a single contig (Figure 4C). In comparison, a 50× PacBio GM12878 dataset with average read length of 4.5 kb assembled with an NG50 contig size of 0.9 Mbp [5]. Newer PacBio assemblies of a human haploid cell line, with mean read lengths greater than 10 kb, have reached contig NG50s exceeding 20 Mbp at 60× coverage [25]. We subsampled this dataset to an equivalent depth as ours (35×) and assembled, resulting in an NG50 of 5.7 Mbp, with the MHC split into >2 contigs. The PacBio assembly is less contiguous, despite a higher average read length and simplified haploid genome.

In addition to assembling the MHC into a single contig, the ultra-long MinION reads enabled the contiguous MHC to be phased. Due to the limited depth of nanopore reads, heterozygous SNPs were called using Illumina data and then phased using the ultra-long nanopore reads to generate two pseudo-haplotypes (Online Methods), from which MHC typing was performed using the approach of Dilthey et al. [40] (Figure 5A). Some gaps were introduced during haplotig assembly due to low phased-read coverage e.g. *HLA-DRB3* was left unassembled on haplotype A, but apart from one *HLA-DRB1* allele, sample HLA types were recovered almost perfectly with an edit distance between 0–1 for true allele versus called allele (Supplementary Table 9). Analysis of parental (GM12891, GM12892) HLA types confirmed the absence of switch errors between the classical HLA typing genes. To our knowledge, this is the first time the MHC has been assembled and phased over its full length in a diploid human genome.

Already published single-molecule human genome assemblies contain multiple contigs that span the MHC [5,41,42] and have not attempted phasing. Instead, MHC surveys have focused on homozygous cell lines [43].

Ultra-long reads close gaps in the human reference genome

Large (>50 kilobase) bridged scaffold gaps remain unresolved in the reference human genome assembly (GRCh38). These breaks in the assembly span tandem repeats and/or long tracts of segmental duplications [44]. Using sequence from our *de novo* assembled contigs we were able to close 12 gaps, each of which was more than 50 kilobases in the reference genome. We then looked for individual ultra-long reads that spanned gaps, and matched the assembly predicted sequence closure for each region (Supplementary Table 10).

The gap closures enabled us to identify 83,980 bp of previously unknown euchromatic sequence. For example, an unresolved 50 kbp scaffold gap on Xq24 marks the site of a human specific tandem repeat that contains a testis/cancer gene family, known as CT47 [45,46]. This entire region is spanned by a single contig in our final assembly (tig00002632). Inspection of this contig using HMM profile modelling of an individual repeat unit containing the *CT47* gene (GRCh38 chrX:120932333-120938697) suggests that there is an array of 8 tandem copies of the CT47 repeat (Figure 5B). In support of this finding, we identified three ultra-long reads that together traverse the entire tandem array (Figure 5B); two reads provide evidence for an array of eight repeat copies and one read supports six copies, suggesting heterozygosity.

Telomere repeat lengths

FISH estimates and direct cloning of telomeric DNAs suggests that telomere repeats (TTAGGG) extend for multiple kilobases at the ends of each chromosome [47,48]. Using HMM profile modeling of the published telomere tract of repeats (M19947.1) we identified 140 ultra-long reads that contain the TTAGGG tandem repeat (Supplementary Table 11). Sequences next to human telomeres are enriched in intra- and inter-chromosomal segmental duplications, which makes it difficult to map ultra-long reads directly to the chromosome assemblies. However, we were able to map 17/140 ultra-long reads to specific chromosome subtelomeric regions. We analysed the mapped regions by identifying the junction or the start of the telomeric array on 17 ultra long reads, and annotating all TTAGGG-repeat sequences to the end of the read to estimate telomeric repeat length. For example, two reads that only mapped to chromosome 21q indicate that there are 9,108 bps of telomeric repeats. Overall, we find evidence for telomeric arrays that span 2 kb-11 kb within 14 subtelomeric regions for GM12878 (Figure 5C,D, Supplementary Table 11).

Discussion

We report sequencing and assembly of a human genome with 99.88% accuracy and an NG50 of 6.4Mb using unamplified DNA and nanopore reads followed by short-read consensus improvement. At 30x coverage we have produced the most contiguous assembly of a human genome to date, using only a single sequencing technology and the Canu assembler [23] . Consistent with the view that the underlying ionic raw current contains additional information, signal based polishing [14] improved the assembly accuracy to 99.44%. Finally, we report that combining signal based polishing and short-read (Illumina) correction [26] gives an assembly accuracy of 99.96%, which is comparable to metrics for other mammalian genomes [9].

Here we report that read lengths produced by the MinION nanopore sequencer are dependent on the input fragment length. We find that careful preparation of DNA in solution using classical extraction and purification methods can yield extremely long reads. The longest read lengths were achieved using the transposase based rapid library kit in conjunction with methods of DNA extraction designed to mitigate shearing. We produced 5× coverage with ultra-long reads, and used this dataset to augment our initial assembly. The final 35× coverage assembly has an NG50 of 6.4Mb. Based on modelling we predict that 30× of ultra-long reads alone would result in an assembly with a contig NG50 in excess of 40 Mb, approaching the contiguity of the current human reference (Figure 4C). We posit that there may be no intrinsic read length limit for pore-based sequencers, other than from physical forces that lead to DNA fragmentation in solution. As such, there is scope to further improve the read length results obtained here, perhaps through solid phase DNA extraction and library preparation techniques such as agar encasement.

The increased single molecule read length that we report here, obtained using a MinION nanopore sequencer, enabled us to analyze regions of the human genome that were previously intractable with state-of-the-art sequencing methods. For example, we were able to phase megabase regions of the human genome in single contigs, to more accurately estimate telomere lengths, and to resolve complex repeat regions. Phasing of 4-5 Mb scaffolds through the MHC has recently been reported using a combination of sequencing and genealogical data [49]. However, the resulting assemblies contained multiple gaps of unknown sequences. We phased the entire MHC, and reconstructed both alleles. Development of tools to automate phasing from nanopore assemblies is now needed.

We also wrote custom software/algorithms (poredb) to track the large number of reads, store each read as an individual file, and enable use of cloud-based pipelines for our analyses (Online Methods).

Our proof-of-concept demonstration of human genome sequencing using a MinION nanopore sequencer reveals the potential of this approach, but identifies specific challenges for future projects. Improvements in real-time base-calling are needed to simplify the workflow. More compact and convenient formats for storing raw and base-called data are urgently required, ideally employing a standardized, streaming compatible serialization format such as BAM/CRAM.

With ultra-long reads we found the longest reads exceeded CIGAR string limitations in the BAM format, necessitating the use of SAM or CRAM (https://github.com/samtools/hts-specs/issues/40). And, we were unable to complete an alignment of the ultra-long reads using BWA-MEM, and needed to adopt other algorithms, including GraphMap and NGM-LR, to align the reads. This required large amounts of compute time and RAM[37,38,50]. Availability of our dataset has spurred the development of Minimap2 [39], and we recommend this long read aligner for use in aligning ultra-long reads on a standard desktop computer.

Nanopore genotyping accuracy currently lags behind short-read sequencing instruments, due to a limited ability to discriminate between heterozygous and homozygous alleles which arises from error-rate and the depth of coverage in our sequencing data. We found that >99% of SNP calls were correct at homozygous reference sites, dropping to 91.4% at heterozygous and homozygous non-reference sites. Nanopore and Illumina SV sites agreed at 90% of heterozygous and homozygous sites.. These results highlight a need for structural variant genotyping tools for long, single-molecule sequencing reads. Using 1D^2 chemistry (which sequences template and complement strands of the same molecule) or modelling nanopore ionic raw current, perhaps by incorporating training data from modified DNA, could potentially produce increased read accuracy. A complementary approach would be to increase coverage.

In summary, we provide evidence that a portable, biological nanopore sequencer could be used to sequence, assemble, and provisionally analyse structural variants and detect epigenetic marks, in point-of-care human genomics applications in the future.

Acknowledgements

We acknowledge the support of Oxford Nanopore Technologies in generating this dataset, in particular R Dokos, O Hartwell, J Pugh, and C Brown. We thank M Akeson for his support and insight. We thank R Poplawski and S Thompson for technical assistance with configuring and using cloud-based file systems with millions of files on CLIMB. We thank W Timp and R Workman for generating the R9.4 methylation training data for nanopolish. We thank T Allers for assistance with PFGE. We thank A Pizarro at Amazon Web Services for hosting the human genome dataset as an Amazon Web Services Open Dataset. This study utilized the computational resources of the Biowulf system at the National Institutes of Health, Bethesda, MD (https://biowulf.nih.gov).

Contributions:

N.J.L, M.L, J.T.S, and J.R.T conceived the study, J.Q developed the long read protocol, A.D.B, M.J, M.L, H.M, S.M, T.N, J.O'G, J.Q, H.R, J.R.T and L.T prepared materials and/or performed sequencing, A.T.D, I.T.F, M.J, S.K, N.J.L, M.L, K.H.M, H.E.O, B.P, B.S.P, A.M.P, A.R.Q, A.C.R, A.R, T.A.S, J.T.S and J.R.T performed bioinformatics analysis and wrote or modified software, I.T.F, M.J, S.K, N.J.L, M.L, K.H.M, J.O'G, H.E.O, B.P, A.M.P, J.Q, A.R.Q, A.C.R, T.A.S, J.T.S, T.P.S and J.R.T wrote and edited the manuscript. All authors approved the manuscript and provided strategic oversight for the work.

Conflict of interest statement

ML, NL, JOG, JTS, JRT, and TPS were members of the MinION access program (MAP) and have received free-of-charge flowcells and kits for nanopore sequencing for this and other studies, and travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. NJL has received an honorarium to speak at an Oxford Nanopore company meeting. SK, ATD, and TAS have received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences. JTS, JOG and ML receive research funding from Oxford Nanopore Technologies.

Funding

This study was partially supported by the UK Antimicrobial Resistance Cross Council Initiative (JOG: MR/N013956/1), Rosetrees Trust (JOG: A749), the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (SK, AD, AR, AMP), the BBSRC (ML: BB/N017099/1 and BB/M020061/1), the Canadian Institutes of Health Research (JRT,TPS: #10677), Brain Canada Multi-Investigator Research Initiative Grant with Genome British Columbia the Michael Smith Foundation for Health Research and the Koerner Foundation (JRT,TPS), the Canada Research Chair in Biotechnology and Genomics-Neurobiology (TPS), the Ontario Institute for Cancer Research and the Government of Canada (JTS: OGI-129), the US National Cancer Institute (ARQ, BSP, TAS: NIH U24CA209999), the Wellcome Trust (ADB: 102732/Z/13/Z, ML: 204843/Z/16/Z), Cancer Research UK (ADB: A23923), the MRC (ADB: MR/M016587/1), the MRC Fellowship in Microbial Bioinformatics as part of CLIMB (NJL) and the NIHR Surgical Reconstruction and Microbiology Research Centre (SRMRC).

References

1. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel

DNA sequencing. *Nature* **452,** 872–876 (2008).

2. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
3. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27,** 847–850 (2009).
4. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).
5. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12,** 780–786 (2015).
6. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9,** 533 (2008).
7. Wevrick, R. & Willard, H. F. Long-range organization of tandem arrays of alpha satellite DNA at the centromeres of human chromosomes: high-frequency array-length polymorphism and meiotic stability. *Proceedings of the National Academy of Sciences* **86,** 9394–9398 (1989).
8. Eichler, E. E., Clark, R. A. & She, X. An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat. Rev. Genet.* **5,** 345–354 (2004).
9. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352,** aae0344 (2016).
10. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Rev. Genet.* **16,** 627–640 (2015).

11. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).

12. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).

13. Quick, J. *et al.* Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **16,** 114 (2015).

14. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12,** 733–735 (2015).

15. Istace, B. *et al.* de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience* **6,** 1–13 (2017).

16. Datema, E. *et al.* The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *bioRxiv* 084772 (2016). doi:10.1101/084772

17. Tyson, J. R. *et al.* Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomic rearrangements using the MinION sequencing device. *bioRxiv* 099143 (2017). doi:10.1101/099143

18. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *bioRxiv* 098913 (2017). doi:10.1101/098913

19. Jansen, H. J. *et al.* Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *bioRxiv* 101907 (2017). doi:10.1101/101907

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

21. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12,** 351–356 (2015).

22. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological sequence analysis*. (1998).

23. 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).

24. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R. & Bergman, N. H. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* (2017). doi:10.1101/gr.215087.116

25. 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).

26. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9,** e112963 (2014).

27. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.* **49,** 643–650 (2017).

28. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods* **12,** 966–968 (2015).

29. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15,** R84 (2014).

30. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

31. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

32. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

33. Wescoe, Z. L., Schreiber, J. & Akeson, M. Nanopores discriminate among five C5-cytosine variants in DNA. *J. Am. Chem. Soc.* **136,** 16582–16587 (2014).

34. Laszlo, A. H. *et al.* Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18904–18909 (2013).

35. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore sequencing. *Nat. Methods* (2017). doi:10.1038/nmeth.4189

36. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* (2017). doi:10.1038/nmeth.4184

37. Sović, I. *et al.* Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7,** 11307 (2016).

38. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using

single molecule sequencing. *bioRxiv* 169557 (2017). doi:10.1101/169557

39. Li, H. Minimap2: fast pairwise alignment for long DNA sequences. *arXiv [q-bio.GN]* (2017).
40. Dilthey, A. T. *et al.* High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. *PLoS Comput. Biol.* **12,** e1005151 (2016).
41. Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat. Commun.* **7,** 12065 (2016).
42. Seo, J.-S. *et al.* De novo assembly and phasing of a Korean human genome. *Nature* **538,** 243–247 (2016).
43. Norman, P. J. *et al.* Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Res.* **27,** 813–823 (2017).
44. Bovee, D. *et al.* Closing gaps in the human genome with fosmid resources generated from multiple individuals. *Nat. Genet.* **40,** 96–101 (2008).
45. Chen, Y.-T. *et al.* Identification of a new cancer/testis gene family, CT47, among expressed multicopy genes on the human X chromosome. *Genes Chromosomes Cancer* **45,** 392–400 (2006).
46. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12,** 351–356 (2015).
47. Moyzis, R. K. *et al.* A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.*

85, 6622–6626 (1988).

48. Kimura, M. *et al.* Measurement of telomere length by the Southern blot analysis of terminal restriction fragment lengths. *Nat. Protoc.* **5,** 1596–1607 (2010).
49. Maretty, L. *et al.* Sequencing and de novo assembly of 150 genomes from Denmark as a population reference. *Nature* (2017). doi:10.1038/nature23264
50. Jain, C., Dilthey, A., Koren, S. & Aluru, S. A fast approximate algorithm for mapping long reads to large reference databases. *Conference on Research …* (2017).

Figure Legends

Figure 1 – Summary of Dataset

a) Read Length N50s by flow cell, colored by sequencing center. Cells: DNA extracted directly from cell culture. DNA: Pre-extracted DNA purchased from Coriell. b) Total yield per flow cell grouped as A. c) Coverage (black line) of GRCh38 reference compared to a Poisson distribution. The depth of coverage of each reference position was tabulated using samtools depth and compared with a Poisson distribution with lambda = 27.4 (dashed red line). d) Alignment identity compared to alignment length. No length bias was observed, with long alignments having the same identity as short ones. e) Correlation between 5-mer counts in reads compared to expected counts in the chromosome 20 reference. f) Chromosome 20 homopolymer length versus median homopolymer base-call length measured from individual Illumina and nanopore reads (Scrappie and Metrichor). Metrichor fails to produce homopolymer runs longer than ~5 bp. Scrappie shows better correlation for longer homopolymer runs, but tends to over-call short homopolymers (between 5 and 15 bp) and under-call long homopolymers (>15 bp). Plot noise for longer homopolymers is due to fewer samples available at that length.

Figure 2 – Structural Variation and SNP Genotyping

a) Structural variant genotyping sensitivity using ONT reads. Genotypes were inferred for a set of 2,414 SVs using both Oxford Nanopore and Platinum Genomes (Illumina) alignments. Using alignments randomly subsampled to a given sequencing depth (n=3), sensitivity was calculated as the proportion of ONT-derived genotypes that were concordant with Illumina-derived genotypes. b) Confusion matrix for genotype calling evaluation. Each cell contains the number of 1000 Genome sites for a particular nanopolish/platinum genotype combination.

Figure 3 - Methylation detection using signal-based methods.

a) SignalAlign methylation probabilities compared to bisulfite sequencing frequencies at all called sites. b) Nanopolish methylation frequencies compared to bisulfite sequencing at all called sites. c) SignalAlign methylation probabilities compared to bisulfite sequencing frequencies at sites covered by at least 10 reads in the nanopore and bisulfite data sets, reads were not filtered for quality. d) Nanopolish methylation frequencies compared to bisulfite sequencing at sites covered by at least 10 reads in the nanopore and bisulfite data sets. A minimum log-likelihood threshold of 2.5 was applied to remove ambiguous reads. n = sample size, r = Pearson Correlation Coefficient.

Figure 4 – Repeat modeling and assembly

a) A model of expected NG50 contig size when correctly resolving human repeats of a given length and identity (Online Methods). The y-axis shows the expected NG50 contig size when repeats of a certain length (x-axis) or sequence identity (colored lines) can be consistently resolved. Nanopore assembly contiguity (GM12878 20×, 30×, 35×) is currently limited by low coverage of long reads and a high error rate, making repeat resolution difficult. These assemblies approximately follow the predicted assembly contiguity. The *projected* assembly contiguity using 30× of ultra-long reads (GM12878 30× ultra) exceeds 30 Mbp. A recent assembly of 65× PacBio P6 data with an NG50 of 26 Mbp is shown for comparison (CHM1 P6). b) Yield by read length ($\log_{10}$) for ligation, rapid and ultra-long rapid library preparations. c) Chromosomes plot illustrating the contiguity of the nanopore assembly boosted with ultra-long reads. Contig and alignment boundaries, not cytogenetic bands, are represented by a color switch, so regions of continuous color indicate regions of contiguous sequence. White areas indicate unmapped sequence, usually caused by N's in the reference genome. Regions of interest including the 12 50+ kb gaps in GRCh38 closed by our assembly as well as the MHC (16 Mbp) are outlined in red.

Figure 5 – Ultra Long Reads, Assembly and Telomeres

a) A 16 Mbp ultra-long read contig and associated haplotigs are shown spanning the full MHC region. MHC Class I and II regions are annotated along with various HLA genes. Below this contig, the MHC region is zoomed showing haplotype A and B coverage tracks for the phased nanopore reads. Nanopore reads were aligned back to the polished Canu contig, with colored lines indicating a high fraction of single-nucleotide discrepancies in the read pileups (as displayed by the IGV browser). The many disagreements indicate the contig is a mosaic of both haplotypes. The haplotig A and B tracks show the result of assembling each haplotype read set independently. Below this, the MHC class II region is enlarged, with haplotype A and

B raw reads aligned to their corresponding, unpolished haplotigs. The few consensus disagreements between raw reads and haplotigs indicate successful partitioning of the reads into haplotypes (Online Methods). b) An unresolved, 50 kb bridged scaffold gap on Xq24 remains in the GRCh38 assembly (adjacent to scaffolds AC008162.3 and AL670379.17, shown in green). This gap spans a ~4.6 kb tandem repeat containing cancer/testis gene family 4 (CT47). This gap is closed by assembly (contig: tig00002632) and has 8 tandem copies of the repeat, validated by alignment of 100 kb+ ultra-long reads also containing 8 copies of the repeat (light blue with read name identifiers). One read has only 6 repeats, suggesting the tandem repeated units are variable between homologous chromosomes. c) Ultra-long reads can predict telomere length. Two 100 kb+ reads that map to the subtelomeric region of the chromosome 21 q-arm, each containing 4.9-9.1 kb of the telomeric (TTAGGG_ repeat). d) Telomere length estimates showing variable lengths between non-homologous chromosomes.

Tables

Table 1 – Summary Assembly Statistics

Assembly	Polishing	Contigs	# Bases (Mbp)	Max Contig (kb)	NG50 (kb)	GRCh38 Identity	GM12878 Identity
WGS Metrichor	N/A	2,886	2,646.01	27,160	2,964	95.20%	95.74%
	Pilon x2		2,763.18	28,413	3,206	99.29%	99.88%
Chr 20 Metrichor	N/A	85	57.83	7,393	3,047	94.90%	95.50%
	Nanopolish		60.35	7,667	5,394	98.84%	99.24%
	Pilon x2		60.58	7,680	5,423	99.33%	99.89%
	Nano + Pilon x2		60.76	7,698	5,435	99.64%	99.95%
Chr 20 Scrappie	N/A	74	59.39	8,415	2,643	97.43%	97.80%
	Nanopolish		60.15	8,521	2,681	99.12%	99.44%
	Pilon x2		60.36	8,541	2,691	99.64%	99.95%
	Nano + Pilon x2		60.34	8,545	2,691	99.70%	99.96%

Summary of assembly statistics. Whole genome assembly (WGA) was performed with reads base called by Metrichor. Chromosome 20 was assembled with reads produced by Metrichor and Scrappie. All datasets contained 30× coverage of the genome/chromosome. The GRCh38 identities were computed based on 1-1 alignments to the GRCh38 reference including alt sites. A GM12878 reference was estimated using an independent sequencing dataset [20] (Methods).

Online Methods

Human DNA

Human genomic DNA from the GM12878 human cell line (CEPH/Utah pedigree) was either purchased from Coriell as DNA (cat no NA12878) or extracted from the cultured cell line also purchased from Coriell (cat no GM12878). Cell culture was performed using EBV transformed B lymphocyte culture from the GM12878 cell line in RPMI-1640 media with 2mM L-glutamine and 15% fetal bovine serum at 37°C.

QIAGEN DNA extraction

DNA was extracted from cells using the QIAamp DNA mini kit (Qiagen). 5×10^6 cells were spun at 300× *g* for 5 minutes to pellet. The cells were resuspended in 200 µl PBS and DNA was extracted according to the manufacturer's instructions. DNA quality was assessed by running 1 µl on a genomic ScreenTape on the TapeStation 2200 (Agilent) to ensure a DNA Integrity Number (DIN) >7 (Value for NA12878 was 9.3). Concentration of DNA was assessed using the dsDNA HS assay on a Qubit fluorometer (Thermo Fisher).

Library preparation (SQK-LSK108 1D ligation genomic DNA)

1.5–2.5 µg human genomic DNA was sheared in a Covaris g-TUBE centrifuged at 5000–6000 rpm in an Eppendorf 5424 (or equivalent) centrifuge for 2× 1 minute, inverting the tube between centrifugation steps.

DNA repair (NEBNext FFPE DNA Repair Mix, NEB M6630) was performed on purchased DNA but not on freshly extracted DNA. 8.5 µl NFW, 6.5 µl FFPE Repair Buffer and 2 µl FFPE DNA Repair Mix were added to the 46 µl sheared DNA. The mixture was incubated for 15 mins at 20 °C, cleaned up using a 0.4× volume of AMPure XP beads (62 µl), incubated at room temperature with gentle mixing for 5 minutes, washed twice with 200 µl fresh 70% ethanol, pellet allowed to dry for 2 mins and DNA eluted in 46 µl NFW or EB (10 mM Tris pH 8.0). A 1 µl aliquot was quantified by fluorometry (Qubit) to ensure ≥1 µg DNA was retained.

End repair and dA-tailing (NEBNext Ultra II End-Repair / dA-tailing Module) was then performed by adding 7 µl Ultra II End-Prep buffer, 3 µl Ultra II End-Prep enzyme mix, and 5 µl NFW. The mixture was incubated at 20 °C for 10 minutes and 65 °C for 10 minutes. A 1× volume (60 µl) AMPure XP clean-up was performed and the DNA was eluted in 31 µl NFW. A 1 µl aliquot was quantified by fluorometry (Qubit) to ensure ≥700 ng DNA was retained.

Ligation was then performed by adding 20 µl Adapter Mix (SQK-LSK108 Ligation Sequencing Kit 1D, Oxford Nanopore Technologies [ONT]) and 50 µl NEB Blunt/TA Master Mix (NEB, cat no M0367) to the 30 µl dA-tailed DNA, mixing gently and incubating at room temperature for 10 minutes.

The adapter-ligated DNA was cleaned-up by adding a 0.4× volume (40 µl) of AMPure XP beads, incubating for 5 minutes at room temperature and resuspending the pellet twice in 140 µl ABB (SQK-LSK108). The purified-ligated DNA was resuspend by adding 25 µl ELB (SQK-LSK108) and resuspending the beads, incubating at room temperature for 10 minutes, pelleting the beads again and transferring the supernatant (pre-sequencing mix or PSM) to a new tube. A 1 µl aliquot was quantified by fluorometry (Qubit) to ensure ≥ 500 ng DNA was retained.

Sambrook and Russell DNA extraction

This protocol was modified from Chapter 6 protocol 1 of Sambrook and Russell [51]. 5×10^7 cells were spun at 4500× *g* for 10 minutes to pellet. The cells were resuspended by pipette mixing in 100 µl PBS. 10ml TLB was added (10mM Tris-Cl pH 8.0, 25mM EDTA pH 8.0, 0.5% (w/v) SDS, 20 µg/ml Qiagen RNase A), vortexed at full speed for 5 seconds and incubated at 37 °C for 1 hr. 50 µl Proteinase K (Qiagen) was added and mixed by slow inversion 10 times followed by 3 hrs at 50 °C with gentle mixing every 1 hour. The lysate was phenol purified using 10 ml buffer saturated phenol using phase-lock gel falcon tubes, followed by phenol:chloroform (1:1), The DNA was precipitated by the addition of 4 ml 5 M ammonium acetate and 30 ml ice-cold ethanol. DNA was recovered with a glass hook followed by washing twice in 70% ethanol. After spinning down at 10,000g, ethanol was removed followed by 10 mins drying at 40 °C. 150 µl EB was added to the DNA and left at 4 °C overnight to resuspend.

Library preparation (SQK-RAD002 genomic DNA)

To obtain ultra-long reads, the standard RAD002 protocol (SQK-RAD002 Rapid Sequencing Kit, ONT) for genomic DNA was modified as follows. 16 µl of DNA from the Sambrook extraction at approximately 1 µg/µl, manipulated with a cut-off P20 pipette tip, was placed in a 0.2 ml PCR tube, with 1 µl removed to confirm quantification value. 5 µl FRM was added and mixed slowly 10 times by gentle pipetting with a cut-off pipette tip moving only 12 µl. After mixing, the sample was incubated at 30 °C for 1 minute followed by 75 °C for 1 minute on a thermocycler. After this, 1 µl RAD and 1 µl Blunt/TA ligase was added with slow mixing by pipetting using a cut-off tip moving only 14 µl 10 times. The library was then incubated at room temperature for 30 minutes to allow ligation of Rapid Adapters (RAD). To load the library, 25.5 µl RBF was mixed with 27.5 µl NFW and this was added to the library. Using a P100 cut-off tip set to 75 µl, this library was mixed by pipetting slowly 5 times. This extremely viscous sample was loaded onto the “spot on” port and entered the flow cell by capillary action. The standard loading beads were omitted from this protocol due to excessive clumping when mixed with the viscous library.

MinION sequencing

MinION sequencing was performed as per manufacturer's guidelines using R9/R9.4 flowcells (FLO-MIN105/FLO-MIN106, ONT). MinION sequencing was controlled using Oxford Nanopore Technologies MinKNOW software. The specific versions of the software used varied from run to run but can be determined by inspection of fast5 files from the dataset. Reads from all sites were copied off to a volume mounted on a CLIMB virtual server (http://www.climb.ac.uk) where metadata was extracted using poredb (https://github.com/nickloman/poredb) and base-calling performed using Metrichor (predominantly workflow ID 1200, although previous versions were used early on in the project) (http://www.metrichor.com). We note that basecalling in Metrichor has now been superseded by Albacore and is no longer available. Scrappie (https://github.com/nanoporetech/scrappie) was used for the chr20 comparisons using reads previously identified as being from this chromosome after mapping the Metrichor reads. Albacore 0.8.4 (available from the Oxford Nanopore Technologies user community) was used for the ultralong read set, as this software became the recommended basecaller for nanopore reads in March 2017. Given the rapid development of upgrades to basecaller software we expect to periodically re-basecall these data and make the latest results available to the community through the Amazon Open Data site.

Modified MinION running scripts

In a number of instances, MinION sequencing control was shifted to customized MinKNOW scripts. These scripts provided enhanced pore utilisation/data yields during sequencing, and operated by monitoring and adjusting flowcell bias-voltage (-180mV to -250mV), and used an event yield dependent (70% of initial hour in each segment) initiation of active pore channel assignment via re-muxing. More detailed information on these scripts can be found on the Oxford Nanopore Technologies user community. In addition, a patch for all files required to modify MinION running scripts compatible with MinKNOW 1.3.23 only is available (Supplementary Code 1).

Live run monitoring

To assist in choosing when to switch from a standard run script to a modified run protocol, a subset of runs were monitored with the assistance of the minControl tool, an alpha component of the minoTour suite of minION run and analysis tools (https://github.com/minoTour/minoTour). minControl collects metrics about a run directly from the grouper software, which runs behind the standard ONT MinKNOW interface. minControl provides a historical log of yield measured in events from a flowcell enabling estimations of yield and the decay rate associated with loss of sequencing pores over time. MinKNOW yield is currently measured in events and is scaled by approximately 1.7 to estimate yield in bases.

Assembly

All “NG” statistics were computed using a genome size of 3,098,794,149 bp (3.1 Gbp), the size of GRCh38 excluding alt sites.

Canu v1.4 (+11 commits) r8006 (4a7090bd17c914f5c21bacbebf4add163e492d54) was used to assemble the initial 20-fold coverage dataset:

```
canu -p asm -d asm genomeSize=3.1g gridOptionsJobName=na12878nano
"gridOptions=--time 72:00:00 --partition norm" -nanopore-raw
rel2*.fastq.gz corMinCoverage=0 corMaxEvidenceErate=0.22
errorRate=0.045
```

These are the suggested low-coverage parameters from the Canu documentation, but with a decreased maximum evidence error rate. This specific parameter was decreased to reduced memory requirements after it was determined that the MinHash overlapping algorithm was under-estimating error rates due to systematic error in the reads. Counterintuitively, this systematic error makes two reads look more similar than they are, because they share more *k*-mers than expected under a random model. Manually decreasing the maximum overlap error rate threshold adjusted for this bias. The assembly took 40K CPU hours (25K to correct and 15K to assemble). This is about twofold slower than a comparable PacBio dataset, mostly due to the higher noise and errors in the nanopore reads.

The same version of Canu was also used to assemble the 30-fold dataset:

```
canu -p asm -d asm genomeSize=3.1g gridOptionsJobName=na12878nano
"gridOptions=--time 72:00:00 --partition norm" -nanopore-raw
rel3*.fastq.gz corMinCoverage=0 corMaxEvidenceErate=0.22
errorRate=0.045 "corMhapOptions=--threshold 0.8 --num-hashes 512 --
ordered-sketch-size 1000 --ordered-kmer-size 14"
```

For this larger dataset, overlapping was again tweaked by reducing the number of hashes used and increasing the minimum overlap identity threshold. This has the effect of lowering sensitivity to further compensate for the bias in the input reads. This assembly required 62K CPU hours (29K to correct, 33K to assemble) and a peak of 120 Gbp of memory, which is about fourfold slower than a comparable PacBio dataset. The assembly ran on a cluster comprised of a mix of 48-thread dual-socket Intel E5-2680 v3 @ 2.50GHz CPUs with 128 Gbp of memory and 8-thread dual-socket Intel CPU E5-2698 v4 @ 2.20GHz CPUs with 1024 Gbp of memory.

The combined dataset incorporating an additional 5× coverage of ultra-long reads was assembled with an updated version of Canu v1.4 (+125 commits) r8120:

```
canu -p asm -d asm genomeSize=3.1g gridOptionsJobName=na12878nano
"gridOptions=--time 72:00:00 --partition norm" -nanopore-raw
rel3*.fastq.gz -nanopore-raw rel4*.fastq.gz "corMhapOptions=--
threshold 0.8 --num-hashes 512 --ordered-sketch-size 1000 --ordered-
kmer-size 14” batOptions=”-dg 3 -db 3 -dr 1 -el 2000 -nofilter
suspicious-lopsided”
```

This assembly required 151K CPU hours (15K to correct, 86K to trim, and 50K to assemble) and a peak of 112 Gbp of memory. These high runtimes are a consequence of the ultra-long reads. In particular, the current Canu trimming algorithm was not designed for reads of this extreme length and high error rate after correction and the algorithms used are not optimal.

Assembly contiguity modeling

Expected assembly contiguity was modeled on repeat tracks downloaded from the UCSC genome browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/).

For a given repeat identity (0%, 90%, 95%, 98%, 99%, and 99.5%), all repeats with a lower identity estimate (genomicSuperDups and chainSelf) were filtered and overlapping repeats were merged. Gaps in the reference were also considered as repeats. To compute the maximum repeat length likely to be spanned by a given sequence distribution, the probability of an unspanned repeat of a fixed length was estimated for all lengths between 1 and 100 kbp in steps of 1 kbp using an equation from http://data-science-sequencing.github.io/lectures/lecture7/ [52–54]:

$$P(at\ least\ one\ repeat\ is\ unbridged)\ \leq \left(e^{-2c+\left(\frac{2N}{G}\right)}\right)\left(\sum_{i=1}^{L-2} \quad a_i\ e^{\frac{2i}{G}}\right)$$

where $G$is the genome size, $L$is the read length, $a_i$is the number of repeats of length $1 \leq i\ \leq L-2$, $N$is the number of reads $\geq L$, and $c$is the coverage in reads $\geq L$. We used the distribution of all repeats for $a_i$and plotted the shortest repeat length such that $P(at\ least\ one\ repeat\ is\ unbridged)\ > 0.05$ for real sequencing length distributions both nanopore and PacBio sequencing runs. Assemblies of the data were plotted at their predicted spanned read length on the x-axis and NG50 on the y-axis for comparison with the model. A 30× run of ultra-long coverage was simulated from the 5× dataset by repeating each ultra-long read six times.

Assembly validation and structural variant analysis

Assemblies were aligned using MUMmer v3.23 with parameters "-l 20 -c 500 -maxmatch" for the raw assemblies and "-l 100 -c 500 -maxmatch" for the polished assemblies. Output was processed with dnadiff to report average 1-to-1 alignment identity. The MUMmer coords file was converted to a tiling using the scripts from Berlin *et al.* [55] with the command:

```
python convertToTiling.py 10000 90 100000
```

and drawn using the coloredChromosomes package [56]. Since the reference is a composite of human genomes and there are true variations between the reference and NA12878, we also computed a reference-free estimate of identity. A 30-fold subset of the Genome In a Bottle Illumina dataset for NA12878 [20] was downloaded from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NIST_NA12878_HG001_HiSeq_300x/RMNISTHS_30xdownsample.bam. Samtools fastq was used to extract fastq paired-end data for the full dataset and for the reads mapping to chromosome 20. The reads were aligned to the whole genome assembly and chromosome 20 assemblies with BWA-MEM 0.7.12-r1039. BWA-MEM is a component of the BWA package and was chosen due to its speed and ubiquitous use in sequence mapping and analysis pipelines. Aside from the difficulties of mapping the ultra-long reads unique to this work, any other mapper could be used instead. Variants were identified using FreeBayes v1.0.2 [57] , a widely-used method originally developed for short-read sequencing but also applicable to long-reads, with the command:

```
freebayes -C 2 -0 -O -q 20 -z 0.10 -E 0 -X -u -p 2 -F 0.6 -b
alignments.bam -v asm.bayes.vcf -f asm.fasta
```

The length of all variants was summed and the total number of bases with at least 3× coverage was summed using samtools depth. QV was computed as$-10\ log_{10}(\frac{length\ of\ variants}{\#\ bases >= 3X\ coverage})$, and identity was computed as $100 * (1 - \frac{length\ of\ variants}{\#\ bases >= 3X\ coverage})$. Dotplots were generated with “mummerplot --fat” using the 1-to-1 filtered matches.

A previously published GM12878 PacBio assembly [5] was aligned as above with MUMmer v3.23. The resulting alignment files were uploaded to Assemblytics [58] to identify structural variants and generate summary figures. Versus GRCh38, the PacBio assembly identified 10,747 structural variants affecting 10.84 Mbp, and reported an equal balance of insertions and deletions (2,361 vs. 2,724), with a peak at approximately 300 bp corresponding to Alu repeats (Supplementary Figure 5 A, Supplementary Table 6). The high error rate of the nanopore assembly resulted in a much larger number of identified variants (69,151) affecting 23.45 Mbp, with a strong deletion bias (3,900 insertions vs. 28,791 deletions) (Supplementary Figure 5 B, Supplementary Table 6). The Illumina-polished assembly reduced the total variants

(47,073) affecting 16.24 Mbp but the deletion bias persisted (2,840 insertions vs. 20,797 deletions) (Supplementary Figure 5 C, Supplementary Table 6).

Base call analysis

Sequences were aligned to the 1000 genome GRCh38 reference (ftp://1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome/GRCh38_full_analysis_set_plus_decoy_hla.fa&sa=D&ust=1502277874429000&usg=AFQjCNHdQDdAYIXe95kV3iM58gv1SCuW0A) using BWA-MEM version 0.7.12-r1039 with the “-x ont2d” option [59]. The BAM alignments were converted to PAF format [60] and cigar-strings parsed to convert alignments to an identity. Summary statistics for each flowcell were tabulated separately and combined. Alignment length versus identity was plotted using smoothScatter in R. Depth of coverage statistics for each flowcell were obtained from “samtools depth -a” and combined. As for the assembly statistics, a genome size of 3,098,794,149 bp was used to compute bases covered. The mean coverage was 25.63 (63.20 sd). The minimum coverage was 0 and the maximum was 44,391. Excluding 0-coverage regions, the mean coverage was 27.41 (64.98 sd). The coverage histogram was plotted compared with randomly-generated Poisson values generated with R’s rpois function with $\lambda = 27.4074$.

Metrichor reads mapping to human chromosome 20 were additionally base-called with Scrappie v0.2.7. Scrappie reads comprised primarily of low-complexity sequence were identified using the sdust program included with Minimap (commit: 17d5bd12290e0e8a48a5df5afaeaef4d171aa133) [60] with default parameters (-w 64 -t 20). The total length of the windows in a single sequence were merged and divided by read length to compute percentage of low-complexity sequence in each read. Any read for which this percentage exceeded 50% was removed from downstream analysis. Without this filtering, BWA-MEM did not complete mapping the sequences after >30 days of runtime on 16-cores. Similar filtering on the Metrichor based reads had only a limited effect on the dataset.

To measure homopolymer accuracy, pairwise read-to-reference alignments were extracted for reads spanning all homopolymers of length 2 or greater. For efficiency, at most 1000 randomly selected instances were considered for each homopolymer length. Each homopolymer so-identified is enclosed by two non-homopolymer "boundary" bases (for example, the T and G in TAAAG). The number of match, mismatch, insertion and deletion alignment operations between the boundary bases was tabulated for each homopolymer, and alignments not anchored at the boundary bases with match/mismatch operations were ignored. Homopolymer call length was reported as the number of inserted bases minus the number of deleted bases in the extracted alignment, quantifying the difference between expected and observed sequence length. All base callers with the exception of Scrappie failed in large homopolymer stretches (e.g. Supplementary Figure 3), consistently capping homopolymers at 5 bp (the *k*-mer length of the model). Scrappie shows significant improvement, but tended to slightly over-call short homopolymers and under-call longer ones (Figure 2B).

To quantify deviations from the expected 50/50 allele ratio at heterozygous sites, 25,541 homozygous and 46,098 heterozygous SNP positions on chromosome 20 were extracted from the Illumina Platinum Genomes project VCF for GM12878, requiring a minimum distance of 10 bp between SNP positions. Scrappie base calls at these positions were extracted using samtools mpileup. Deviation from the expected allelic ratio was defined as d = abs(0.5 - [allele A coverage]/[allele A coverage + allele B coverage]). Averaged over all evaluated heterozygous SNPs, d = 0.13 and 90% of SNPs have d <= 0.27 (corresponding to approximately >= 25% coverage on the minor allele). Results were similar when stratified by SNP type.

Assembly polishing with Nanopolish

We ran the nanopolish consensus calling algorithm [14] on the chromosome 20 assemblies described above. For each assembly we sampled candidate variants from the base-called reads used to construct the contigs (using the “--alternative-basecalls” option) and input the original fast5 files (generated by the basecaller in the Metrichor computing platform) into a hidden Markov model, as these files contained the annotated events that the HMM relies on. The reads were mapped to the draft assembly using BWA-MEM with the “-x ont2d” option.

Each assembly was polished in 50,000 bp segments and the individual segments were merged into the final consensus. The nanopolish jobs were run using default parameters except the “--fix-homopolymers” and “--min-candidate-frequency 0.01” options were applied.

Assembly annotation

Comparative Annotation Toolkit (CAT) (https://github.com/ComparativeGenomicsToolkit/Comparative-Annotation-Toolkit commit c9503e7) was run on both the polished and unpolished assemblies. CAT uses whole genome alignments to project transcripts from a high-quality reference genome to other genomes in the alignment [61]. The gene finding tool AUGUSTUS is used to clean up these transcript projections and a combined gene set is generated [62].

To guide the annotation process, human RNA-seq data were obtained from SRA for a variety of tissues (Supplementary Table 7) and aligned to both GRCh38 and the two assembly versions. GENCODE V24 was used as the reference annotation. Two separate progressiveCactus [63] alignments were generated for each assembly version with the chimpanzee genome as an outgroup.

The frequency of frameshifting insertions or deletions (Indels) in transcripts was evaluated by performing pairwise CDS sequence alignments using BLAT in a translated protein parameterization. Alignments were performed both on raw transMap output as well as on the final consensus transcripts.

Paralogous alignments of a source transcript were resolved through a heuristic combination of alignment coverage, identity and synteny. Synteny is measured by counting how many gene projections near the current projection match the reference genome. In the case where multiple isoforms of a gene end up in different loci as the result of this process, a rescuing process is performed that chooses the highest scoring locus to place all isoforms at so that isoforms do not end up on different contigs. Through this process, a 1-1 orthology relationship is defined.

MHC analysis

The ultra-long assembly contains the MHC region between positions 2–6 Mb within a single 16 Mbp contig (tig01415017). Heterozygous sites were extracted by mapping Illumina reads to the polished assembly using BWA-MEM with default parameters. Alignments were post-processed according to the GATK 3.7 whole-genome variant calling pipeline, except for the “-T IndelRealigner” step using “--consensusDeterminationModel USE_READS”. The -T HaplotypeCaller parameter was used for variant calling. WhatsHap [64] was used to phase the Illumina variants with Nanopore reads reported to be contained in the contig by Canu. WhatsHap was modified to accept CRAM (http://genome.cshlp.org/content/21/5/734.long, https://bitbucket.org/skoren/whatshap) output since BAM files could not represent long CIGAR strings at the time of this analysis (https://github.com/samtools/hts-specs/issues/40). First, WhatsHap was run excluding any ultra-long sequences. This generated 18 phase blocks across the MHC. When ultra-long sequences were included the result was a single phase block comprising the entire MHC, supporting the utility of ultra-long reads in resolving haplotypes across large, complex regions in the genome. Nanopore reads were aligned back to the assembly using NGM-LR [38] and the combined VCF file used for phasing. Reads with more than 1 phasing marker were classified as haplotype A or B when >55% of their variants were in agreement (Figure 5A). A new assembly was generated for haplotypes A and B using only reads assigned to each haplotype as well as reads marked homozygous. The assemblies were polished by Pilon 1.21 [26] using the SGE pipeline at https://github.com/skoren/PilonGrid. Pilon was given all reads mapping to the MHC.

Exon sequences belonging to the six classical HLA genes were extracted from the phased assembly, and HLA types called at G group resolution. These results were compared to GM12878 HLA type reference data. For the class I and II HLA genes, with the exception of one DRB1 haplotype, there was good agreement between the best-matching reference type and the alleles called from the assembly (edit distance 0–1). Detailed examination of HLA-DRB1, however, showed that one exon (exon 2) is different from all reference types in the assembly, a likely error in the assembly sequence.

GM12878 G group HLA types for HLA-A, -B, -C, -DQA1, -DQB1 and -DRB1 are from ref. [65]; the presence of exactly one HLA-DRB3 allele is expected due to linkage with HLA-DRB1 (DRB1*03 is associated with HLA-DRB3, and DRB1*01 has no DRB3/4/5 association).

Genotyping SNPs using Nanopolish

Nanopolish was used for genotyping the subset of reads that mapped to human chromosome 20. The 1000 Genomes phase 3 variant set for GRCh38 was used as a reference and filtered to include only chromosome 20 SNPs that were not singletons (AC ≥ 2). This set of SNPs was input into "nanopolish variants" in genotyping mode ("--genotype"). The genotyping method extends the variant calling framework previously described [12] to consider pairs of haplotypes, allowing it to be applied to diploid genomes (option "--ploidy 2"). To evaluate their accuracy, genotype calls were compared to the "platinum calls" generated by Illumina [23]. When evaluating the correctness of a nanopore call, we required the log-likelihood ratio of a variant call (heterozygous or homozygous non-reference) to be at least 30, otherwise we considered the site to be homozygous reference.

Estimating SV genotyping sensitivity

Previously identified high confidence GM12878 SVs, validated with Moleculo and/or PacBio long reads, were used to determine genotyping sensitivity [29]. Using LUMPY

[28], we re-called SVs in the Platinum Genomes NA12878 Illumina dataset (paired-end reads; European Nucleotide Archive, Run Accession ERR194147), intersected these calls with the aforementioned high confidence set, and genotyped the resulting calls using SVTyper [28] and the same Platinum alignments, generating a set 2,414 high confidence duplications and deletions with accompanying genotypes. Nanopore reads from all flowcells were mapped using BWA-MEM (`bwa mem -k15 -W30 -r10 -B2 -O2 -L0`), and then merged into release-specific BAM files. Merged BAM files were subsampled using Samtools (`samtools view -s $COVERAGE_FRACTION`) to approximate coverage values as shown in Figure 2A. SVs were then genotyped in each subsampled BAM file using a modified version of SVTyper (http://github.com/tomsasani/svtyper). Generally, long nanopore reads are subject to higher rates of mismatches, insertions, and deletions than short Illumina reads. These features can result in "bleed-through" alignments, where reads align past the true breakpoint of an SV [66]. The modifications to SVTyper attempt to correct for the "bleed-through" phenomenon by allowing reads to align past the breakpoint, yet still support an alternate genotype. All modifications to SVTyper are documented in the source code available at the GitHub repository listed above (commit ID: d70de9c) (Supplementary Code 2). Nanopore and Illumina derived genotypes were then compared as a function of subsampled nanopore sequencing coverage.

The false-discovery rate of our SVTyper genotyping strategy was estimated by randomly permuting the genomic locations of the original SVs using BEDTools "shuffle" [67]. Centromeric, telomeric, and “gap” regions (as defined by the UCSC Genome Browser) were excluded when assigning randomly selected breakpoints to each SV. The randomly shuffled SVs were then genotyped in Illumina and nanopore data in the same manner as before. It is expected that the alignments at shuffled SV intervals would almost always support a homozygous reference genotype. So, all instances in which Illumina data supported a homozygous reference genotype, yet the nanopore data called a non-homozygous reference genotype, were considered false positives. SV coordinates were shuffled and genotyped 1000 times and the average false discovery rate over all iterations was 6.4%.

Nanopore and PacBio genotyping sensitivity was compared at a subset of our high confidence SV set. Because our high confidence set includes only “DUP” and “DEL” variants, and the Genome in a Bottle (GIAB) PacBio SV VCF (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz) does not report “DUP” variants, we compared genotypes at deletions with genomic coordinates that shared reciprocal overlap of at least 0.5 between the GIAB VCF and our high confidence SV VCF. We then compared nanopore genotypes (as determined by SVTyper) with the genotypes reported in the GIAB SV VCF. Importantly, the GIAB VCF was derived from a ~44× coverage dataset, whereas our dataset (containing data from both releases) represents only about ~32× coverage of the genome. Additionally, all nanopore data used in this analysis were aligned using BWA, while GIAB PacBio data were aligned using BLASR [69].

Scaling marginAlign and signalAlign data analysis pipelines

To handle the large data volume, the original marginAlign and signalAlign algorithms were ported to cloud infrastructures using the Toil batch system [68]. Toil allows for computational resources to be scaled horizontally and vertically as a given experiment requires and enables researchers to perform their own experiments in identical conditions. All of the workflows used and the source code is freely available from https://github.com/ArtRand/toil-signalAlign and https://github.com/ArtRand/toil-marginAlign. Workflow diagrams are shown in Supplementary Figure 10.

Generating a controlled set of methylated control DNA samples

For signalAlign, DNA methylation control standards were obtained from Zymo Research (cat. Number D5013). The standards contain a whole-genome-amplified (WGA) DNA substrate that lacks methylation and a WGA DNA substrate that has been enzymatically treated so all CpG dinucleotides contain 5-methyl cytosines. The two substrates were sequenced independently on two different flowcells using the sequencing protocol described above. Otherwise, training for signalAlign and nanopolish was carried out as previously described [35,36].

5-methyl cytosine detection with signalAlign

The signalAlign algorithm uses a variable order hidden Markov model combined with a hierarchical Dirichlet process (HMM-HDP) to infer base modifications in a reference sequence using the ionic current signal produced by nanopore sequencing [69]. The ionic current signal is simultaneously influenced by multiple nucleotides as the strand passes through the nanopore. Correspondingly, signalAlign models each ionic current state as a nucleotide *k*-mer. The model allows a base in the reference sequence to have any of multiple methylation states (in this case 5-methy cytosine or canonical cytosine). The model ties the probabilities of consistently methylated *k*-mers by configuring the HMM in a variable order meta-structure that allows for multiple paths over a reference *k*-mer depending on the number of methylation possibilities. To learn the ionic current distributions for methylated *k*-mers, signalAlign estimates the posterior mean density for each *k*-mer's distribution of ionic currents using a Markov chain Monte Carlo (MCMC) algorithm given a set of *k*-mer-to-ionic current assignments. Using the full model, the posterior for each methylation status is calculated for all cytosines in CpG dinucleotides.

5-methyl cytosine detection with nanopolish

Previous work describes using nanopolish to call 5-methylcytosine in a CpG context using a hidden Markov model [36]. The output of the nanopolish calling procedure is a log-likelihood ratio, where a positive log-likelihood ratio indicates evidence for methylation. Nanopolish groups nearby CpG sites together and calls the group jointly, assigning the same methylation status to each site in the group. To allow comparison to the bisulfite data each such group was broken up into its constituent CpG sites, which all have the same methylation frequency. Percent-methylation was calculated by converting the log-likelihood ratio to a binary methylated/unmethylated call for each read, and calculating the fraction of reads classified as methylated. A filtered score was also computed by first filtering reads where the absolute value of the log-likelihood ratio was less than 2.5 to remove ambiguous reads.

Data Availability Statement

Sequence data including raw signal files (FAST5), event-level data (FAST5), basecalls (FASTQ) and alignments (BAM) are available as an Amazon Web Services Open Dataset for download from https://github.com/nanopore-wgs-consortium/NA12878. Nanopore raw signal files and the 35x assembly are additionally archived and available from the European Nucleotide Archive under accession PRJEB23027.

Online Methods References

51. Sambrook, J. & Russell, D. W. *Molecular cloning: a laboratory manual*. (Cold Spring Harbor Laboratory Press, 2001).

52. Shomorony, I., Courtade, T. & Tse, D. Do read errors matter for genome assembly? in *2015 IEEE International Symposium on Information Theory (ISIT)* 919–923 (2015).

53. Bresler, G., Bresler, M. ’ayan & Tse, D. Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics* **14 Suppl 5,** S18 (2013).

54. Ukkonen, E. Approximate string-matching with q-grams and maximal matches. *Theor. Comput. Sci.* **92,** 191–211 (1992).

55. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33,** 623–630 (2015).

56. Böhringer, Stefan Gödde, René Böhringer, Daniel Schulte, Thorsten Epplen, Jörg T. A software package for drawing ideograms automatically. *Online J Bioinformatics* **1,** (2002).

57. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

58. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32,** 3021–3023 (2016).

59. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).

60. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long

sequences. *Bioinformatics* **32,** 2103–2110 (2016).

61. Zhu, J. *et al.* Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput. Biol.* **3,** e247 (2007).
62. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24,** 637–644 (2008).
63. Paten, B. *et al.* Cactus graphs for genome comparisons. *J. Comput. Biol.* **18,** 469–481 (2011).
64. Patterson, M. *et al.* WhatsHap: Haplotype Assembly for Future-Generation Sequencing Reads. in *Research in Computational Molecular Biology* 237–249 (Springer, Cham, 2014).
65. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47,** 682–688 (2015).
66. Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R. & Timp, W. Nanopore sequencing detects structural variants in cancer. *Cancer Biol. Ther.* **17,** 246–253 (2016).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
68. Vivian, J. *et al.* Rapid and efficient analysis of 20,000 RNA-seq samples with Toil. (2016). doi:10.1101/062497
69. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore

sequencing. *Nat. Methods* (2017). doi:10.1038/nmeth.4189

69. Chaisson, M. & Tesler, G. Mapping single molecule sequencing reads using Basic Local Alignment with Successive Refinement (BLASR): Theory and Application. *BMC Bioinformatics* (2012).

Figure 1 - Summary of Dataset

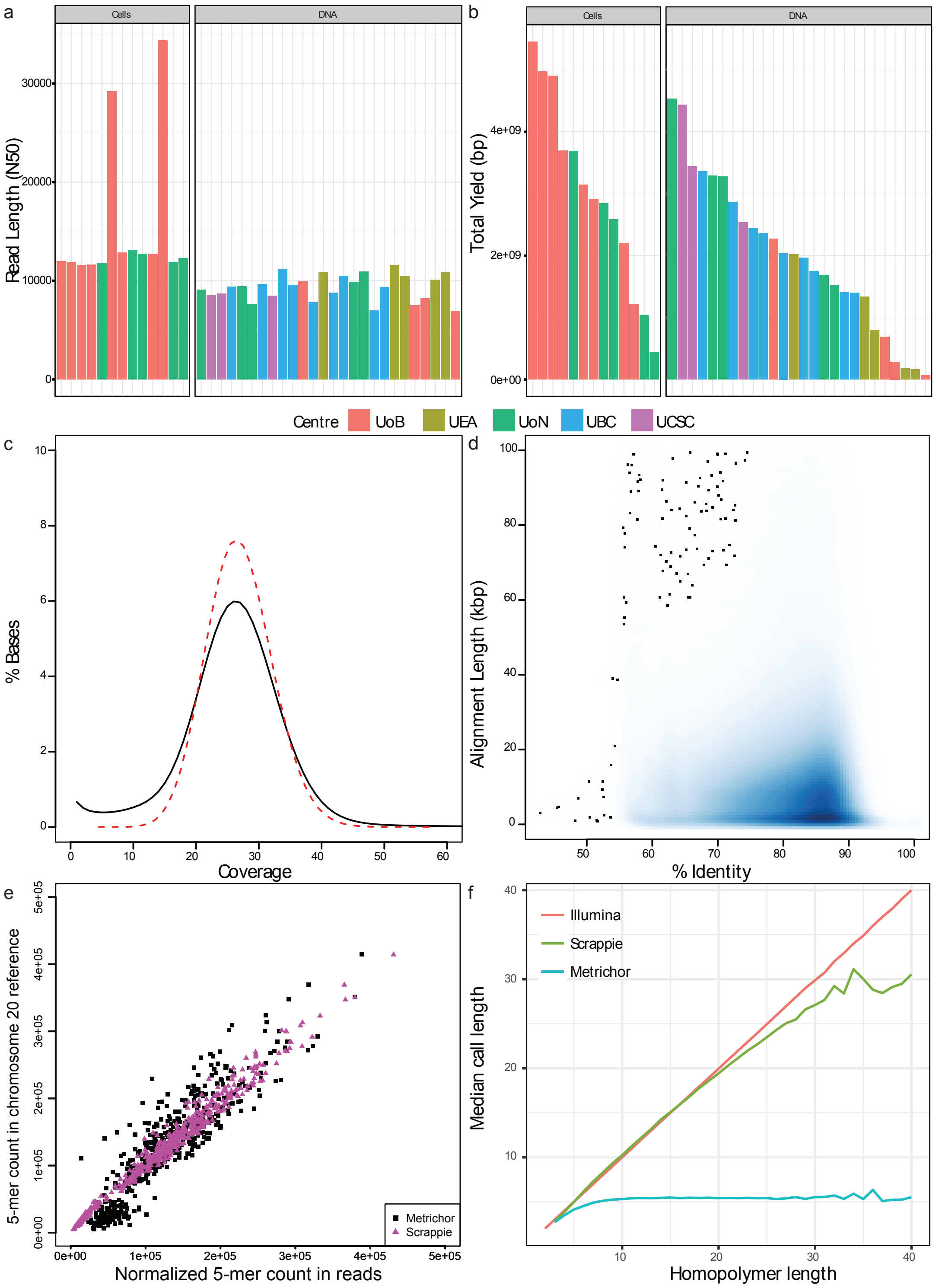
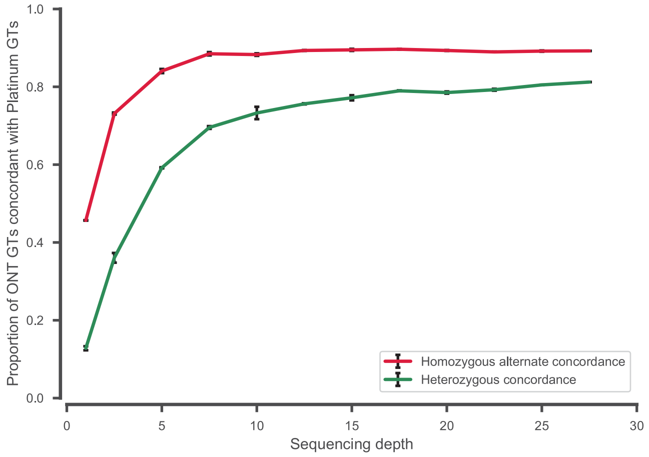


Figure 2 - Structural Variation/Genotyping

a



b

		Platinum (Illumina) Genotype		
		0/0	0/1	1/1
Nanopolish Genotype	0/0	727598	1730	75
	0/1	3217	29096	914
	1/1	601	49	21718

Figure 3 - Methylation Detection

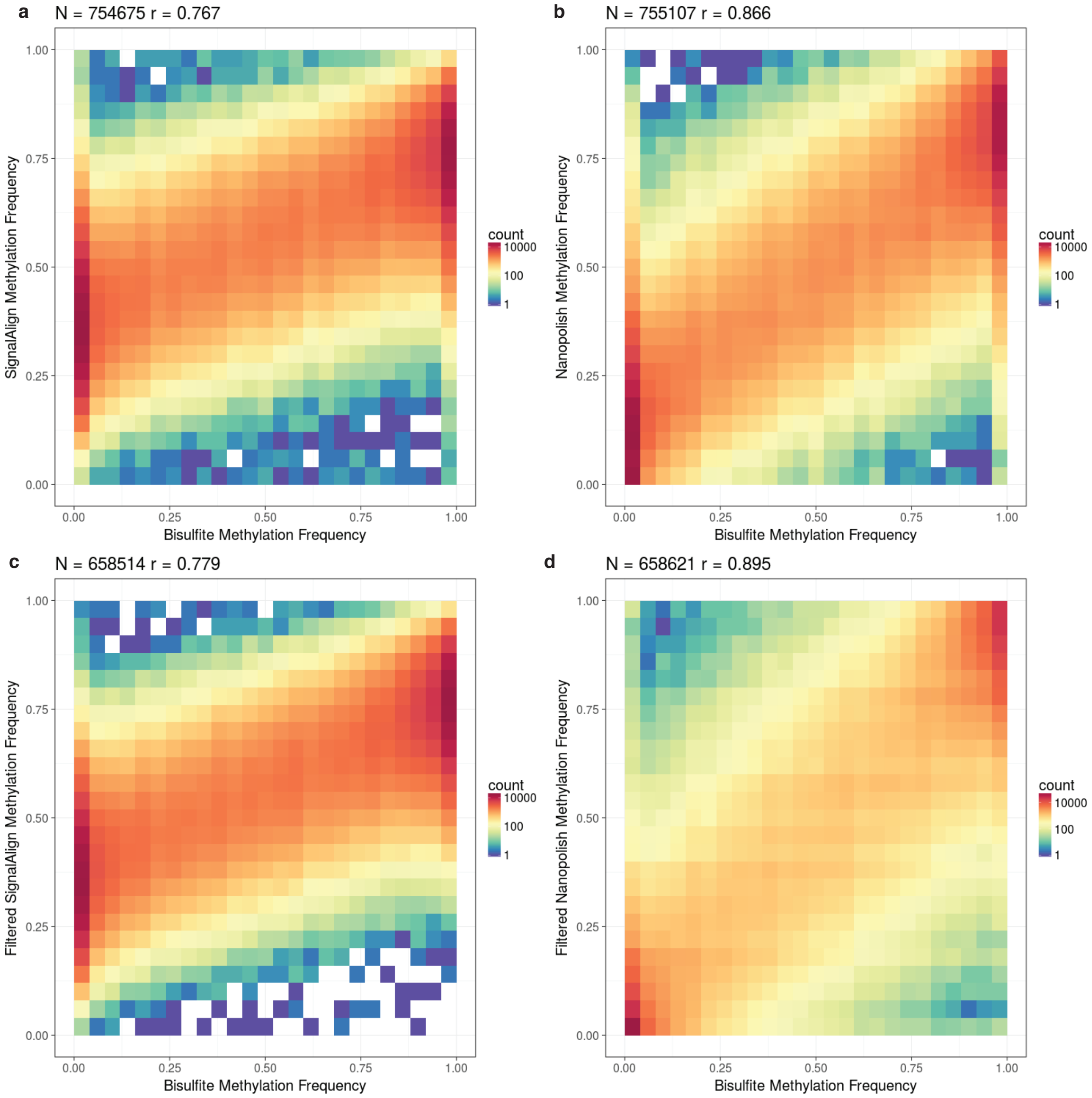


Figure 3 - Methylation Detection

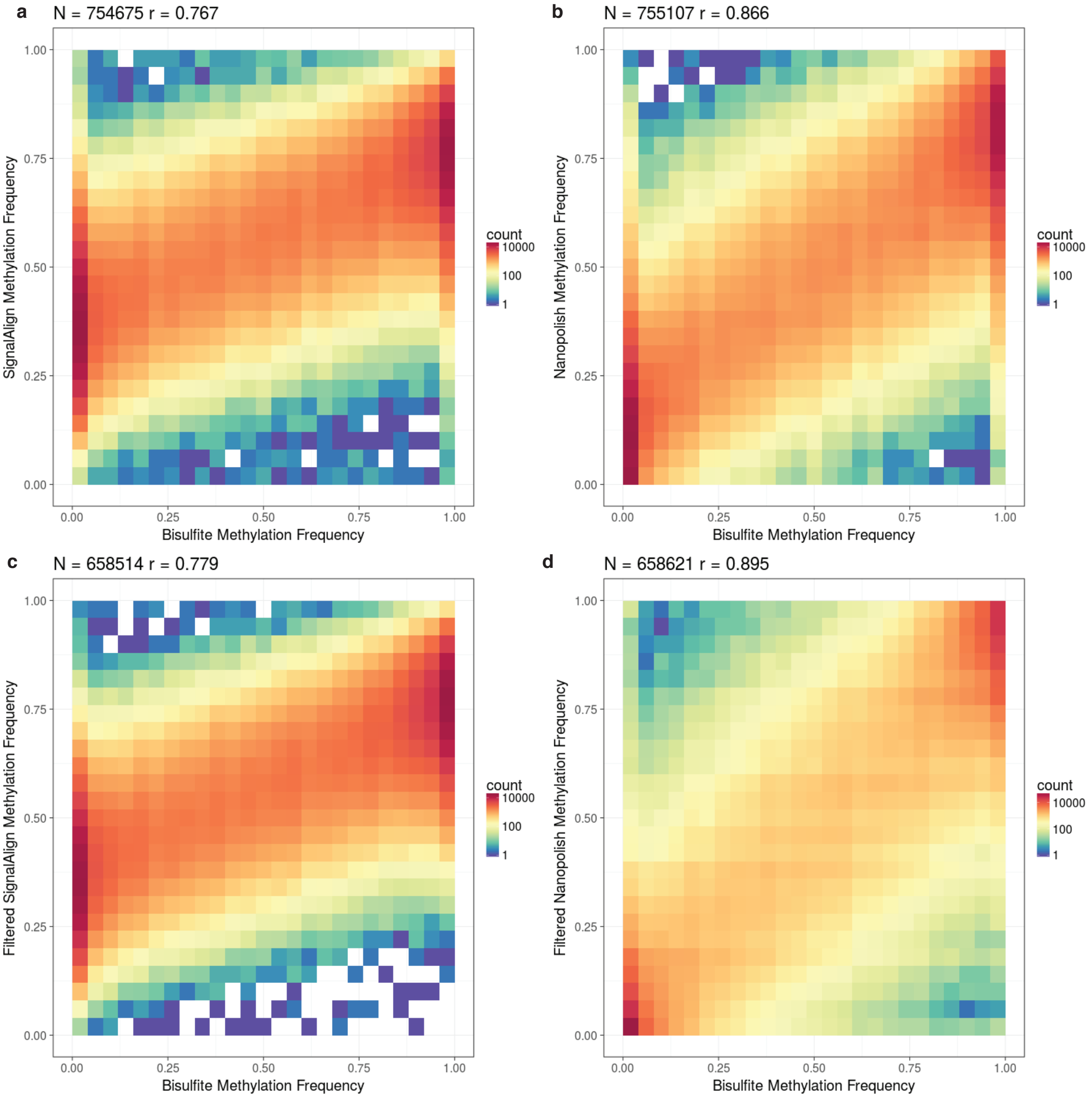


Figure 4 - Repeat modeling and assembly

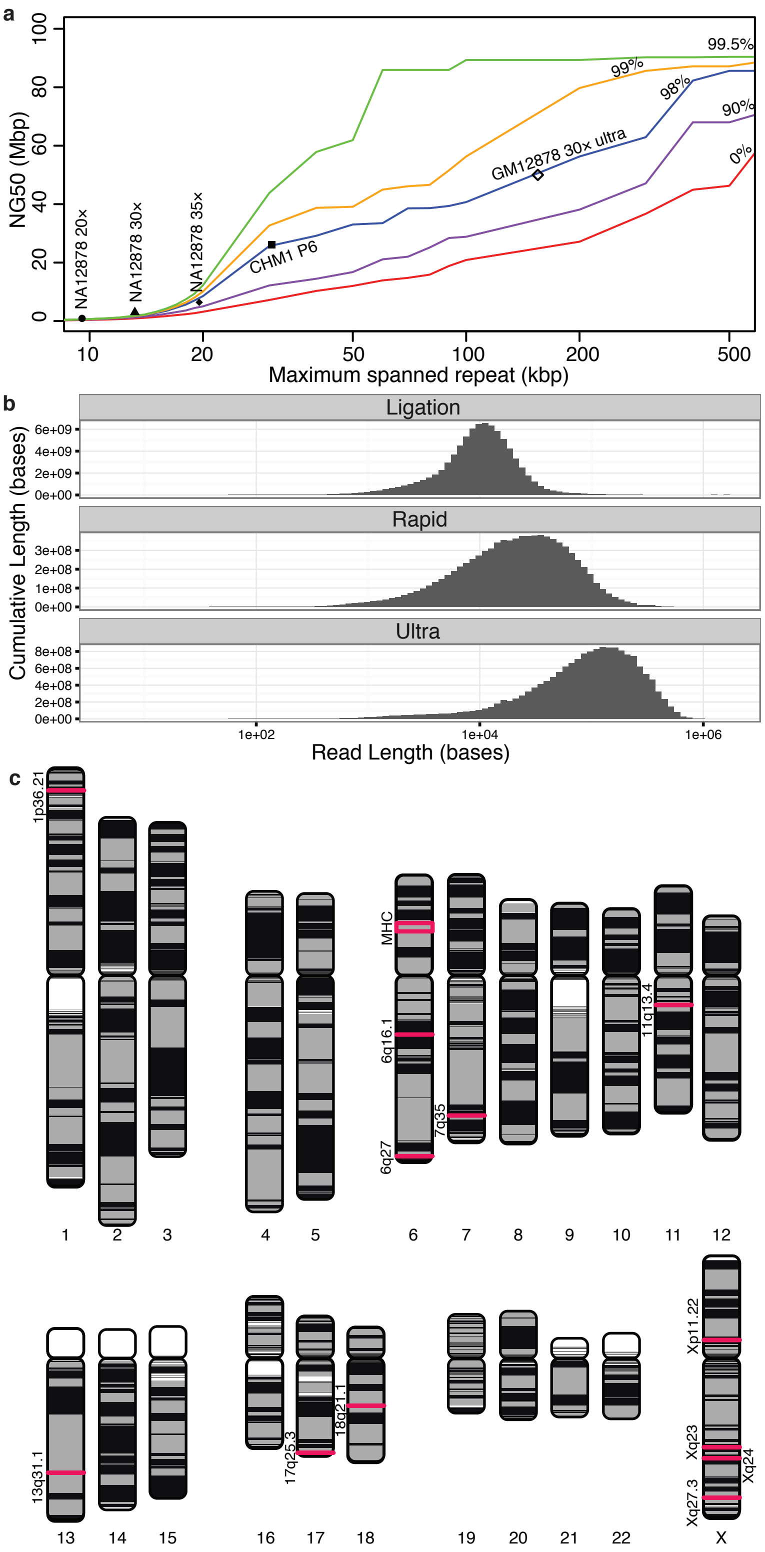
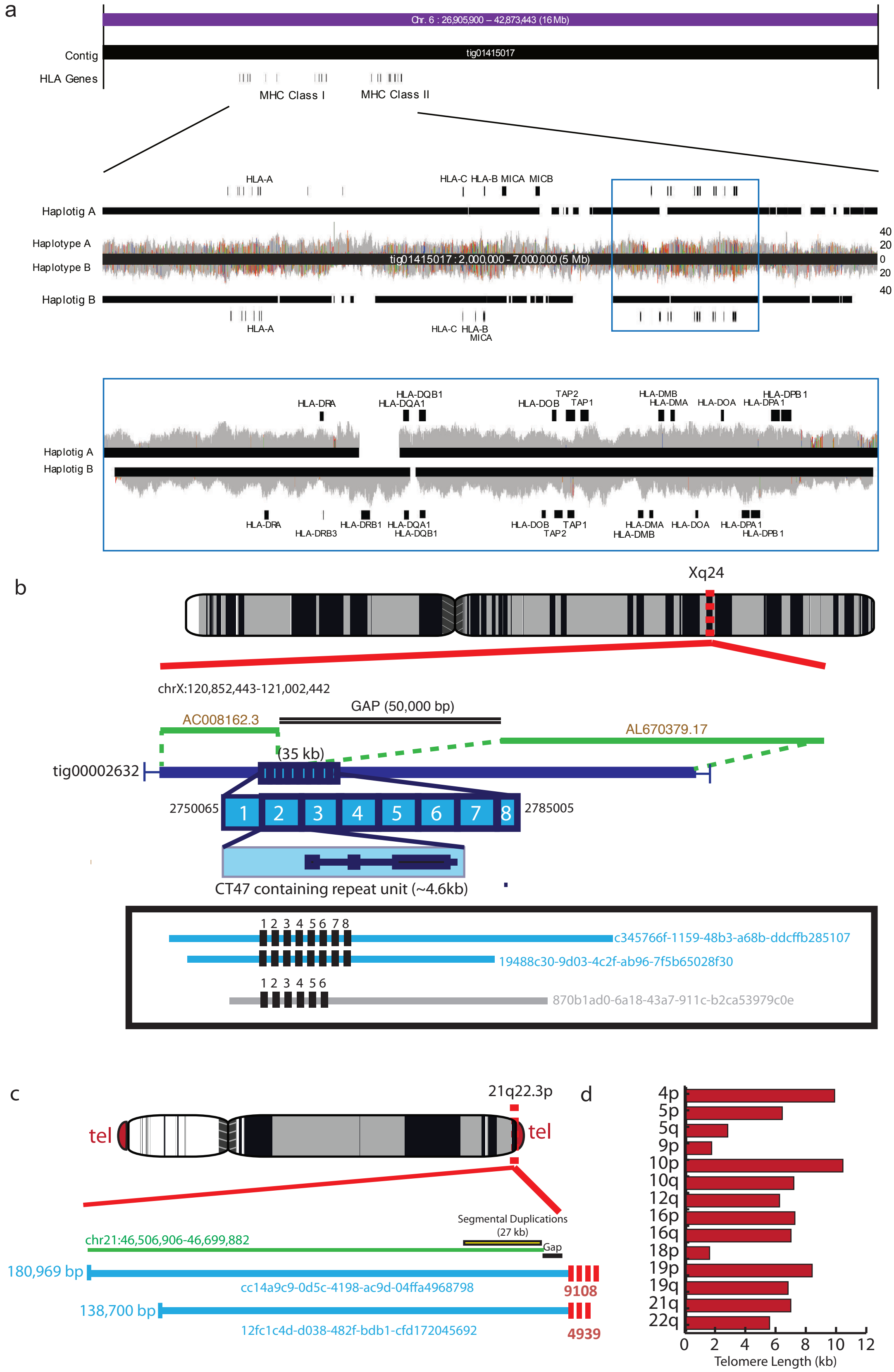


Figure 5 - Ultra Long Reads; Assembly and Telomeres


a
Chr. 6 : 26,905,900 – 42,873,443 (16 Mb)
Contig
tig01415017
HLA Genes
MHC Class I
MHC Class II
HLA-A
HLA-C HLA-B MICA MICB
Haplotig A
Haplotype A
Haplotype B
tig01415017 : 2,000,000 - 7,000,000 (5 Mb)
40
20
0
20
40
Haplotig B
HLA-A
HLA-C HLA-B
MICA
HLA-DRA
HLA-DQB1
HLA-DQA1
TAP2
HLA-DOB
TAP1
HLA-DMB
HLA-DMA
HLA-DOA
HLA-DPB1
HLA-DPA1
Haplotig A
Haplotig B
HLA-DRA
HLA-DRB3
HLA-DRB1
HLA-DQA1
HLA-DQB1
HLA-DOB
TAP1
TAP2
HLA-DMA
HLA-DMB
HLA-DOA
HLA-DPA1
HLA-DPB1
b
Xq24
chrX:120,852,443-121,002,442
GAP (50,000 bp)
AC008162.3
AL670379.17
(35 kb)
tig00002632
2750065
1
2
3
4
5
6
7
8
2785005
CT47 containing repeat unit (~4.6kb)
1 2 3 4 5 6 7 8
c345766f-1159-48b3-a68b-ddcffb285107
19488c30-9d03-4c2f-ab96-7f5b65028f30
1 2 3 4 5 6
870b1ad0-6a18-43a7-911c-b2ca53979c0e
c
21q22.3p
tel
tel
Segmental Duplications
(27 kb)
chr21:46,506,906-46,699,882
Gap
180,969 bp
cc14a9c9-0d5c-4198-ac9d-04ffa4968798
9108
138,700 bp
12fc1c4d-d038-482f-bdb1-cfd172045692
4939
d
4p
5p
5q
9p
10p
10q
12q
16p
16q
18p
19p
19q
21q
22q
0
2
4
6
8
10
12
Telomere Length (kb)