Supplemental material for Vollger, Logsdon et al:

Supplemental Inventory

Supplemental Figure Legends

Figure S1. Distribution of the CHM13 HiFi and CLR read lengths and quality values (QVs).

A) Histogram of the CHM13 HiFi (blue) and CLR (green) read lengths. The read N50 of each dataset is shown.

B) Histogram of the CHM13 HiFi (blue) and CLR (green) read QVs. QVs were estimated by aligning reads from each dataset to the curated, telomere-to-telomere assembly of the CHM13 X chromosome (Miga et al., 2019) and counting the differences in the alignments as errors in the reads. More than half (54.6%) of the HiFi reads are QV 30 or greater.

C) Table listing the CHM13 HiFi and CLR read QV and accuracy. Mean and median values were calculated by aligning reads from each dataset to the CHM13 telomere-to-telomere X chromosome assembly (Miga et al., 2019). Values were calculated with and without indels to demonstrate the high indel error rate in CLR data.

Figure S2. Assessment of QV score of each genome assembly with varying levels of polishing.

The QV score histogram was derived from the alignment of 31 BACs to the indicated assembly. Each BAC clone accession name is indicated, and the overall and median QV scores for each genome assembly are shown.

Figure S3. Assessment of a base mismatch between a BAC clone and the HiFi assembly.

Out of all 31 BACs assessed, there was a single mismatch observed between one BAC and the HiFi assembly (highlighted here in red and boxed with a dashed line). However, both the HiFi and Illumina data supported the base in the HiFi assembly (‘C’ rather than ‘A’), indicating a divergence in sequence between BAC clone AC275285.1 and the CHM13 cell line (see **Supplemental Notes**).

Figure S4. Example of a misjoined contig in the HiFi assembly.

Shown is an example of a misjoined contig in the HiFi assembly**.** Reads mapping 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 S5. Assessment of continuity in the pericentromeric regions in the HiFi and CLR assemblies.

A) Plot of the number of contigs in the 1 Mbp regions flanking each centromere in the HiFi and CLR assemblies. The majority of the pericentromeric regions in the HiFi assembly (52.2%) 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 (39.1%) in the HiFi assembly relative to the CLR assembly.

B) Histogram of the length of contigs in the 1 Mbp regions 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 145.8 kbp in the HiFi assembly and 177.6 kbp in the CLR assembly.

C) Plot of 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. The HiFi assembly contains an additional 5.03 Mbp of pericentromeric sequence missing in the CLR assembly.

Figure S6. SVs discovered in the HiFi assembly are supported by published CHM13 calls.

We intersected SVs with published CHM13 SVs excluding tandem repeat and segmental duplication (SD) loci, where variant comparisons are more challenging. Both insertions (left) and deletions (right) are strongly supported. Each Venn area is annotated with the total number of variants (n) along with the

mean and median variant size, respectively, in brackets. For both insertions and deletions, the HiFi assembly calls more variants around 700 bp to 1 kbp, but the published variants have more calls in the 50-100 bp range as well as more larger calls (10+ kbp). These disagreements are reflected in the mean and medians, and they may be due to assembly errors or differences in mapping and assembly methods used in each study.

Figure S7. Disrupted genes in the HiFi assembly supported by the CLR assembly.

For all loci where the polished HiFi and CLR assemblies had a single alignment to the reference (left), all but two genes disrupted by the HiFi assembly have support in the CLR assembly. We observe 50 genes disrupted in CLR without HiFi support (29% of CLR-disrupted genes). When we restrict our analysis to SDs (right), the percentage of genes disrupted in the CLR assembly without HiFi support increases to 59%.

Figure S8. Repeat content of unassembled reads.

Bar plot of 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). SINE, small interspersed nuclear element; LINE, long interspersed nuclear element; LTR, long terminal repeat.

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 40.4 to 45.0). In addition, we observed that the QV achieved using these Racon parameters was greater than that achieved with Arrow (which used >1 TB of CCS subreads). Polishing a second time with Racon further increased the QV and significantly reduced the number of gene-disrupting indels. Adding Pilon polishing did not change the median QV but significantly reduced the total QV across all the BACs because it introduced a 660 bp insertion that appears to be an error relative to the AC275297.1 BAC. Additionally, Pilon polishing only reduced the number of indels genome wide by 645 out of 683,564 (0.094%) and resolved only one additional gene-disrupting event in unique sequence. 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 or Pilon.

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 (**Fig. 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 an 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 an 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 (**Fig. S3**). 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.

Additional assemblies. To determine whether the improvements in genome assembly quality observed in the HiFi assembly (**Figs. 1-3, Tables 1-4**) are due to the assembler or the data type, we performed two control experiments. First, we generated a HiFi assembly using FALCON rather than Canu (hereafter termed "HiFi,FALCON"); second, we also produced a CLR assembly using Canu rather than FALCON [using a downsampled CLR dataset that has the equivalent coverage as the HiFi dataset (24-fold rather than 77-fold); hereafter termed "CLR,Canu"] (**Table S1**). Our results suggest that the improvements in the HiFi genome assembly quality are due to the data type and not the assembler.

For the HiFi,FALCON assembly, we find that it is highly comparable to the HiFi,Canu assembly in size (3.00 Gbp vs. 3.03 Gbp), quality (median BAC QV of 44.45 vs. 45.25), and compute time (~4,400 CPU hours vs. ~2,800 CPU hours). The contiguity of the HiFi,FALCON assembly is improved compared to the HiFi,Canu assembly (N50 31.92 vs. 25.51) and slightly exceeds the CLR,FALCON assembly (N50 29.26). Additionally, error correction with Arrow on the HiFi,FALCON assembly resulted in a higher quality assembly than when Quiver was applied to the CLR,FALCON assembly (QV 43.54 vs. 40.73), and two rounds of polishing with Racon performs better than polishing with Arrow on the HiFi,FALCON assembly (QV 44.45 vs. 43.54). Quiver polishing is not supported for sequencing data generated on the Sequel II, so it was not possible to generate a direct comparison on Quiver polishing. Compute time for the HiFi,FALCON assembly (~4,400 CPU hours) is slightly longer than the HiFi,Canu assembly (~2,800 CPU hours); however, the HiFi,FALCON assembly is still more than ten times faster than the CLR,FALCON assembly (>50,000 CPU hours). Overall, the HiFi,FALCON and HiFi,Canu assemblies are similar in terms of size (3.00 vs. 3.03 Gbp), median BAC QV (44.45 vs. 45.25), compute time (~4,400 vs. ~2,800 CPU hours), and contiguity (31.92 vs. 25.51).

For the CLR,Canu assembly, we find that it is not comparable to the HiFi,Canu assembly, and this is mainly due to the data type used to generate the assembly. When CLR coverage is downsampled to

that of HiFi coverage (24-fold) and assembled with the same assembler as HiFi data (Canu), it produces a much smaller assembly (2.48 Gbp vs. 3.03 Gbp) with a much lower contiguity (N50 0.31 Mbp vs. 25.51) and many more contigs (21,267 vs. 5,296) while taking much longer compute time (~37,300 vs. ~2,800 CPU hours). Additionally, the CLR,Canu assembly has a median BAC QV that is much lower than the HiFi,Canu assembly (26.50 vs. 40.41). Overall, we find that the downsampled CLR,Canu assembly is much lower quality than the HiFi,Canu assembly, and this is largely due to the data type.