Improved assembly and variant detection of a haploid human genome using single-molecule, high-fidelity long reads

Mitchell R. Vollger[1,9], Glennis A. Logsdon[1,9], Peter A. Audano[1], Arvis Sulovari[1], David Porubsky[1], Paul Peluso[2], Aaron M. Wenger[2], Gregory T. Concepcion[2], Zev N. Kronenberg[2], Katherine M. Munson[1], Carl Baker[1], Ashley D. Sanders[3], Diana C.J. Spierings[4], Peter M. Lansdorp[4,5,6], Urvashi Surti[7], Michael W. Hunkapiller[2], Evan E. Eichler[1,8,#]

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA
[2]Pacific Biosciences of California, Inc., Menlo Park, CA 94025, USA
[3]European Molecular Biology Laboratory, Genome Biology Unit, 69117, Heidelberg, Germany
[4]European Research Institute for the Biology of Ageing, University of Groningen, University Medical Center Groningen, 9713 AV Groningen, The Netherlands
[5]Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada
[6]Department of Medical Genetics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[7]Department of Pathology, University of Pittsburgh School of Medicine and University of Pittsburgh Medical Center, Pittsburgh, PA, USA
[8]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[9]These authors contributed equally to this work.

[#] Correspondence to:
Evan E. Eichler, Ph.D.
Department of Genome Sciences
University of Washington School of Medicine
Foege S413A, Box 355065
3720 15th Ave NE
Seattle, WA 98195-5065
Phone: 206-543-9526
E-mail: eee@gs.washington.edu

ABSTRACT

The sequence and assembly of human genomes using long-read sequencing technologies has revolutionized our understanding of structural variation and genome organization. We compared the accuracy, continuity, and gene annotation of genome assemblies generated from either high-fidelity (HiFi) or continuous long-read (CLR) datasets from the same complete hydatidiform mole human genome. We find that the HiFi sequence data assemble an additional 10% of duplicated regions and more accurately represent the structure of large tandem repeats, as validated with orthogonal analyses. Additionally, the HiFi genome assembly was generated in significantly less time with fewer computational resources than the CLR assembly. Although the HiFi assembly has significantly improved continuity and accuracy in many complex regions of the genome, it still falls short of the assembly of centromeric DNA and the largest regions of segmental duplication using existing assemblers. Despite these shortcomings, our results suggest that HiFi may be the most effective stand-alone technology for *de novo* assembly of human genomes.

INTRODUCTION

Recent advances in long-read sequencing technologies, including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have revolutionized the assembly of highly contiguous mammalian genomes (Bickhart et al., 2017; Chaisson et al., 2015; Gordon et al., 2016; Huddleston et al., 2017; Jain et al., 2018; Kronenberg et al., 2018; Low et al., 2019; Seo et al., 2016; Steinberg et al., 2016). For example, individual laboratories can now accurately assemble >90% of mammalian euchromatin in less than 1,000 contigs within a few months. However, the generation of high-quality datasets is costly and requires computational resources unavailable to most researchers. Long-read *de novo* assemblies of human samples typically require 20,000-50,000 CPU hours (Chin et al., 2016; Koren et al., 2017) and terabytes of data storage.

With the recent introduction of high-fidelity (HiFi) sequence data from PacBio and the development of the SMRT Cell 8M, the accessibility of *de novo* assembly using Single-Molecule, Real-Time (SMRT) sequencing data has significantly improved. With 28-fold sequence coverage of the Genome in a Bottle (GIAB) Ashkenazim sample HG002, Wenger and colleagues demonstrated that it is possible to create a *de novo* assembly comparable to previous long-read assemblies with half the data and one-tenth the compute power (Wenger et al., 2019). While compute time and throughput have improved, there is little comparison of the HiFi assembly quality of HG002 to a previous continuous long-read (CLR) HG002 genome assembly and limited assessment of the more difficult regions of the genome.

Here, we generate 24-fold sequence coverage and produce a *de novo* assembly of a complete hydatidiform mole human genome (CHM13) with HiFi data. We directly compare it to a previous assembly of CHM13 produced with CLR data (Kronenberg et al., 2018). The accurate assembly of the CHM13 genome is valuable for several reasons. First, due to its single-haplotype nature, it allows for better resolution of highly duplicated sequences, including segmental duplications (SDs) and tandem repeats. This 5-8% portion of the genome represents some of the most challenging regions to resolve. Second, its monoallelic nature permits the detection and unambiguous resolution of structural variants that are crucial in disease and evolution. Finally, it allows for complete and absolute deduction of the sequence accuracy of a genome assembly (i.e., quality value (QV)) because there is only one haplotype for comparison. As a result, large-insert clone BAC sequences from the same source

material can be expected to align at near 100% sequence identity and therefore be used to reliably compute the accuracy of different sequencing platforms and assembly approaches.

RESULTS

Whole-genome assembly with HiFi versus CLR reads. To assess the utility of PacBio's HiFi technology (Wenger et al., 2019) for *de novo* assembly, we set out to compare assemblies of the CHM13 genome using either HiFi data (generated on the Sequel II platform) or CLR data (generated on the RSII platform). To do this, we generated 24-fold HiFi circular consensus sequence (CCS) data from four SMRT Cells 8M. Each SMRT Cell produced, on average, 19.1 Gbp of QV > 20 sequence data (range 14-25 Gbp) with an average consensus read length of 10.95 kbp. The long-read sequence data were of high quality, with 61.2% of the quality-filtered CCS reads showing an estimated QV > 30 (**Figure 1**). The generation of HiFi data using the CCS algorithm took on average 12,500 CPU hours for each SMRT Cell 8M.

Using Canu (Koren et al., 2017) (**Methods**), we generated a *de novo* assembly (assembly FTP) using the HiFi CCS data (hereafter termed "HiFi assembly") and compared it to a previous FALCON assembly of CHM13 (accession GCA_002884485.1) generated with 77-fold CLR data (hereafter termed "CLR assembly") (**Figure 2**). The HiFi assembly required only 2,800 CPU hours, whereas the CLR assembly required more than 50,000 CPU hours. This reduction in runtime is because the N × N correction step common to both FALCON and Canu can be skipped with adequate input read quality. It might be expected that the shorter read length of the HiFi data (N50 10.9 vs. 17.5 kbp) might lead to a less continuous assembly; however, we observed that the HiFi assembly had an N50 of 29.5 Mbp, which is slightly higher than the N50 of the CLR assembly (29.3 Mbp; **Figure 2**, **Table 1**).

To determine assembly base-pair accuracy, we sequenced and assembled the inserts (**Methods**) of 31 randomly selected BACs from a genomic library produced from the CHM13 cell line (VMRC59). We estimated assembly accuracy by aligning these sequence inserts to the HiFi and CLR assemblies (**Methods**). We found that, before any polishing, the consensus accuracy of the HiFi assembly was much higher than the CLR assembly (median BAC QV 34.4 vs. 27.5). Next, we polished the CLR assembly using 77-fold coverage of CLR reads with Quiver and the HiFi assembly using 355-fold coverage of CCS subreads with Arrow. In this experiment, once again, the HiFi assembly was superior to the CLR assembly with respect to accuracy (median QV 44.3 vs. 40.7) (**Table 1**).

While the initial assembly of the HiFi data was relatively rapid (2,800 CPU hours), subsequent polishing with Arrow required an additional 7,200 CPU hours. We were curious if we could reduce the polishing time by not incorporating subread information and using only the HiFi data . To do this, we applied Racon (Vaser et al., 2017) to polish our assembly with only the HiFi CCS reads. This Racon-based polishing step finished in only 135 CPU hours (100 for alignment and 35 for polishing) and offered improved accuracy over Arrow (median QV 45.0 vs. 44.3). After a second round of Racon polishing, there was only one single-nucleotide difference between the HiFi assembly and the BACs excluding indels. Using Illumina WGS data as a third orthogonal platform, we determined that this difference is likely not a sequence error but rather a *bona fide* mutational change that occurred since the CHM13 cell line diverged from the propagated VMRC59 BAC (**Figure S1**). With the exception of remaining single

basepair indels, this finding suggests that the quality values reported here should be considered lower bounds due to subsequent propagation errors in BAC DNA (**Supplemental Note, Figure S2**).

To evaluate the global contiguity of the respective assemblies, we generated and applied 2.8-fold sequencing data from strand-specific sequencing (Strand-seq) of the CHM13 cell line. Strand-seq is able to preserve structural contiguity of individual homologs by tracking the read directionality and, therefore, can be used for detection of misassembled contigs in *de novo* assemblies (Falconer et al., 2012; Sanders et al., 2017) (**Figure 3**). Using this analysis, we detected seven misassembled contigs that contain eight breakpoints in the HiFi assembly. In contrast, we detected a slightly lower number of misassembled contigs (5) and breakpoints (5) in the CLR assembly (**Table 2**). However, given the number of assembled contigs, these results demonstrate that both assemblies are highly accurate, with <0.5% misassembly.

Segmental duplication analyses. SDs are often recalcitrant to genome assembly due to their high (>90%) sequence identity, length (>1 kbp), and complex modular organization. Therefore, the accuracy and completeness of SDs is a particularly useful metric for assembly quality since these most often correspond to the last gaps in the euchromatic portions of long-read assemblies (Chaisson et al., 2015). We performed a number of analyses to assess the SD resolution in the HiFi and CLR assemblies. First, we compared the percentage of SDs resolved in both genome assemblies, as well as the human reference genome and several recently published long-read assemblies (**Methods**) (Vollger et al., 2019). Requiring that SDs are anchored contiguously with unique flanking sequence, we found that, on average, 43% of SDs are resolved in the CHM13 HiFi assembly compared to 34% in the CLR assembly. Although the majority of human SDs remain unassembled, this is the highest fraction of resolved SDs for any of the published assemblies analyzed thus far (Huddleston et al., 2017; Jain et al., 2018; Seo et al., 2016; Shi et al., 2016), with an average 15% increase over even the ultra-long ONT assembly of NA12878 (**Figure 4**). Additionally, the number of bases with significantly elevated coverage (mean + three standard deviations) (Vollger et al., 2019) in the HiFi assembly was reduced by 16% as compared to the CLR assembly (27.0 vs. 32.1 Mbp). This indicates that the HiFi assembly has fewer collapsed sequences compared to the CLR assembly, with multiple SDs now represented by a single contig.

Next, we specifically focused on the pericentromeric regions of the genome where megabases of interchromosomal duplications have accumulated during the course of great ape evolution (She et al., 2004, 2006). We first assessed the contiguity and coverage within the 1 Mbp regions flanking each centromere by calculating a pericentromere-specific NG50. We found that the HiFi assembly had an NG50 of 480.6 kbp, whereas the CLR assembly had a NG50 of only 191.5 kbp (**Figure 5**). Next, we assessed contiguity within the pericentromeric regions by counting the number of contigs within the 1 Mbp region flanking the centromeres for each assembly (**Figure S3**). Assemblies with fewer contigs have increased contiguity and improved assembly; therefore, we expected that the HiFi assembly would have fewer contigs within many of these regions. Indeed, we found that the HiFi assembly had reduced or the same number of contigs at 54.3% (25/46) of the 1 Mbp pericentromeric regions when compared to the CLR assembly [32.6% (15/46) of the pericentromeric regions had fewer contigs, and 21.7% (10/46) had the same number of contigs in both assemblies]. The remaining pericentromeric regions were split between having no contig representation (8.7%; 4/46) and an increased number of contigs (37.0%; 17/46) in the HiFi assembly relative to the CLR assembly. We hypothesized that the

increased number of contigs in these regions in the HiFi assembly may be indicative of fragmented sequences not found in the CLR assembly (**Figure S4**). When we tested this hypothesis by summing up the total contig coverage in the 1 Mbp windows flanking the centromeres, we found that, indeed, the HiFi assembly had an additional 5 Mbp of pericentromeric sequence missing from the CLR assembly (**Figure S5**).

To assess the sequence accuracy and contiguity within SD regions, we compared HiFi and CLR assemblies to 310 sequenced and assembled large-insert BAC clones of CHM13 origin (**Figure 6**). Once again, we found that the assembled HiFi assembly is more accurate (median QV 34.0) than the CLR assembly (median QV 31.2) against BACs that align along at least 95% of their length. We suspect the increased QV is due to the inability of the correction step in FALCON to correctly resolve paralog-specific reads into different groups. Although the HiFi shows higher QV, it should be noted that both assemblies are far less accurate for SDs than unique regions of the genome. Additionally, we find that the HiFi-assembled contigs are more continuous within the sampled SD regions: in 252 of the 310 (82%) BACs, the alignment length to the HiFi assembly is greater than or equal to the alignment length to the CLR assembly (**Figure 6**).

A significant fraction of high-identity duplications remain collapsed and unassembled in both the CLR and HiFi assemblies. However, we recently developed a method, Segmental Duplication Assembler (SDA), that can resolve collapsed duplications by taking advantage of long reads that share multiple paralog-specific variants (PSVs) and then grouping them using correlation clustering (Vollger et al., 2019). The algorithm depends on the length of the underlying reads, and since HiFi reads are significantly shorter (N50 10.9 vs. 17.5 kbp), we were concerned that SDA would be limited. To test the ability of HiFi and CLR to resolve collapses, we selected five problematic gene-rich regions of biomedical and biological importance and directly compared the potential of correlation clustering to partition and assemble such regions (**Table 3**; these regions contained the *OPN1LW, NOTCH2NL, SRGAP2, FCGR2/3, KANSL1* genes). Of the five regions: two were resolved more accurately by the CLR reads (*OPN1LW, KANSL1*), one was equivalent between HiFi and CLR (*SRGAP2*), and two were better resolved by the HiFi reads (*NOTCH2NL, FCGR2/3*). These results are encouraging since SDA was optimized to handle CLR data (Vollger et al., 2019), and we believe that future improvements to SDA that take advantage of the high-quality SNVs embedded within the HiFi data will resolve even more collapsed regions of genomes.

Tandem repeat resolution. Since tandem repeat sequences are often difficult to resolve for both length and content, we assessed whether short tandem repeats (STRs) and variable number of tandem repeats (VNTRs) were correctly assembled in the HiFi and CLR assemblies. We identified 3,074 tandem repeats that were ≥1 kbp, on average, across the six Human Genome Structural Variation Consortium (HGSVC) haplotype-resolved assemblies (Chaisson et al., 2019). For each locus, we compared the mean length of the region in the HiFi and CLR assemblies against an orthogonal set of ultra-long ONT reads generated from CHM13 (**Methods**). A total of 3,002 (97.7%) and 2,935 (95.5%) of the tandem repeats assembled with HiFi and CLR reads, respectively. Both HiFi and CLR assemblies had a high length concordance with ONT reads (Pearson's correlation coefficients $\rho = 0.853$ and $\rho = 0.838$, respectively) over tandem repeats that were resolved in at most a single contig by each assembly and spanned by more than one ONT read (n = 2,900). When we compared loci within each assembly to the mean length of the region in ultra-long ONT reads (with at least one spanning read)

(**Figure 7**), we found that the HiFi contigs had a lower root-mean-square (RMS) error of 0.86 kbp, while the CLR contigs had an RMS error of 0.95 kbp.

Further restricting the analysis to VNTRs present in HiFi but completely absent from the CLR assembly (n = 90), 53% (n = 48) of the loci agreed in length with the ONT reads. Inversely, restricting the analysis to VNTRs present in CLR but completely absent from the HiFi assembly (n = 23), 52% (n = 12) of the loci agreed in length with the ONT reads. The 48 validated HiFi-only tandem repeats were significantly larger than the 12 validated CLR-only tandem repeats, with average sizes of 2,024 bp and 1,270 bp, respectively (Wilcoxon rank sum test p-value = 0.03). This pattern suggests that HiFi reads accurately assemble large tandem repeats that may be inaccessible to CLR. Some of these loci were genic, such as a 20 kbp VNTR composed of 53 bp repeat units in the 3' UTR of *OR2T1* (**Figure 8**), or an expansion of an 83 bp protein-encoding VNTR in *MUC12* that increased from 95 (CLR) to 232 (HiFi) tandem repeat copies (**Figure 8**). Overall, the HiFi assembly more accurately represented the content and sequence length of the tandem repeats, particularly in previously unrepresented or collapsed regions of the CLR assembly, based on orthogonal validation experiments.

Structural variant analyses. Since errors in an assembly will lead to false-positive variant calls, we assessed the utility of assembled HiFi data as a variant discovery tool and used it as a metric to evaluate assembly quality. For each assembly, we called insertions and deletions against GRCh38 from contig alignments (**Methods**). We generated a callset for each assembly before and after polishing using a variety of tools, including Racon, Quiver, Arrow, Pilon, and a FreeBayes-based indel correction pipeline (Chin et al., 2013; Kronenberg et al., 2018; Vaser et al., 2017; Walker et al., 2014). We found that structural variant (SV, indels ≥ 50 bp) calls were largely consistent among assemblies (**Table 4**). Although HiFi read quality is substantially higher, polishing was required to reduce the number of false positive indel calls (**Table 5**). Overall, we found that the number of insertions and deletions were comparable between polished HiFi and CLR assemblies.

Gene open reading frame annotations. Long-read sequencing platforms exhibit high indel error rates due to missed and erroneous incorporations during real-time sequencing. As a result, predicted open reading frames are often disrupted, leading to potential problems in gene annotation (Watson and Warr, 2019) unless additional error correction steps are employed (Kronenberg et al., 2018). We compared the SV and indel callsets to human RefSeq annotations and identified likely gene-disruptive events. In the unpolished HiFi assembly, we found 19,278 SVs and indels putatively disrupting 4,756 of 18,037 RefSeq genes within the assembly consensus regions (26%), which reduced to 128 after polishing with two rounds of Racon (0.71%) (**Table 6**). Before polishing, these predicted gene-disruptive SVs and indels were overwhelmingly single-base-pair errors (96%; 18,456 of 19,278), which were greatly reduced after polishing (57%; 99 of 174). As expected, the CLR assembly had more likely-disrupted genes before polishing (64%; 11,591 of 17,991 genes in its consensus region), but this declined to 202 after polishing (1.1%). We found fewer predicted disrupted genes outside of repetitive events in the HiFi assembly (54 HiFi vs. 57 CLR), and this trend increases inside SDs where short reads may not polish as effectively (35 HiFi vs. 99 CLR). It is worth noting that 2,412 protein coding genes (13%) have exons in SDs, and this difference between the HiFi and CLR assemblies represents 2.7% of these duplicated protein coding genes.

DISCUSSION

The generation and assembly of HiFi and CLR long-read sequence data from the same haploid source material allows us to directly compare the accuracy and contiguity of these technologies without the added complication of disentangling haplotypes needed to resolve SV alleles. We conclude that there are three key strengths of the HiFi technology over CLR technology. First, the time to generate the *de novo* assembly is reduced 10-fold, and it will likely be reduced further as HiFi assemblers are developed and optimized. This not only makes *de novo* assembly of human genomes accessible to a larger number of research groups but also paves the ways for larger cohorts of individuals to be sequenced and assembled. Although assembly time is drastically reduced, the compute time required to generate HiFi data by the CCS algorithm is substantial (~50,000 CPU hours in total).

Second, our analyses confirm that, both in terms of quality and continuity, the HiFi assembly is generally superior or at least comparable to the CLR assembly despite the shorter read lengths and effectively reduced genome coverage (Wenger et al., 2019). One significant advance is that HiFi assemblies can be polished without reverting to the underlying subreads, which saves approximately 1 terabyte of subread data and 7,000 hours of additional compute time. Polishing remains an absolute requirement to reduce indel errors and obtain a high-quality final assembly. Human CLR datasets ultimately require orthogonal Illumina data, and our results show that HiFi datasets alone achieve a greater level of accuracy for annotated protein-coding genes.

Finally, we demonstrate that, in some of the most difficult regions of the genome (i.e., SDs, large tandem repeats, and pericentromeric regions), the HiFi assembly shows improved continuity and representation, but modest accuracy improvements. Highly accurate HiFi data allows for the assembly of an additional 10% of duplicated sequences and better recovers the structure of tandem repeats such that they more exactly reflect the genomic length of VNTRs and STRs as confirmed by orthogonal analyses. We note, however, that the accuracy of the duplicated and tandem repeat regions is still lower than that of unique regions of the genome. Follow-up procedures such as SDA, which are designed to target and further resolve collapsed regions, show mixed results especially among the most highly identical human duplications. Our analysis suggests that this is a limitation of the shorter read lengths of HiFi (N50 of 10.9 vs. 17.5 kbp), which reduces the power needed to phase PSVs and assign collapsed reads to their respective duplicated loci. Nevertheless, we believe the results are encouraging since methods such as SDA were optimized to handle CLR data (Vollger et al., 2019). Future improvements to SDA that take advantage of the high-quality SNVs embedded within the HiFi data in duplicated regions will resolve even more collapsed regions of assembled genomes.

Next steps involve benchmarking and optimization of performance within diploid genome assemblies. Much of the recent advances in improving the contiguity of genome assemblies from telomere to telomere (Miga, Koren et al., unpublished) have been based on the same haploid source material analyzed here. It is clear that current HiFi genome assemblies are not as contiguous as those generated with high-coverage ultra-long ONT or with combinations of PacBio and ONT. While the haploid source material has been extremely useful for benchmarking, the ultimate challenge is the accurate assembly of human diploid genomes where both chromosomal haplotypes are fully resolved. Incorporation of linking-read technologies, such as Strand-seq, Hi-C, and 10X Genomics, or trio-binning approaches have been shown to significantly improve phasing and SV sequence and assembly (Chaisson et al., 2019; Koren et al., 2018; Kronenberg et al., 2019). It is likely that such approaches

could be combined with HiFi datasets to enhance telomere-to-telomere phasing and improve the accuracy of more complex repeats. Alternatively, the use of ultra-long-read datasets coupled with HiFi sequencing on the same samples will likely enhance both the phasing and accuracy of diploid genome assemblies. A useful standard for diploid genome assembly will be to repeat these analyses for two haploid source genomes in order to model the effect and accuracy of *in silico* diploid genomes as we (Huddleston et al., 2017) and others (Li et al., 2018) have shown.

Notwithstanding these advances, significant challenges remain for complete genome assembly, including large SDs, centromeric satellites, and acrocentric DNA. For example, although the CHM13 HiFi assembly we generated is highly contiguous (N50 29.5), an analysis of the unmappable reads shows an abundance of repetitive DNA (69.8%; **Figure 9**). Of these sequences, 45.6% consist of various classes of satellite repeats, which populate centromeres and the acrocentric portions of human chromosomes. Given the accuracy of these unmapped sequence reads, they will be quite valuable in obtaining the first overview of the sequence content and composition of these more complex heterochromatic regions. Obtaining even longer HiFi reads than used in this assembly (i.e., >11 kbp average used here) will be necessary to accurately anchor and sequence-resolve these repeat regions in future genome assemblies. Coupled with advances from other long-read technologies, such as ONT, it is clear that highly accurate telomere-to-telomere assemblies of diploid genomes will soon be achievable.

FIGURES

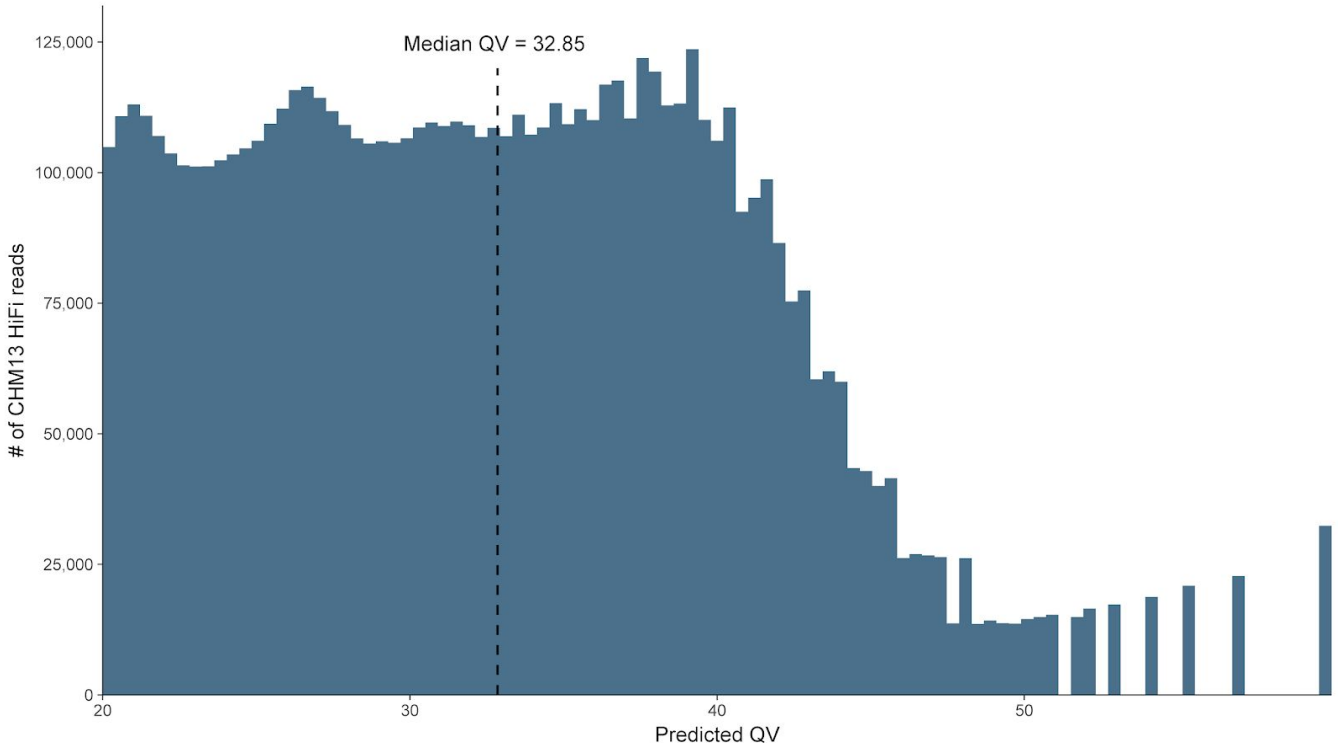

Figure 1. Distribution of read QV within the HiFi dataset. PacBio-predicted read quality of HiFi reads with a QV > 20 used for *de novo* assembly and all other analysis. The median QV of the reads over QV 20 was 32.85.

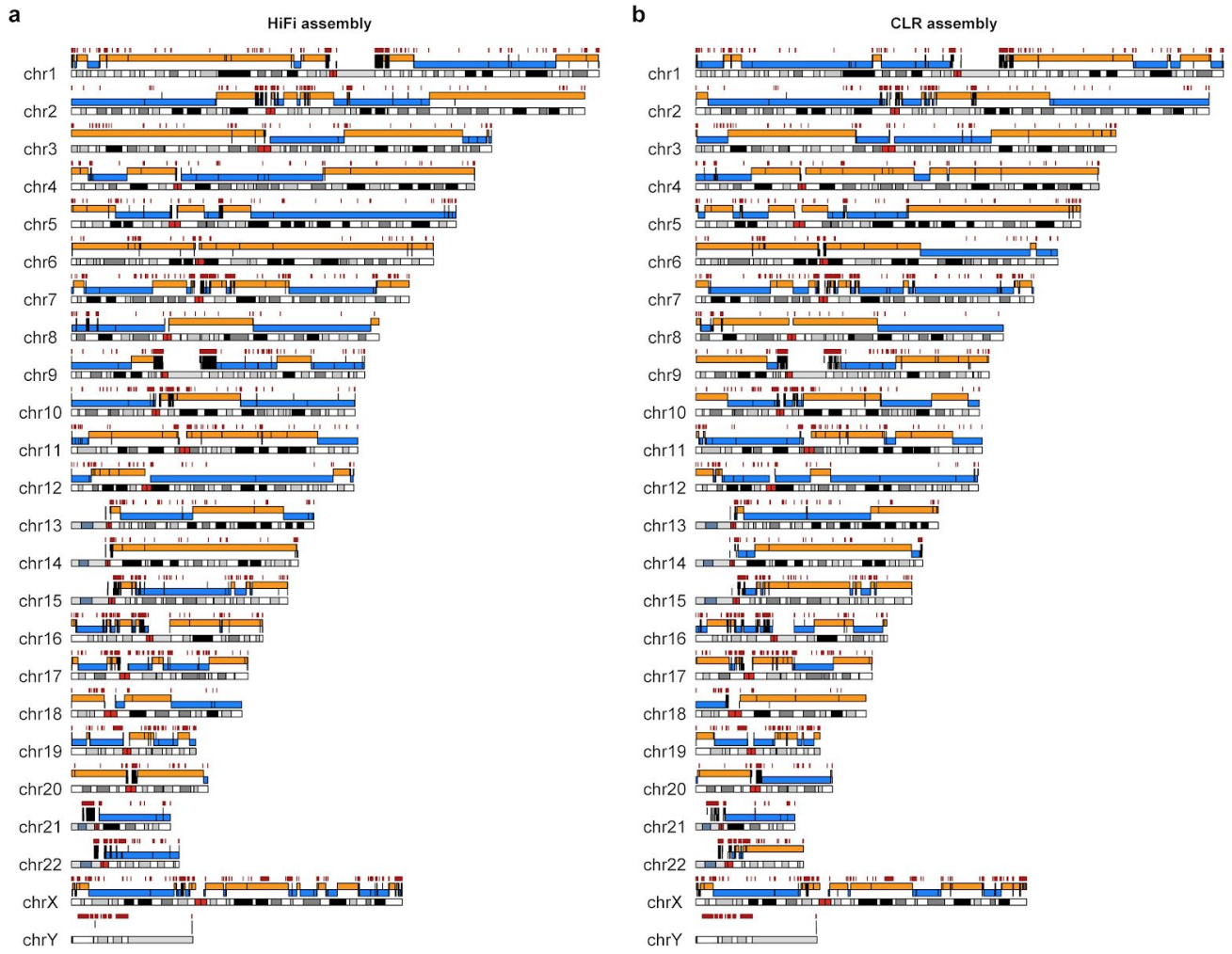

Figure 2. Comparison between the HiFi and CLR genome assemblies. Shown are alignments (blue and orange) of the HiFi assembly (**a**) and the CLR assembly (**b**) to GRCh38, as well as SD blocks greater than 10 kbp in length (red). The alignments are colored by contig name such that when the contig name changes, so does the alignment color. Black bars within a solid color block represent a break in the alignment within the same contig name. These are likely to be locations of structural variants between CHM13 and GRCh38. The large majority of contigs alignments over 100 kbp in length end within 50 kbp of a SD [158/166 (95%) in HiFi and 177/182 (97%) in CLR].

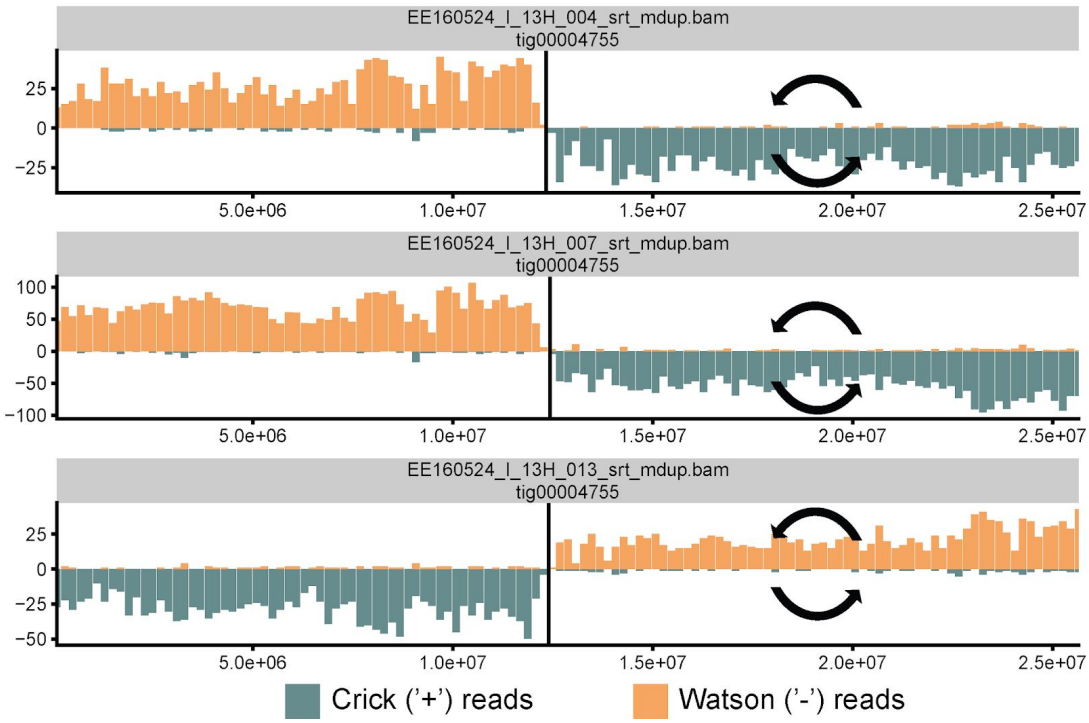

Figure 3. Example of a misjoined contig in the HiFi assembly. Reads mapped to the plus (Crick; teal) or minus (Watson; orange) strand of the reference genome are plotted as vertical bars along the contig. Each row shows one Strand-seq library. A recurrent change in read directionality in the middle of the contig suggests that left and right portions of this contig have flipped orientation with respect to each other and have likely been misjoined during the assembly process.

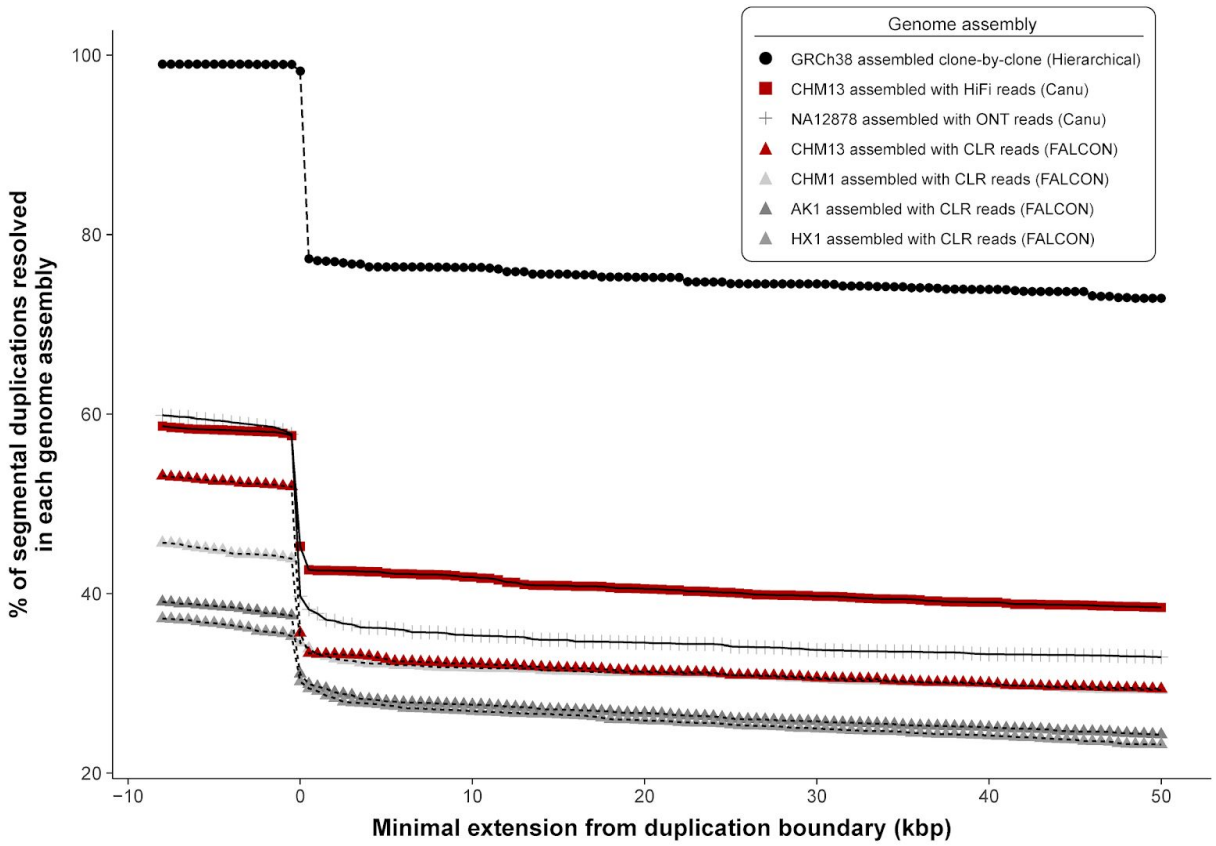

Figure 4. Percentage of GRCh38 SDs resolved across assemblies. Shown is the percent of resolved SDs as defined in GRCh38 across different *de novo* assemblies. To be considered resolved, the alignment of the *de novo* assembly must extend some number of base pairs beyond the annotated duplication block (x-axis) on either side. GRCh38 is not 100% resolved after a minimum extension of zero base pairs because many SDs in GRCh38 are flanked by gaps.

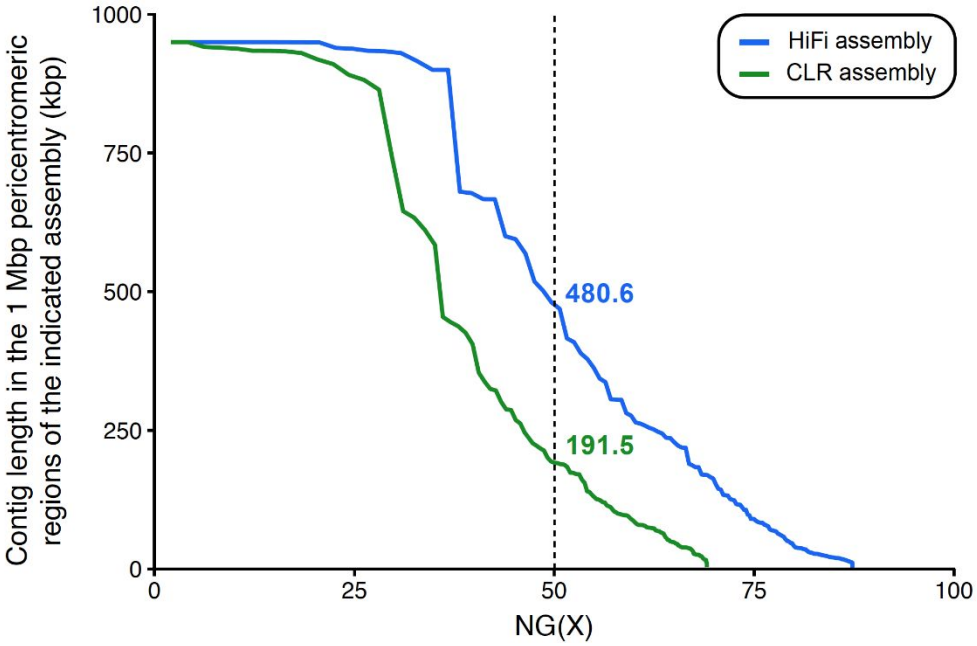

Figure 5. Increased contiguity and representation in the pericentromeric regions of the HiFi assembly. Plot showing the NG(X) of the HiFi and CLR assemblies in the 1 Mbp regions flanking the centromeres. NG(X) is defined as the sequence length of the shortest contig at X% of the total pericentromeric region length, which is 46 Mbp (1 Mbp for each pericentromere). The HiFi assembly has an NG50 2.5-fold greater than the CLR assembly in these regions.

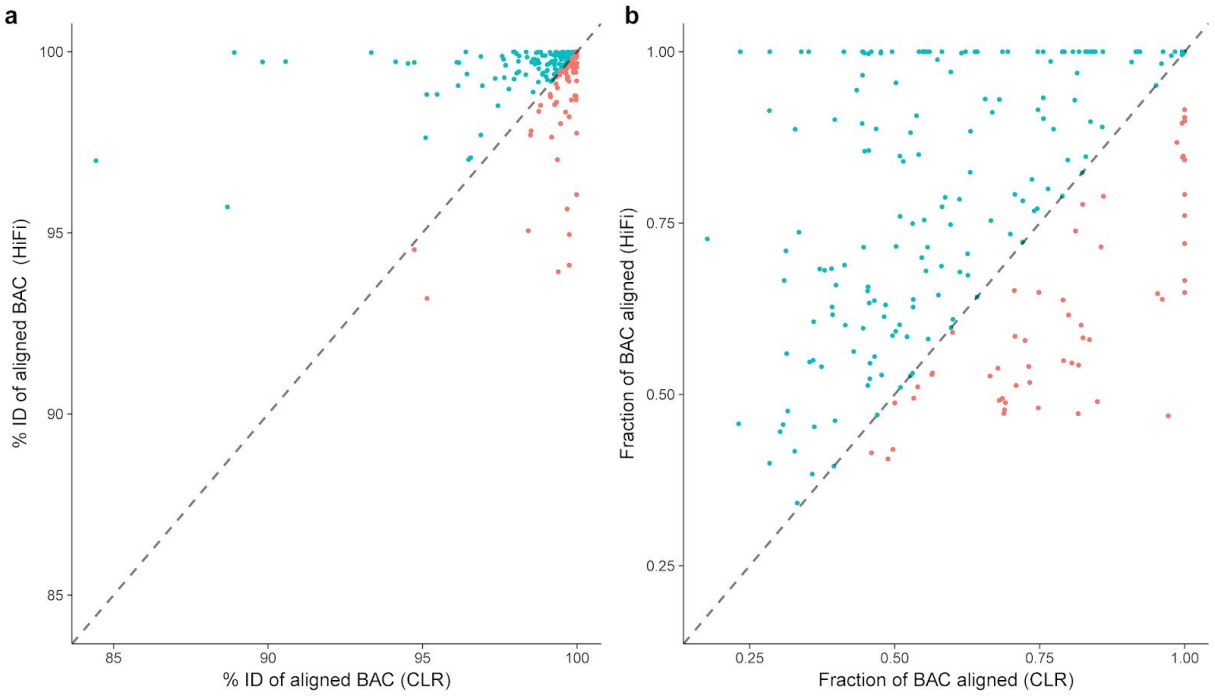

Figure 6. Evaluation of continuity and accuracy of segmental duplications. For 310 BACs aligning to SDs, we show the percent identity of the aligned sequence to the HiFi assembly (y-axis) and CLR assembly (x-axis) (**a**). Similarly, we show the fraction of the BAC that can be aligned to the HiFi assembly (y-axis) and the CLR assembly (x-axis) (**b**). Points above the diagonal (blue) have better identity (**a**) or aligned fraction (**b**) to the HiFi assembly, whereas points below the diagonal (red) show a better match to the CLR assembly. In 252 of the 310 (82%) BACs, the alignment length to the HiFi assembly is greater than or equal to the alignment length to the CLR assembly. Additionally the accuracy of the HiFi assembly (median QV 34.0) is increased compared to the CLR assembly (median QV 31.2) although both are significantly reduced compared to unique regions of the CHM13 genome.

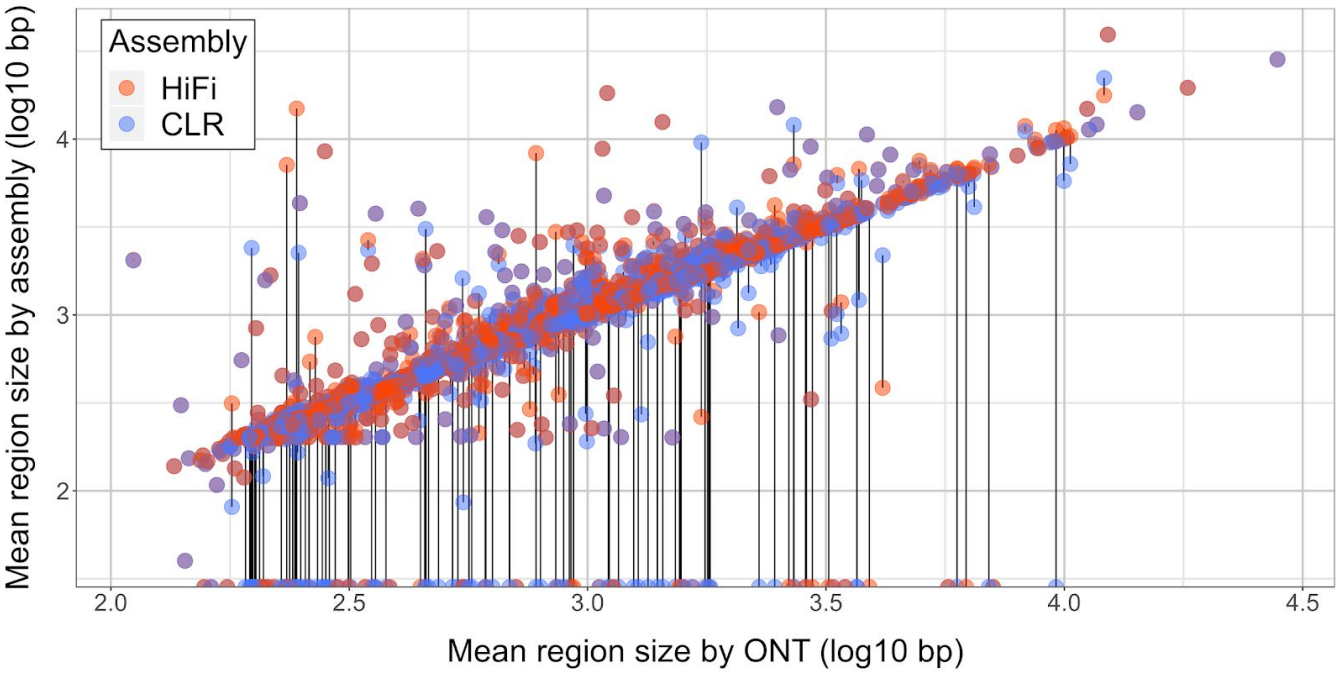

Figure 7. Tandem repeat length resolution. The length of tandem repeat loci in each assembly (vertical axis, red = HiFi, blue = CLR) is compared to the predicted size by ONT sequencing (horizontal axis). Discordancy between HiFi and CLR map off the diagonal, with dropouts clustering as points along on the horizontal axis. More dropouts occur in the CLR vs. HiFi assembly. For this plot, we include only regions with more than one spanning ONT read and no more than one spanning contig in either assembly (n = 2,900 regions).

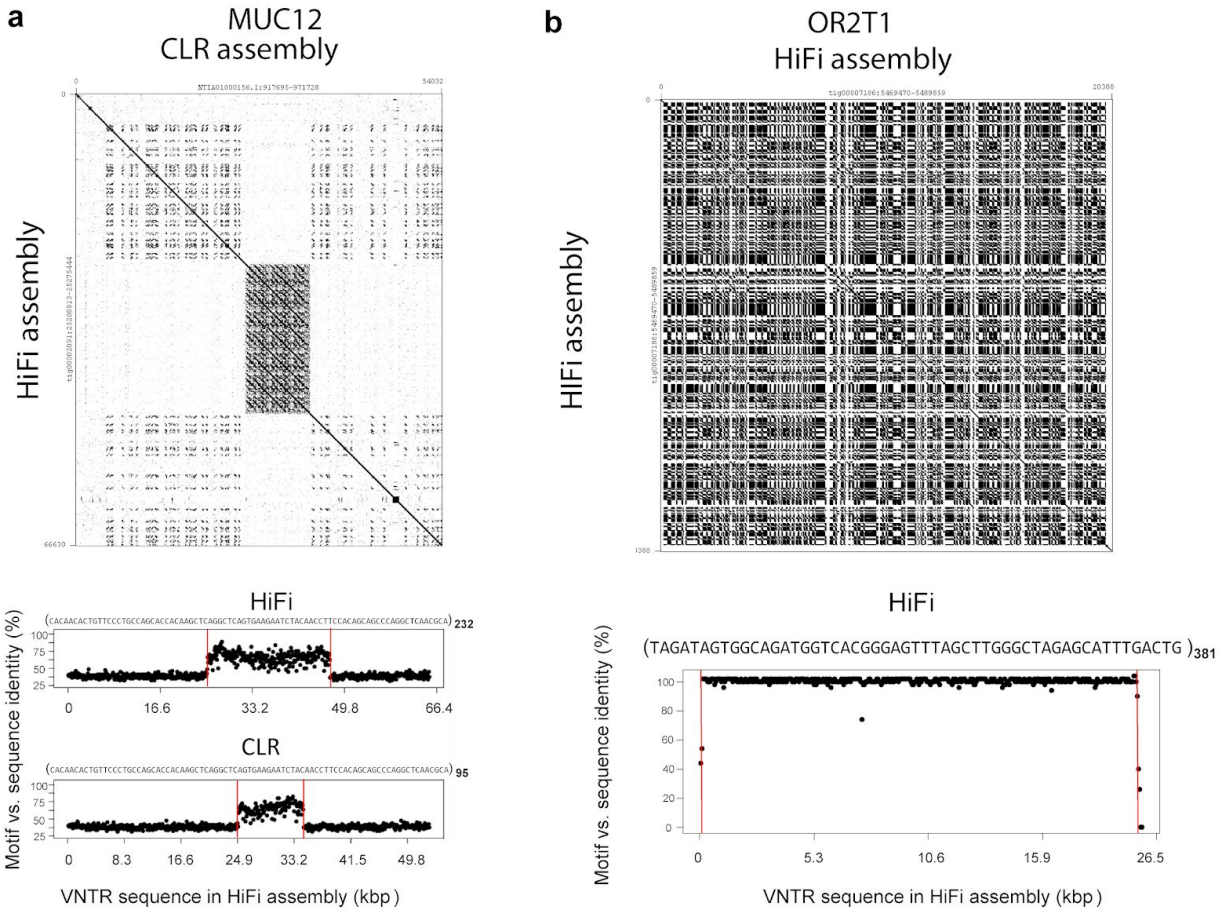

Figure 8. **Genic VNTR expansions.** Two loci that differed between CLR and HiFi assemblies were selected for in-depth analysis of their repeat content. **a)** Mucin gene *MUC12* (GRCh38 coordinates: chr7:100966000-101024949) contains a large VNTR in its largest exon, which was collapsed in the CLR assembly relative to HiFi. Specifically, the number of tandem repeat copies of the 83 bp repeat unit increased from 95 to 232 tandem copies in the HiFi assembly. The large amount of variation between tandem copies of this VNTR is shown in the dotplot as well as in sequence between the red lines in the motif homology plots. **b)** A 20 kbp VNTR in the 5' UTR of *OR2T1* (chr1:248407967-248409225) was resolved in the HiFi assembly only, while the CLR assembly contained a gap over this region. The overall structure and length of this VNTR was supported by the ONT reads mapping to this location which averaged 20,013 +/- 470 bp (n = 10 ultra-long ONT reads). Interestingly, the motif homology plots indicate that the content of the *OR2T1* VNTR is relatively pure, with 92% of the tandem copies >92% identical to the repeat unit sequence shown at the top of each plot. The read vertical lines indicate the start and end position of the VNTR. The dotplot in panel B contains 200 bp of non-repetitive sequence in both flanking regions.

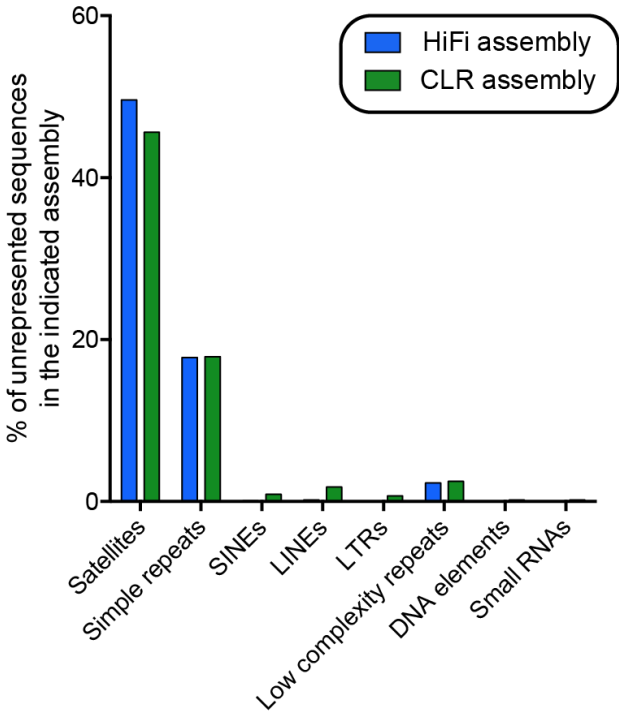

Figure 9. Repeat content of unassembled reads. Plot showing the repeat composition of sequences not incorporated into the HiFi and CLR assemblies. Most of the unrepresented sequences consist of satellite repeats mapping to heterochromatin or pericentromeric DNA (centromeres, acrocentric DNA and secondary constrictions of chromosomes).

TABLES

Table 1. Statistics of the HiFi and CLR genome assemblies.

Genome assembly	Total size (Gbp)	N50 (Mbp)	Median QV	No. of contigs	No. of CPU hours for assembly
HiFi assembly	3.03	29.52	34.43	5,206	~2,800
HiFi + Arrow polish	3.03	29.52	44.28	5,206	~10,000
HiFi + Racon polish	3.03	29.52	44.95	5,206	~2,950
HiFi + 2 x Racon polish	3.03	29.52	44.95	5,206	~3,100
CLR assembly	2.88	29.26	27.49	1,916	>50,000
CLR + Quiver	2.88	29.26	40.73	1,916	>55,000
CLR + Quiver + Illumina	2.88	29.26	42.70	1,916	>55,000

Median QV: Median QV over 31 BACs.

Table 2. False joins identified by Strand-seq within *de novo* assembly contigs.

	Contig name	Start	End	No. of false joins
HiFi				
	tig00000013	39053067	39099948	1
	tig00004755	12224208	12401058	1
	tig00005189	22935488	22975507	1
	tig00021540	923701	984042	1
	tig00003880	3053003	3203850	1
	tig00003880	10092230	10210470	1
	tig00002677	2544892	2612057	1
	tig00006896	801381	805905	1
CLR				
	NTIA01000004.1	51797117	51869304	1
	NTIA01000039.1	16068231	16072573	1
	NTIA01000061.1	12087358	12166683	1
	NTIA01000067.1	4639556	4640945	1
	NTIA01000093.1	746665	783795	1

Table 3. Comparison of PSV linking with SDA in HiFi and CLR.

Locus	Expected no. of paralogs	Expected no. of phased bases (kbp)	No. of paralogs by SDA (HiFi, CLR)	No. of phased bases by SDA (HiFi, CLR) (kbp)	Average length of SDA phased block (HiFi, CLR) (kbp)	No. of PSVs (HiFi, CLR)
OPN1LW	2	70	0, 1	0, 20	0, 20	0, 14
NOTCH2NL	5	500	9, 4	443, 345	49, 86	728, 482
SRGAP2	4	520	5, 5	493, 494	99, 99	1194, 821
FCGR2/3	3	220	3, 4	141, 140	47, 35	611, 139
KANSL1	2	280	4, 3	54, 150	13, 50	48, 92

Where applicable, results are shown for the HiFi data first and then the CLR data (i.e., HiFi, CLR).

Table 4. Summary of SV calls against GRCh38 post- and pre-polishing.

	Insertions			Deletions			All		
Polishing	**N**	**Mean**	**kbp**	**N**	**Mean**	**kbp**	**N**	**Mean**	**kbp**
HiFi									
None	10,727	571	6,123	6,266	482	3,017	16,993	538	9,141
Racon x2	10,700	572	6,116	6,272	479	3,002	16,972	537	9,118
Arrow	10,673	574	6,121	6,270	485	3,042	16,943	541	9,163
CLR									
None	10,655	558	5,948	6,405	471	3,019	17,060	526	8,966
Quiver	10,664	559	5,960	6,497	476	3,095	17,161	528	9,055
Quiver+	10,627	560	5,951	6,992	469	3,276	17,619	524	9,227

Racon x2: Two rounds of Racon.
Quiver+: Quiver, Pilon, and FreeBayes based indel correction.
Mean: Mean SV length.
kbp: Sum of all SV lengths.
SV: indel ≥ 50 bp
Excludes SVs mapping to pericentromeric regions (see **Methods**).

Table 5. Summary of indels against GRCh38 post- and pre-polishing.

	Insertions				Deletions			
Polishing	N	1 bp	Mean	kbp	N	1 bp	Mean	kbp
HiFi								
None	1,144,318	82%	1.81	2,068	1,558,924	88%	1.60	2,491
Racon x2	343,951	50%	3.44	1,183	340,148	48%	3.68	1,253
Arrow	340,247	50%	3.46	1,176	345,766	48%	3.66	1,265
CLR								
None	934,936	75%	1.99	1,861	3,616,964	82%	1.46	5,272
Quiver	350,924	50%	3.41	1,196	509,229	62%	2.87	1,460
Quiver+	353,245	50%	3.39	1,198	392,657	52%	3.41	1,338

Racon x2: Two rounds of Racon.
Quiver+: Quiver, Pilon, and FreeBayes based indel correction.
Mean: Mean indel length.
kbp: Sum of all indel lengths.

Table 6. Number of disrupted RefSeq gene models.

Polishing	All	No TR/SD	SD
HiFi			
None	4,756	2,835	426
Racon	146	61	36
Racon x2	126	53	34
Arrow	128	54	35
CLR			
None	11,591	6,685	1,249
Quiver	647	260	157
Quiver+	202	57	99

All: All variants within the assembly consensus regions. Total gene count is 18,037 (HiFi) and 17,991 (CLR).
No TR/SD: Excluded genes with exons intersecting SDs or tandem repeats. Total gene count is 10,858 (HiFi) and 10,850 (CLR).
SD: Only genes with exons intersecting SDs. Total gene count is 1,988 (HiFi) and 1,951 (CLR).
Racon x2: Two rounds of Racon.
Quiver+: Quiver, Pilon, and FreeBayes based indel correction.

ACKNOWLEDGEMENTS

The authors thank T. Brown for assistance in editing this manuscript. This work was supported, in part, by grants from the U.S. National Institutes of Health (NIH grants HG002385 and HG010169 to E.E.E.) and an Advanced Grant from the European Research Council (P.M.L.). M.R.V. was supported by a National Library of Medicine (NLM) Big Data Training Grant for Genomics and Neuroscience (5T32LM012419-04). A.S. was supported by a National Human Genome Research Institute (NHGRI) Training Grant (5T32HG000035-23). E.E.E. is an investigator of the Howard Hughes Medical Institute.

CONTRIBUTIONS

M.R.V., G.A.L., and E.E.E. wrote the manuscript; M.R.V., G.A.L, P.A.A, A.S, and D.P. produced the display items; M.R.V. performed the assembly and polishing with suggestions from Z.N.K. and A.M.W.; M.R.V. and A.M.W. performed the QV analysis; D.P. performed the Strand-seq analysis; M.R.V. and G.A.L. performed the SD analyses; A.S. and P.A.A. performed the tandem repeat analysis; P.A.A performed the structural variant and gene annotation analyses; G.A.L. performed the unassembled sequence analysis; M.R.V. and G.A.L. organized the supplementary material; P.P., G.T.C, K.M.M., C.B., and M.W.H generated the PacBio genome sequence data; U.S. developed and supplied the homozygous CHM13hTERT cell line.

CONFLICTS OF INTEREST STATEMENT

E.E.E. is on the scientific advisory board (SAB) of DNAnexus, Inc. and was a SAB member of Pacific Biosciences, Inc. (2009–2013). P.P., A.M.W., G.T.C., Z.N.K., and M.W.H. are employees and shareholders of Pacific Biosciences.

REFERENCES

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., Cantsilieris, S., Sorensen, M., Welch, A.E., Dougherty, M.L., Nelson, B.J., Shah, A., Dutcher, S.K., et al. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. Cell *176*, 663–675.e19.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Bickhart, D.M., Rosen, B.D., Koren, S., Sayre, B.L., Hastie, A.R., Chan, S., Lee, J., Lam, E.T., Liachko, I., Sullivan, S.T., et al. (2017). Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. Nat. Genet. *49*, 643–650.

Chaisson, M.J.P., Huddleston, J., Dennis, M.Y., Sudmant, P.H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., et al. (2015). Resolving the complexity of the human genome using single-molecule sequencing. Nature *517*, 608–611.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner, E.J., Rodriguez, O.L., Guo, L., Collins, R.L., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. *10*, 1784.

Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., et al. (2016). Phased diploid genome assembly with Single Molecule Real-Time Sequencing. Nat. Methods *13*, 056887.

Falconer, E., Hills, M., Naumann, U., Poon, S.S.S., Chavez, E.A., Sanders, A.D., Zhao, Y., Hirst, M., and Lansdorp, P.M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. Nat. Methods *9*, 1107–1112.

Gordon, D., Huddleston, J., Chaisson, M.J.P., Hill, C.M., Kronenberg, Z.N., Munson, K.M., Malig, M., Raja, A., Fiddes, I., Hillier, L.W., et al. (2016). Long-read sequence assembly of the gorilla genome. Science *352*, aae0344.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., Sudmant, P.H., Graves, T.A., Alkan, C., Dennis, M.Y., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. *24*, 688–696.

Huddleston, J., Chaisson, M.J.P., Steinberg, K.M., Warren, W., Hoekzema, K., Gordon, D., Graves-Lindsay, T.A., Munson, K.M., Kronenberg, Z.N., Vives, L., et al. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. *27*, 677–685.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation.

Genome Res. 1–33.

Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., and Phillippy, A.M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol. *36*, 1174–1182.

Kronenberg, Z.N., Fiddes, I.T., Gordon, D., Murali, S., Cantsilieris, S., Meyerson, O.S., Underwood, J.G., Nelson, B.J., Chaisson, M.J.P.P., Dougherty, M.L., et al. (2018). High-resolution comparative analysis of great ape genomes. Science *6343*.

Kronenberg, Z.N., Rhie, A., Koren, S., Concepcion, G.T., Peluso, P., Munson, K.M., Hiendleder, S., Fedrigo, O., Jarvis, E.D., Phillippy, A.M., et al. (2019). Extended haplotype phasing of de novo genome assemblies with FALCON-Phase.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics *23*, 1026–1028.

Li, H. (2017). Minimap2: pairwise alignment for nucleotide sequences. *34*, 3094–3100.

Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., and MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat. Methods *15*, 595–597.

Low, W.Y., Tearle, R., Bickhart, D.M., Rosen, B.D., Kingan, S.B., Swale, T., Thibaud-Nissen, F., Murphy, T.D., Young, R., Lefevre, L., et al. (2019). Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. Nat. Commun. *10*, 260.

Sanders, A.D., Falconer, E., Hills, M., Spierings, D.C.J., and Lansdorp, P.M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat. Protoc. *12*, 1151–1176.

Seo, J.-S.S., Rhie, A., Kim, J.J.J.J., Lee, S., Sohn, M.-H.H., Kim, C.C.-U.U., Hastie, A., Cao, H., Yun, J.-Y.Y., Kim, J.J.J.J., et al. (2016). De novo assembly and phasing of a Korean human genome. Nature *538*, 243–247.

She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpern, A.L., and Eichler, E.E. (2004). Shotgun sequence assembly and recent segmental duplications within the human genome. Nature *431*, 927–930.

She, X., Liu, G., Ventura, M., Zhao, S., Misceo, D., Roberto, R., Cardone, M.F., Rocchi, M., NISC Comparative Sequencing Program, Green, E.D., et al. (2006). A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. Genome Res. *16*, 576–583.

Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., Fu, A., Li, Q., Li, N., Gong, S., et al. (2016). Long-read sequencing and de novo assembly of a Chinese genome. Nat. Commun. *7*, 12065.

Smit, A.F.A., Hubley, R., and Green, P. (1996). RepeatMasker.

Steinberg, K.M., Graves-Lindsay, T., Schneider, V.A., Chaisson, M.J.P., Tomlinson, C., Huddleston, J.L., Minx, P., Kremitzki, M., Albrecht, D., Magrini, V., et al. (2016). High-quality assembly of an individual of yoruban descent. bioRxiv 067447.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly

from long uncorrected reads. Genome Res. *27*, 737–746.

Vollger, M.R., Dishuck, P.C., Sorensen, M., Welch, A.E., Dang, V., Dougherty, M.L., Graves-Lindsay, T.A., Wilson, R.K., Chaisson, M.J.P., and Eichler, E.E. (2019). Long-read sequence and assembly of segmental duplications. Nat. Methods *16*, 88–94.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One *9*.

Watson, M., and Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. Nat. Biotechnol. *37*, 124–126.

Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Highly-accurate long-read sequencing improves variant detection and assembly of a human genome.

METHODS

CCS library preparation. High-molecular-weight DNA was isolated from cultured CHM13 cells using a modified Qiagen Gentra Puregene Cell Kit protocol (Huddleston et al., 2014). A HiFi library with an average insert length of ~11 kbp was generated according to the protocol in Wenger et al., 2019 and sequenced on four SMRT Cells 8M on a Sequel II instrument using Sequel II Sequencing Chemistry 1.0, 12 hour pre-extension, and 30 hour movies. Raw data was processed using the CCS algorithm (version 3.4.1, parameters: `--minPasses 3 -- minPredictedAccuracy 0.99 --maxLength 21000`) to yield 75.7 Gbp in 6.9 million reads with an average read length of 10.95 kbp and estimated median quality value of 32.85. Sequence data is available via NCBI SRA (https://www.ncbi.nlm.nih.gov/sra/SRX5633451). Average run time for the CCS algorithm was ~12,500 CPU core hours per SMRT Cell (~50,000 total).

Strand-seq library preparation. Cultured CHM13 cells were pulsed with BrdU and used for preparation of single-cell Strand-seq libraries as described (Sanders et al., 2017).

BAC clone insert sequencing. BAC clones from the VMRC59 clone library were hybridized with probes targeting complex or highly duplicated regions of GRCh38 (n = 310), or selected from random regions of the genome not intersecting with SD (n = 31). DNA from positive clones was isolated, screened for genome location, and prepared for long-insert PacBio sequencing as previously described (Vollger et al., 2019). Libraries were sequenced on the PacBio RS II and Sequel platforms with the P6-C4 or Sequel2.1/Sequel3.0 chemistries, respectively. We performed *de novo* assembly of pooled BAC inserts using Canu v1.5 (Koren et al., 2017). After assembly, we removed vector sequence (pCCBAC1), restitched the insert, and then polished with Quiver or Arrow. Canu is specifically designed for assembly with long error-prone reads, whereas Quiver/Arrow is a multi-read consensus algorithm that uses the raw pulse and base call information generated during SMRT sequencing for error correction. We reviewed PacBio assemblies for misassembly by visualizing the read depth of PacBio reads in Parasight (http://eichlerlab.gs.washington.edu/jeff/parasight/index.html), using coverage summaries generated during the resequencing protocol.

Genome assembly. Canu version 1.7.1 was applied with the following parameters to generate the HiFi *de novo* assembly: `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`.

Assemblies were mapped to GRCh38 with minimap2 (Li, 2017) version 2.15 using the following parameters: `--secondary=no -a --eqx -Y -x asm20 -m 10000 -z 10000,50 -r 50000 --end-bonus=100 -O 5,56 -E 4,1 -B 5`. These alignments were used for downstream SV calling and ideogram visualizations.

Error correction with Quiver, Arrow, Pilon, and indel correction was done as previously described (Chin et al., 2013; Kronenberg et al., 2018; Vaser et al., 2017; Walker et al., 2014). Error correction with Racon was executed with the following steps:

```
minimap2 -ax map-pb --eqx -m 5000 -t {threads} --secondary=no {ref}
{fastq} | samtools view -F 1796 - > {sam}
```

```
racon {fastq} {sam} {ref} -u -t {threads} > {output.fasta}
```

QV calculations. QV calculations were made by alignments to 31 sequenced and assembled BACs falling within unique regions of the genome (> 10 kbp away from the closest SD) where at least 95% of the BAC sequence was aligned. The following formula was used to calculate QV, and gaps of size N were counted as N errors:

$$QV = -10\, log_{10}(\, 1 - percent\ identity/100\,)$$

QV calculations within SDs were done in the same manner but against 310 BACs that overlap with SD regions.

Segmental duplication analyses. SDs were defined as resolved or unresolved based on their alignments to GRCh38 using the minimap2 parameters described above. Alignments that extended a minimum number of base pairs beyond the annotated SDs were considered to be resolved. This minimum extension varied from -10,000 to 50,000 bp and the average difference between assemblies was used to define the percent difference reported.

The number of collapsed bases was determined by aligning the CLR reads to both the CLR and the HiFi assemblies. Regions were defined as collapsed if they met the following conditions: coverage greater than the mean coverage plus three standard deviations, 15 kbp of consecutive increased coverage or more, and less than 80% repeat content as defined by RepeatMasker.

Pericentromeric analyses. The number of contigs within each pericentromeric region was calculated by first aligning the contigs from the HiFi or CLR assemblies to GRCh38 using the minimap2 parameters described above. Alignments were limited to be within 1 Mbp on either side of the centromere decoys and then unique contigs names were counted.

The representation within the pericentromeric regions was calculated using bedtools to collapse all filtered contigs within the pericentromeric region for the HiFi and CLR assemblies. The resulting size of the collapsed contigs within the CLR assembly was subtracted from the size calculated in the corresponding region in the HiFi assembly.

The pericentromere-specific NG50 statistic was calculated using a G of 46 Mbp (accounting for the 1 Mbp size of each pericentromeric region on the 23 chromosomes).

Tandem repeat analyses. Tandem Repeat Finder (Benson, 1999) was run on the six haplotype-resolved assemblies (Chaisson et al., 2019) as well as the CLR CHM13 assembly using the following parameters: `2 7 7 80 10 50 2000 -h -d -ngs`. After identifying all tandem repeats not represented or collapsed in CLR relative to the six human haplotypes, we obtained a final set of 3,074 large tandem repeats, all of which were anchored in GRCh38. Second, we retrieved sequence from each of these loci using the two assemblies and our orthogonal CHM13 ONT data source. For each region in both assemblies and aligned ultra-long ONT reads, we extracted the sequence that mapped from the start of the region to the end using the alignment CIGAR strings as a guide. Since multiple sequences may map to a region, we recorded the number of alignments and computed the

average length of the region for each dataset. Concordance with ONT reads was defined by allowing ≤5% variation in the average ONT read length. For our in-depth sequence analysis of the two VNTR loci, we used repeat homology plots, which were constructed using a pairwise alignment between the motif and assembled sequence in every tiling window of the same length as the repeat unit length (i.e., 53 bp and 83 bp, respectively, for the two VNTRs in **Figure 8**). At any given window, the repeat unit (i.e., the motif) was circularized in 1 bp increments, and the maximal sequence identity was reported at each tiling window. The dotplots were generated using Gepard (Krumsiek et al., 2007).

Structural variant analysis. For assembly in each polishing stage, contigs mapped to GRCh38 were used to create a consensus region, which included all loci with exactly one aligned contig. Next, we called indels and SVs from the alignments using a previously validated method (Chaisson et al., 2015) implemented in PrintGaps.py distributed in the SMRT-SV v2 pipeline (https://github.com/EichlerLab/smrtsv2). We then filtered for variants within the assembly's consensus region. We further filtered out variants in pericentromeric loci where callsets are difficult to reproduce (Audano et al., 2019). This process was repeated for each assembly in each polishing stage.

Gene annotation. With custom code using the SV and indel callset, the number of bases in coding regions of RefSeq annotations (retrieved 2019-04-24 from UCSC RefSeq track on GRCh38) were quantified. Briefly, if an insertion was located in a coding region, its entire length was taken as the number of coding bases it affects. For deletions, the number of bases falling inside the coding region were quantified. From these results, we obtained a set of genes where at least one variant inserts or deletes a number of bases that is not a multiple of three within any isoform of the gene. For this analysis, we excluded RefSeq ncRNA annotations.

We intersected RefSeq exons with tandem repeats (TRs) (UCSC hg38 "Simple Repeats" track) and SDs (UCSC hg38 "Segmental Dups" track) to annotate them as either containing or absent of SDs or TRs. For each assembly, we calculated results using only RefSeq genes that are fully contained within its consensus region.

RepeatMasker analysis of unmappable sequences. All HiFi sequence reads were mapped to the *de novo* assemblies using the following minimap2 parameters: `-x asm20 -m 4000 --secondary=no --paf-no-hit`. Reads that that did not map to the *de novo* assemblies were subjected to RepeatMasker analysis (Smit et al., 1996) to determine their repeat content.

SUPPLEMENTAL FIGURES

Figure S1. The HiFi assembly only has one mismatch error across 31 BACs. Across the 31 BACs, we found only one mismatched base when compared to the HiFi assembly. At this base, both the HiFi data and Illumina data support the HiFi assembly, indicating 100% accuracy within this assembly excluding indels.

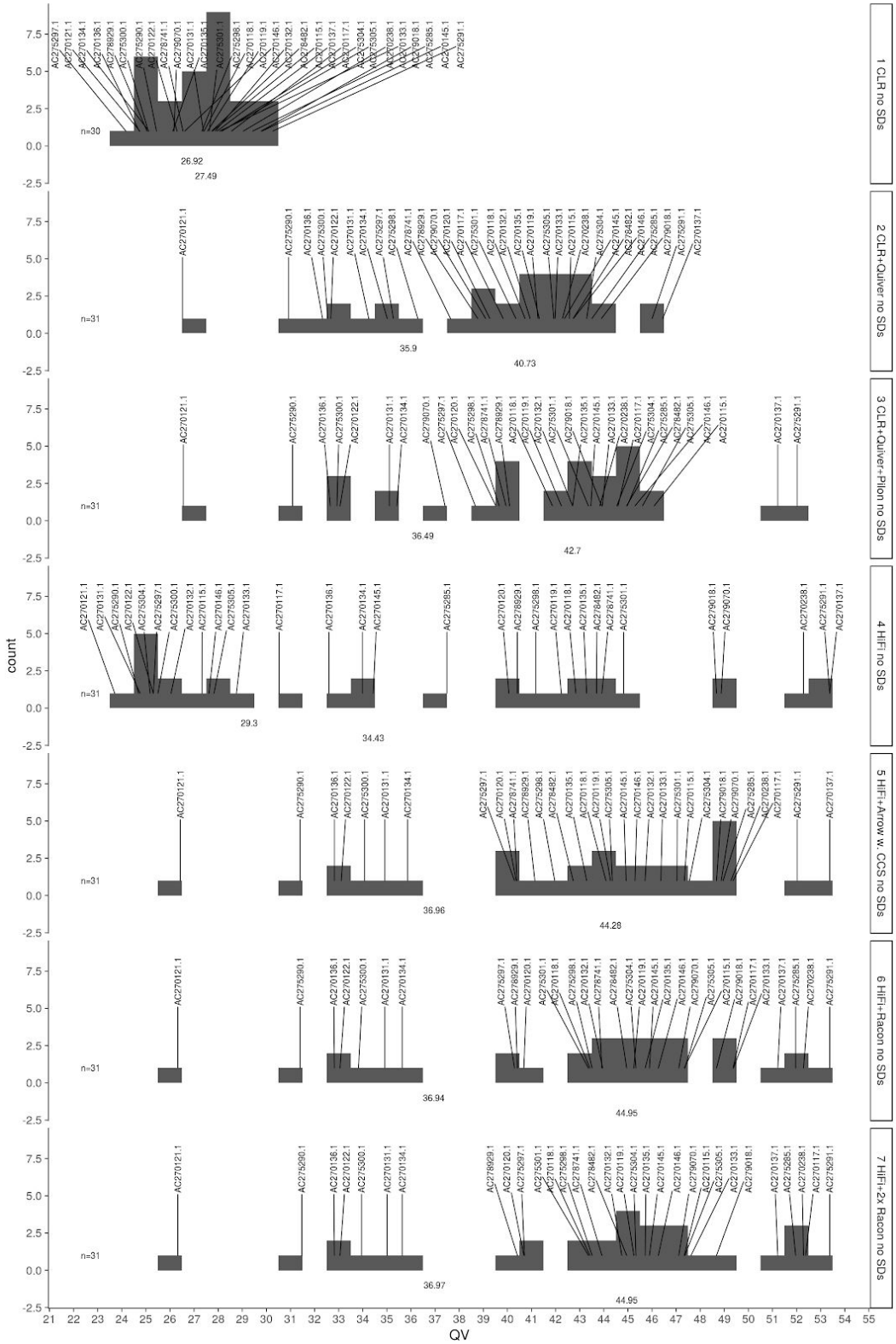

Figure S2. Histogram of QVs across polishing steps and assemblies for 31 BACs. Shown in text are the number of BACs displayed, the mean QV, the median QV, and the accession names for all the BACs.

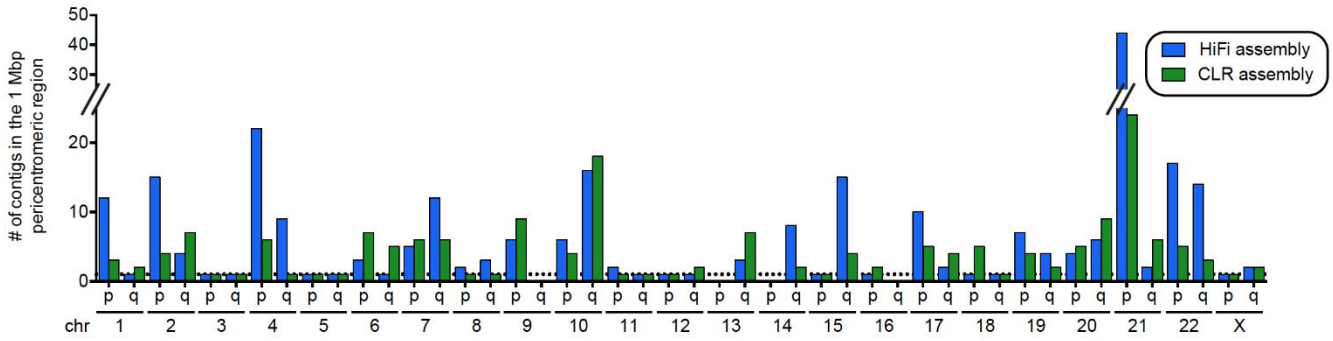

Figure S3. Assessment of continuity in the pericentromeric regions in the HiFi and CLR assemblies. Plot showing the number of contigs in the 1 Mbp regions flanking each centromere in the HiFi and CLR assemblies; 54.3% of the pericentromeric regions in the HiFi assembly contained either a reduced number of contigs or the same number of contigs. The remaining pericentromeric regions either contained no contig (8.7%) or an increased number of contigs (40%) in the HiFi assembly relative to the CLR assembly.

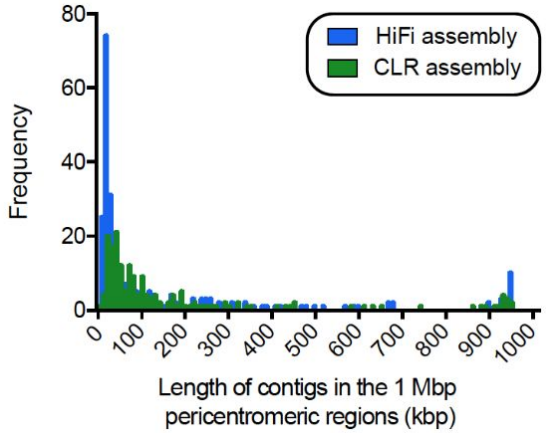

Figure S4. Assessment of contig length in the pericentromeric regions of HiFi and CLR assemblies. Histogram of the length of contigs in the 1 Mbp region flanking the centromeres for each assembly. The HiFi assembly has more contigs than the CLR assembly overall, with an increase in the number of small contigs (<100 kbp) and large contigs (900-1000 kbp). The average contig length is 289.6 kbp in the HiFi assembly and 283.3 kbp in the CLR assembly.

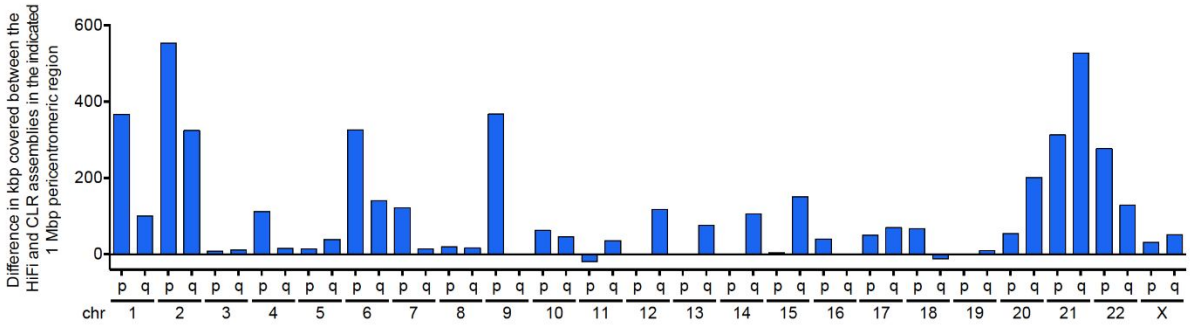

Figure S5. The HiFi assembly contains an additional 5 Mbp of pericentromeric sequence missing in the CLR assembly. Plot showing the difference in sequence coverage in the 1 Mbp regions flanking each centromere for the HiFi and CLR genome assemblies. Nearly all pericentromeric regions contain additional sequences in the HiFi assembly relative to the CLR assembly.

SUPPLEMENTAL NOTES

Polishing the assemblies. Initially, the HiFi assembly was polished with Racon using the approximate alignments from minimap2 (i.e. the PAF output generated using the `-x asm5` option) and fasta input of HiFi reads. However, we found that this polishing step only modestly increased the QV (by <0.01). When we polished the HiFi assembly with the exact alignments (i.e. the SAM output generated using the `-ax map-pb` option) and fastq input, we observed a large increase in the median QV (from 34.4 to 45.0). In addition, we observed that the QV achieved using these Racon parameters was greater than that achieved with Arrow (which used > 1TB of CCS subreads). It is, therefore, our suggestion to polish assemblies generated with HiFi data with two rounds of Racon using the parameters described above rather than with Arrow.

BAC divergence. In all of our polished assemblies (HiFi or CLR; Table 1), we noticed that the same two BACs (AC270121.1 and AC275290.1) had the lowest QV values of those assessed (**Figure S2**). We examined the alignments of these BACs to all the assemblies and to the HiFi reads and found that these BACs had contractions in tandem repeats relative to the CHM13 cell line. In AC270121.1, there was a 338 bp deletion of a (TCCCCC)n repeat, and in AC275290.1, there was a 80 bp deletion in a (GGCTGAGG)n repeat. In addition, AC270136.1 showed a 62 bp expansion in a poly(T) tract, where the HiFi data supported the HiFi assembly, and AC270122.1 showed a 83 bp insertion, where both Illumina and HiFi data supported the HiFi assembly. Across all 31 BACs used for calculating QV, there was only one mismatched base (AC275285.1:148688-148688). This base appears to be correct in the HiFi assembly, as it was observed in both the HiFi and Illumina data (**Figure S1**). In combination, these results indicate that many of the BACs with QV < 40 are diverged in sequence when compared to the CHM13 genome due to a mutation that likely arose during BAC generation and/or clonal propagation and do not represent an error in the assemblies. For this reason, our QV values should be interpreted as a lower bound of the true QV.