nature methods

ARTICLES

https://doi.org/10.1038/s41592-022-01440-3

Check for updates

Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies

Ann M. Mc Cartney[1,17], Kishwar Shafin[2,17], Michael Alonge[3,17], Andrey V. Bzikadze[4], Giulio Formenti[5], Arkarachai Fungtammasan[6], Kerstin Howe[7], Chirag Jain[1,8], Sergey Koren[1], Glennis A. Logsdon[9], Karen H. Miga[2,10], Alla Mikheenko[11], Benedict Paten[2], Alaina Shumate[12], Daniela C. Soto[13], Ivan Sović[14,15], Jonathan M. D. Wood[7], Justin M. Zook[16], Adam M. Phillippy[1]✉ and Arang Rhie[1]✉

Advances in long-read sequencing technologies and genome assembly methods have enabled the recent completion of the first telomere-to-telomere human genome assembly, which resolves complex segmental duplications and large tandem repeats, including centromeric satellite arrays in a complete hydatidiform mole (CHM13). Although derived from highly accurate sequences, evaluation revealed evidence of small errors and structural misassemblies in the initial draft assembly. To correct these errors, we designed a new repeat-aware polishing strategy that made accurate assembly corrections in large repeats without overcorrection, ultimately fixing 51% of the existing errors and improving the assembly quality value from 70.2 to 73.9 measured from PacBio high-fidelity and Illumina *k*-mers. By comparing our results to standard automated polishing tools, we outline common polishing errors and offer practical suggestions for genome projects with limited resources. We also show how sequencing biases in both high-fidelity and Oxford Nanopore Technologies reads cause signature assembly errors that can be corrected with a diverse panel of sequencing technologies.

Genome assembly is a foundational practice of quantitative biological research with increasing utility. By representing the genomic sequence of a sample of interest, genome assemblies enable researchers to annotate important features, quantify functional data and discover/genotype genetic variants in a population[1–6]. Modern draft eukaryotic genome assembly graphs are typically built from a subset of four whole-genome shotgun (WGS) sequencing data types: Illumina short reads[7,8], Oxford Nanopore Technologies (ONT) long reads[9,10], PacBio continuous long reads (CLRs) and PacBio high-fidelity (HiFi) long reads[9,11], all of which have been extensively described[7–9,11]. However, we note that even the high-accuracy technologies produce sequencing data with some noise caused by platform-specific technical biases that require careful validation and polishing[1,11–14].

Current genome assembly software attempts to reconstruct an individual or mosaic haplotype sequence from a subset of the above WGS data types. Some assemblers do not attempt to correct sequencing errors[15], while others attempt to remove errors at various stages of the assembly process[16–20]. Regardless, technology-specific sequencing errors usually lead to distinct assembly errors[14,21]. Additionally, suboptimal assembly of specific loci often causes small and large errors in draft assemblies[22,23]. Here, we define 'polishing' as the process of removing these errors from draft genome assemblies. Most polishing tools use an approach that is similar to sequence-based genetic variant discovery. Specifically, reads from the same individual are aligned to a draft assembly, and putative 'variant'-like sequence edits are identified[23,24]. For diploid genomes, heterozygous 'alternate' alleles are interpreted as genuine heterozygous variants, while homozygous alternate alleles are interpreted as assembly errors to be corrected. Some polishing tools, such as Quiver/Arrow, Nanopolish, Medaka, DeepVariant and PEPPER leverage specialized models and previous knowledge to correct errors caused by technology-specific bias[25–29]. Others, such as Racon[30], use generic methods to correct assembly errors with a subset of sequencing technologies[30–32]. These generic tools can utilize multiple data types to synergistically overcome technology-specific assembly errors.

The telomere-to-telomere (T2T) consortium recently convened an international workshop to assemble the first-ever complete sequence of a human genome. Because heterozygosity can complicate assembly algorithms, the consortium chose to assemble the highly homozygous genome of a complete hydatidiform mole

[1]Genome Informatics Section, Computational and Statistical Genomics Branch, NHGRI, NIH, Bethesda, MD, USA. [2]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA. [3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [4]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA. [5]Laboratory of Neurogenetics of Language and The Vertebrate Genome Lab, The Rockefeller University, New York, NY, USA. [6]DNAnexus, Mountain View, CA, USA. [7]Wellcome Sanger Institute, Cambridge, UK. [8]Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India. [9]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [10]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA. [11]Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, Saint Petersburg State University, Saint Petersburg, Russia. [12]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [13]Genome Center, MIND Institute, Department of Biochemistry and Molecular Medicine, University of California, Davis, CA, USA. [14]Pacific Biosciences, Menlo Park, CA, USA. [15]Digital BioLogic d.o.o., Ivanić-Grad, Croatia. [16]Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA. [17]These authors contributed equally: Ann M. Mc Cartney, Kishwar Shafin, Micheal Alonge. ✉e-mail: adam.phillippy@nih.gov; arang.rhie@nih.gov

cell line (CHM13hTERT, hereafter CHM13). Primarily using HiFi reads and supplemented with ONT reads, the consortium built a highly accurate and complete draft assembly (CHM13v0.9) that resolved all repeats with the exception of the ribosomal RNA genes (rDNAs)[1]. CHM13v0.9 contained about one error in every 10.5 Mb (Q70.22), and while this was highly accurate by traditional standards, we, as part of the consortium, sought to correct all lingering errors and omissions, including those within repeats, in this first truly complete assembly of a human genome.

Alignment-based validation and polishing commonly underperform within genomic repeats where alignments are ambiguous and inaccurate. For example, this challenge was identified while validating the first complete centromere and satellite repeats of the X chromosome, requiring a customized conservative marker-assisted alignment[33]. To address this challenge, specialized repeat-aware alignment methods were recently developed, such as Winnowmap2 (refs. [34,35]) and TandemMapper[36]. However, to the best of our knowledge, no studies have utilized such methods to reliably validate and polish an entire genome assembly, including the most notoriously repetitive regions.

Here, we describe techniques developed to carefully evaluate the accuracy and completeness of a complete human genome assembly using multiple complementary WGS data types. Our evaluation of the initial draft CHM13 assembly discovered a number of assembly errors; therefore, we created a custom polishing pipeline that was robust to genomic repeats and technology-specific biases. By applying this polishing pipeline to CHM13v0.9, we made 1,457 corrections, replacing a total of 12,234,603 bp of sequence with 10,152,653 bp of sequence, ultimately leading to the landmark CHM13v1.1 assembly representing the first complete human genome ever assembled. Our edits increased the estimated quality value (QV) to Q73.94 while mitigating haplotype switches. Further, we extended the truncated p arm of chromosome 18 to encompass the complete telomere, and polished all telomeres with a new specialized PEPPER-DeepVariant model. Our careful evaluation of CHM13v1.1 confirmed that polishing did not overcorrect repeats (including rDNAs) nor did it cause false-positive edits causing invalid coding sequence reading frames. Additionally, we identified a comprehensive list of putatively heterozygous loci in the CHM13 cell line, as well as sporadic loci where read alignments still indicated exceptionally low coverage. Finally, we uncovered common mistakes made by standard automated polishing pipelines and provide best practices for other genome assembly projects.

Results

Initial evaluation of CHM13v0.9. The T2T consortium has collected a comprehensive and diverse set of publicly available WGS sequencing and genomic map data (Illumina PCR-free, PacBio HiFi, PacBio CLR, ONT and Bionano optical maps) for the nearly completely homozygous CHM13 cell line (https://github.com/marbl/CHM13/). As part of the consortium, we drew upon these sequencing data to generate a custom pipeline (Fig. 1) to evaluate, identify and correct lingering errors in CHM13v0.9.

We first derived *k*-mer-based quality estimations ($k=21$ bp) of CHM13v0.9 using Merqury[37] with both Illumina and HiFi reads. The *k*-mer size was chosen to limit the collision rate to 0.5% given the estimated genome size of 3.05 Gb of CHM13 (ref. [38]). While estimating the Illumina reads QVs, we found 15,723 *k*-mers present in the assembly and not the reads (erroneous *k*-mers), leading to an estimated base quality of Q66.09. Using HiFi reads, we found 6,881 error *k*-mers (Q69.68; Fig. 2a). To test how technical sequencing bias may have influenced this QV estimation, we examined the *k*-mer multiplicity and sequence content of assembly *k*-mers absent from one technology but present in the other. Here, our results indicated that *k*-mers missing from Illumina reads were present with expected frequency in HiFi and were enriched for G/C bases. Conversely, *k*-mers missing in HiFi were present with higher frequency in Illumina reads with A/T base enrichment (Fig. 2b). However, we identified no particular enrichment pattern in the number of GAs or CTs within the *k*-mers, possibly due to the short *k*-mer size chosen (Extended Data Fig. 1a). Most of the *k*-mers absent from HiFi reads were located in patches derived from a previous ONT-based assembly (CHM13v0.7)[33], which were included to overcome regions of HiFi coverage dropout[1] (Extended Data Fig. 1b,c). These findings highlighted that platform-specific sequencing biases were underestimating the QV when measured from a single sequencing platform. To overcome this, we created a hybrid *k*-mer database that combined these platforms to be used for QV estimation (Extended Data Fig. 1d). Unlike the default QV estimation in Merqury, we removed low-frequency *k*-mers to avoid overestimated QVs caused by excessive noise accumulated from both platforms. We estimated base-level accuracy as Q70.22 with 6,073 missing *k*-mers (Table 1). We note that this estimate does not account for the rarer case of *k*-mers present in the reads but misplaced or falsely duplicated in the assembly.

Despite the high accuracy of CHM13v0.9 (Q70.22), we expected to find consensus sequence errors related to the systematic presence of homopolymer-specific or repeat-specific issues in HiFi reads[9,39]. To detect these, we generated self-alignments by aligning CHM13 reads to CHM13v0.9 for each WGS sequencing technology. Although each data type required technology-specific alignment methods (Methods), we highlight our use of Winnowmap2 that enabled robust alignment of long reads to both repetitive and non-repetitive regions of CHM13v0.9 (refs. [34,35]). To understand the homopolymer length differences between the assembly and the reads, we derived a confusion matrix from Illumina read alignments showing discordant representation of long homopolymers between the Illumina reads and the assembly (Fig. 2c). Altogether, the QV and homopolymer analysis suggested that CHM13v0.9 required polishing to maximize the accuracy of a complete human genome.

Identification and correction of assembly errors. To address assembly flaws identified during evaluation, we aimed to establish a customized polishing pipeline that would avoid false-positive polishing edits (especially in repeats) and maintain local haplotype consistency (Fig. 1b and Extended Data Fig. 2). We identified and corrected small errors (≤50 bp) using several small-variant calling tools from self-alignments of Illumina, HiFi and ONT reads to CHM13v0.9. To call both single-nucleotide polymorphisms (SNPs) and small insertions and deletions (INDELs), we applied a hybrid mode of DeepVariant[27] that exploited both HiFi and Illumina read alignments[40]. Simultaneously, we used PEPPER-DeepVariant[28] to generate additional SNP calls with ONT reads as it can yield high-quality SNP variants in difficult regions of the genome[28,40] (see Supplementary Fig. 1 in ref. [28] for more details). We rigorously filtered all calls using a genotype quality (GQ) score (GQ < 30 for the hybrid calls and GQ < 25 for ONT SNP calls) and variant allele frequency (VAF < 0.5) to exclude any low-frequency false-positive calls (Extended Data Fig. 2). We chose VAF < 0.5 to avoid including heterozygous variants, and the GQ threshold was chosen based on the previously reported calibration plot of DeepVariant, which shows that calls that have quality scores above 25 or 30 are highly unlikely to result in false positives[27,28]. We then filtered all of the suggested alternate corrections with Merfin[41], a tool concurrently developed by members of the T2T consortium, to avoid introducing error *k*-mers (Figs. 1b and 3c). Finally, we ignored variants near the distal or proximal rDNA junctions on the short arms of the acrocentric chromosomes to avoid homogenizing the alleles from the unassembled rDNAs. After merging all variant calls, we identified 993 small variants (≤50 bp) that represented potential assembly errors and heterozygous sites. From these 993 assembly edits, about two-thirds were homopolymer corrections (512)

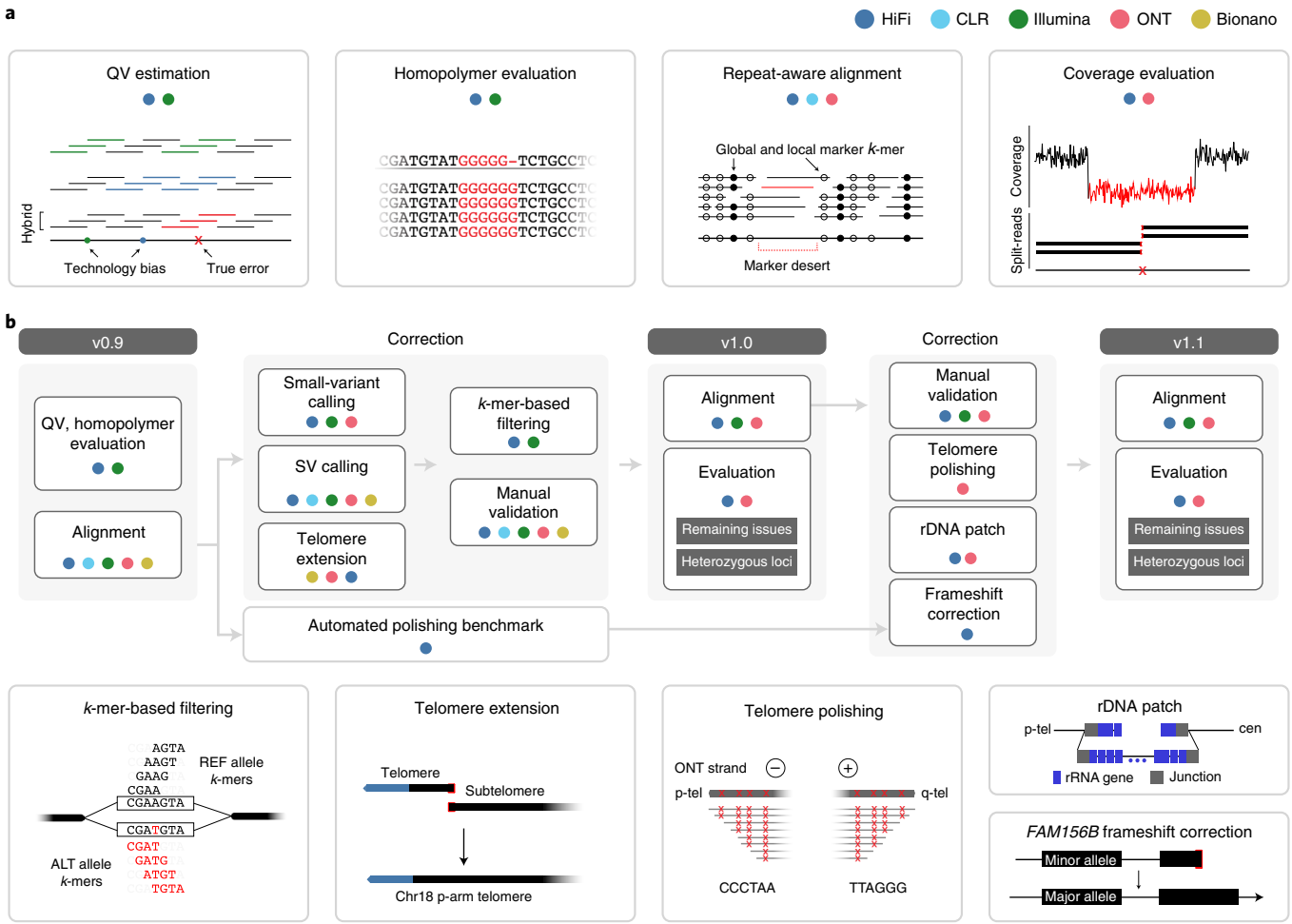


Fig. 1 | An overview of the evaluation and polishing strategy developed to achieve a complete human genome assembly. a, The evaluation strategies used to assess genome assembly accuracy before (CHM13v0.9) and after (CHM13v1.0 and CHM13v1.1) polishing. **b**, The 'do no harm' polishing strategy developed and implemented to generate CHM13v1.0 and CHM13v1.1.

or low-complexity microsatellite repeats composed of two distinct bases in homopolymer-compressed space (hereby noted as '2-mer') consistent with previous observations of HiFi sequence errors or bias[17]. Across all 617 loci, we evaluated the edit distribution using both Illumina and HiFi reads and found that the majority of Illumina reads supported the longer homopolymer or 2-mer repeat lengths compared to HiFi reads, thereby uncovering systemic biases in both homopolymer and 2-mer length in HiFi reads[17] that caused the propagation of these errors into the consensus assembly sequence (Fig. 3d).

We used Parliament2 (ref. [42]) and Sniffles[43] to identify medium-sized (>50 bp) assembly errors and heterozygous structural variants (SVs). Parliament2 runs six SV callers[42] using short-read data, while the Sniffles detects SVs using one of the long-read technologies (HiFi, ONT and CLR). To improve specificity, we only considered Parliament2 calls supported by at least two SV callers and Sniffles calls supported by at least two long-read technologies. Similarly to small-variant detection, we excluded SVs called in the partial rDNA arrays and the HSat3 satellite repeat on chromosome 9. This pipeline identified a relatively small number of SV calls (66; Extended Data Fig. 2) that we were able to manually curate via genome browsing. In total, we corrected three medium-sized assembly errors (replacing 1,998 bp of CHM13v0.9 sequence with 151 bp of new sequence), and we identified 44 heterozygous SVs (Fig. 3a and Extended Data Fig. 3b). We also identified a missing telomere sequence on the p arm of chromosome 18—a potential result of the string graph simplification process and confirmed through Bionano mapping (Figs. 1b and 3b). To correct this omission, we used the CHM13v0.9 graph to identify a set of HiFi reads expected to cover this locus[1] and found ONT reads that mapped to the corresponding subtelomere and contained telomeric repeats. We used the ONT reads to derive a consensus chromosome 18 extension that was subsequently polished with the associated HiFi reads. After patching this telomere extension, we used Bionano alignments to confirm the accuracy of this locus (Fig. 3b). Altogether, the small and medium-sized variant calls along with the chromosome 18 telomere patch were combined into two distinct VCF files: a polishing edits file (homozygous ALT variants and the telomere patch) and a file for heterozygous variants (all other variants). We created the polished CHM13v1.0 assembly by incorporating these edits into the CHM13v0.9 with bcftools[44].

We ensured polishing accuracy by extensive manual validation through visual inspection of the repeat-aware alignments, error *k*-mers, marker *k*-mers and marker-assisted alignments. Here, we defined 'marker' *k*-mers as *k*-mers that occur only once in the assembly and in the expected single-copy coverage range of the read *k*-mer database and are highly likely to represent unique regions of the assembly (Extended Data Fig. 3b–d)[33]. To generate marker-assisted alignments, we filtered Winnowmap2 (ref. [34]) alignments to exclude any alignments that did not span marker *k*-mers

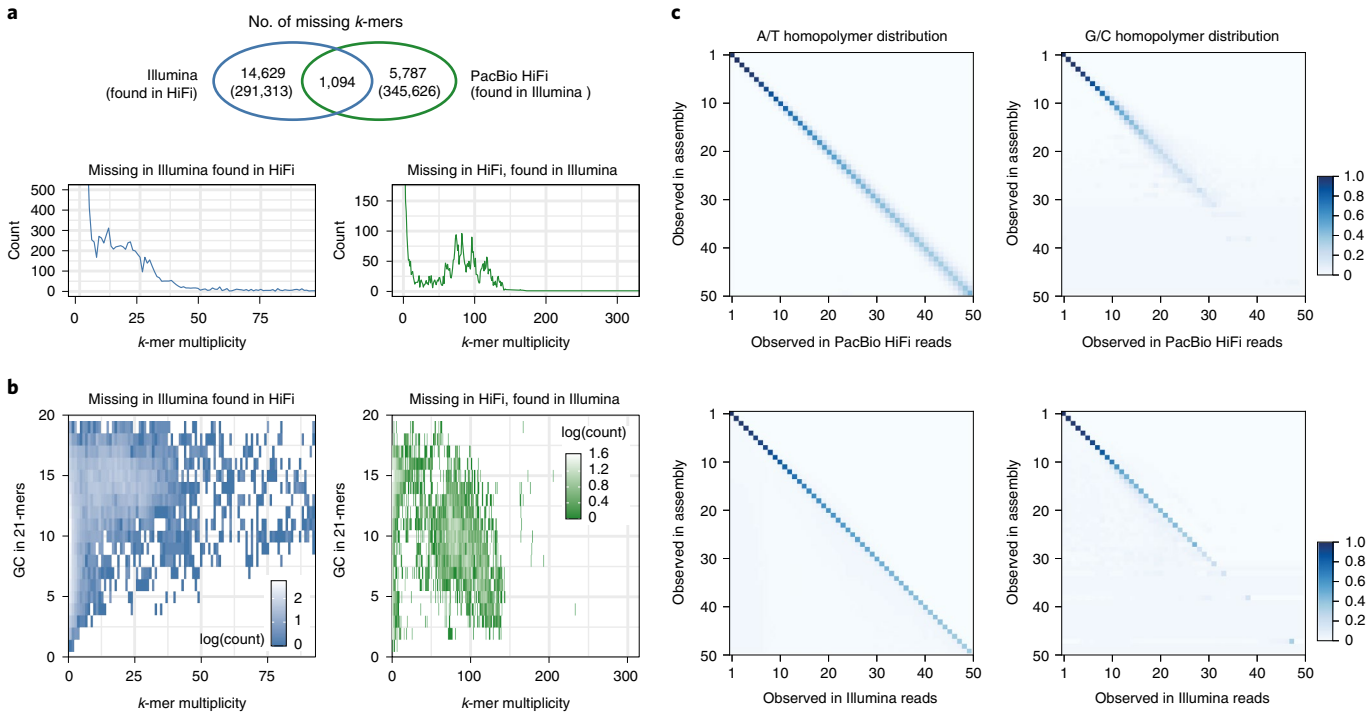


Fig. 2 | Sequencing biases in PacBio HiFi and Illumina reads. a, Venn diagram of the 'missing' *k*-mers found in the assembly but not in the HiFi reads (green) or Illumina reads (blue). Except for the 1,094 *k*-mers that were absent from both HiFi and Illumina reads, error *k*-mers were found in the other sequencing platform with expected frequency, matching the average sequencing coverage (lower). **b**, Missing *k*-mers from **a** with their GC content, colored by the frequency observed. Low-frequency erroneous *k*-mers did not have a clear GC bias. *k*-mers found only in HiFi had a higher GC percentage, while higher-frequency *k*-mers tended to have more AT-rich sequences in Illumina. **c**, Homopolymer length distribution observed in the assembly and in HiFi reads (upper) or Illumina reads (lower) aligned to that position. Longer homopolymers in the consensus are associated with length variability in HiFi reads especially in the GC homopolymers. The majority of the Illumina reads were concordant with the consensus.

(https://github.com/arangrhie/T2T-Polish/tree/master/marker_assisted/). Our findings supported that most genomic loci contained a deep coverage of marker *k*-mers to facilitate marker-assisted alignment, except for a few highly repetitive regions (11.3 Mb in total) that lacked markers (termed 'marker deserts'; Fig. 1a and Extended Data Fig. 3c,d). In parallel, we used TandemMapper[36] to detect structural errors in all centromeric regions, including identified marker deserts. TandemMapper[36] used locally unique markers for the detection of marker order and orientation discrepancies between the assembly and associated long reads. We manually validated all large polishing edits and heterozygous SVs, and many small loci were validated ad hoc.

Validation of CHM13v1.0. Given the high completeness and accuracy standards of the T2T consortium, and knowing that polishing may introduce additional errors[41], we took extra precautions to validate polishing edits and to ensure that edits did not degrade the quality of CHM13v0.9. First, we repeated self-alignment variant calling methods on CHM13v1.0, confirming that all edits made were correct (Fig. 3a). Through Bionano optical map alignments, we validated the structural accuracy of the chromosome 18 telomere patch and confirmed that all 46 telomeres were represented in CHM13v1.0 (Fig. 3b). Notably, our polishing led to a marked improvement in the distribution of GQ and VAF of small-variant calls (Fig. 3c and Extended Data Fig. 4a). Our approach also increased the base-level consensus accuracy from Q70.22 in CHM13v0.9 to Q72.62 in CHM13v1.0. Further, we found that error *k*-mers were uniformly distributed along each chromosome, suggesting that remaining errors were not clustered within certain genomic regions (Extended Data Fig. 4b,c). Upon reevaluation of the homopolymers and 2-mers, most of the biases that we found in CHM13v0.9 from HiFi reads had been accurately removed, achieving an improved concordance with Illumina reads (Fig. 3d). Polishing did not induce invalid open reading frames (ORFs) in CHM13v0.9 transcripts with valid ORFs, and polishing corrected 16 invalid CHM13v0.9 ORFs (Supplementary Table 1).

Overall, we made a total of 112 polishing edits (impacting 267 bp) in centromeric regions[45], with 15 (35 bp) of these edits occurring specifically in centromeric alpha-satellite higher-order repeat arrays. We made 134 edits (4,975 bp) in non-satellite segmental duplications[2]. Moreover, the polishing edits were neither

Table 1 | *k*-mer-based consensus quality evaluation

	PacBio HiFi	Illumina	Hybrid
QV			
v0.9	69.68	66.09	70.22
v1.0	69.88	67.28	72.62
v1.1	69.80	67.86	73.94
k-mers found only in assembly (error *k*-mers)			
v0.9	6,881	15,723	6,073
v1.0	6,581	11,961	3,496
v1.1	6,724	10,497	2,591
k-mers found in both assembly and reads			
v0.9	3,045,438,411	3,045,438,411	3,045,438,411
v1.0	3,045,440,942	3,045,440,942	3,045,440,942

From each sequencing dataset and assembly versions, 21-mers were collected and compared with Merqury[37].

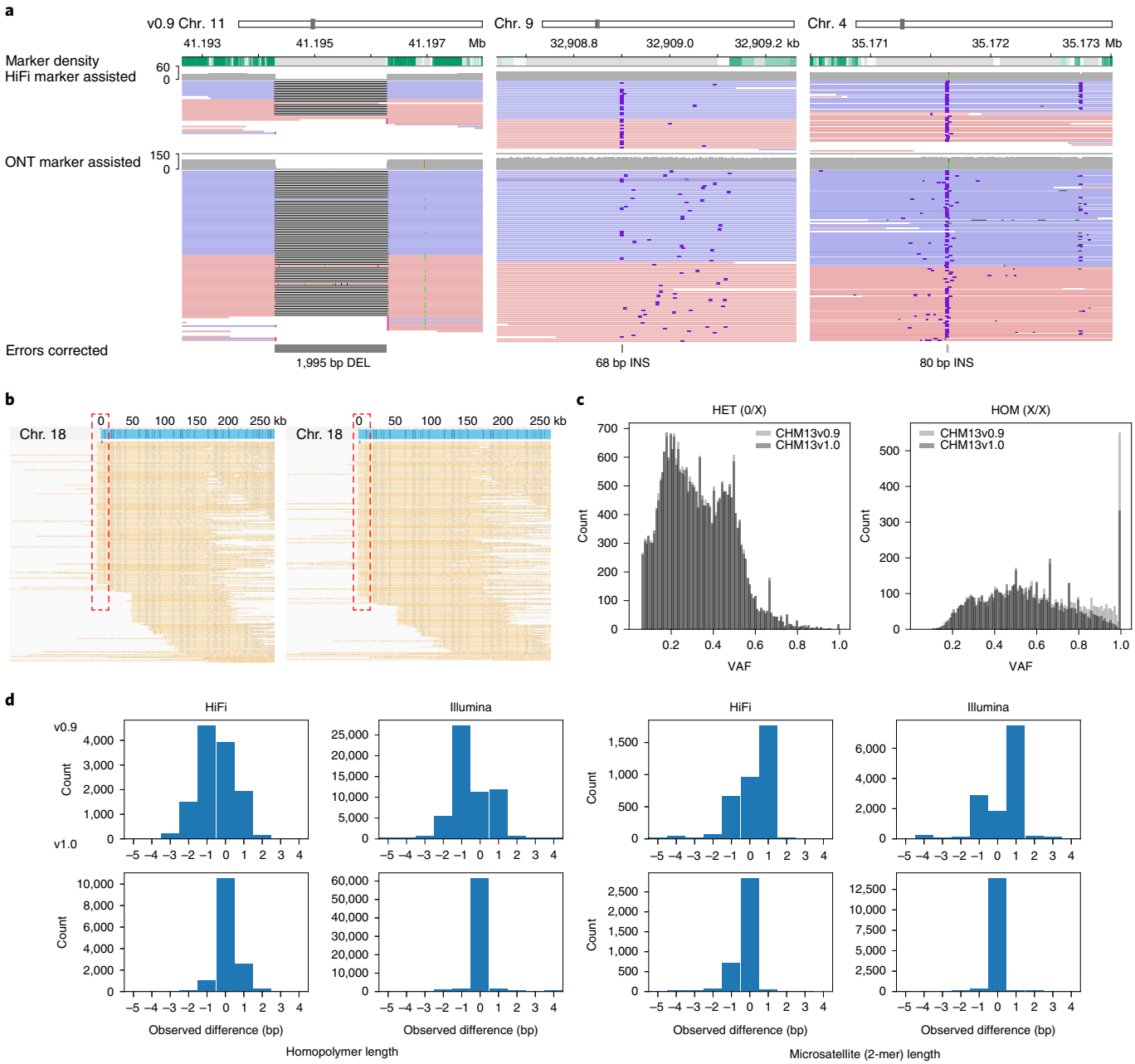


Fig. 3 | Errors corrected after polishing. a, Three corrected SV-like errors. **b**, Bionano optical maps indicating the missing telomeric sequence on the p arm of chromosome 18 (left) with a higher-than-average mapping coverage. This excessive coverage was removed after adding the missing telomeric sequence (right) and most of the Bionano molecules end at the end of the sequence. **c**, VAF of each variant called by DeepVariant hybrid (HiFi + Illumina) mode, before and after polishing. Most of the high-frequency variants (errors) were removed after polishing, which were called 'homozygous' variants. **d**, Total number of reads in each observed length difference (bp) between the assembly and the aligned reads at each edit position. Positive numbers indicate that more bases are found in the reads, while negative numbers indicate fewer bases in the reads. Both the homopolymer and microsatellite (2-mers in homopolymer-compressed space) length difference became 0 after polishing.

enriched nor depleted in satellite repeats and segmental duplications ($P = 0.85$, permutation test), suggesting that non-masked repeats were not overcorrected or undercorrected compared to the rest of the genome (Extended Data Fig. 4c). Finally, through extensive manual inspection, we confirmed the reliability of the alignments for the three SV-associated edits incorporated into CHM13v1.0 (Extended Data Fig. 5), and these efforts uncovered some heterozygous loci in the centromeres. These regions are under active investigation by the T2T consortium to both ensure their structure and understand their evolution[45].

As an additional validation, we investigated potential rare or false collapses as well as rare or false duplications in CHM13v1.0. Here, based on *k*-mer estimates from both GRCh38 and CHM13v1.0 and from Illumina reads for 268 Simon's Genome Diversity Project (SGDP) samples, we identified regions in CHM13v1.0 with a lower or higher copy number than both GRCh38 and 99% of the SGDP samples[2]. We found six regions of rare collapses in CHM13v1.0 that were not in GRCh38 (covering 205 kb, four from one single segmental duplication family). Both our HiFi read-depth and Illumina *k*-mer-based copy number estimates suggest these six regions

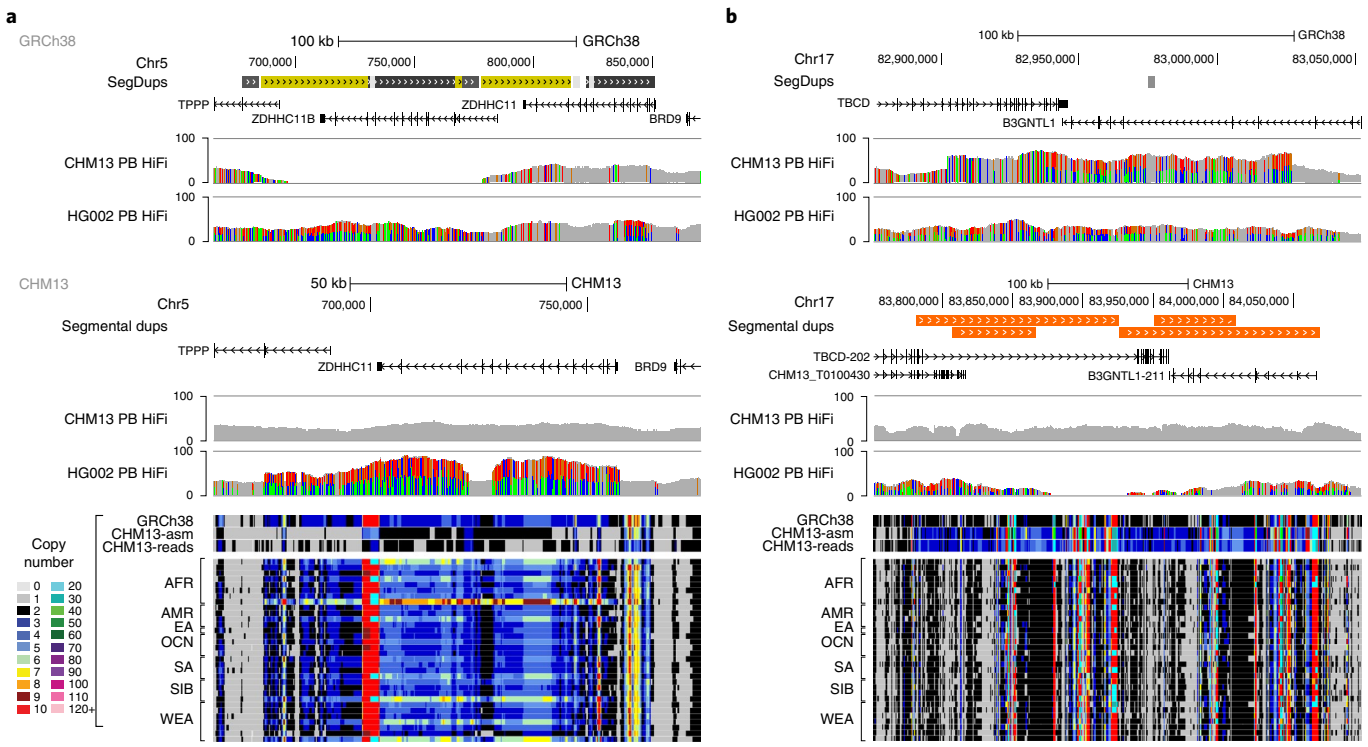


Fig. 4 | Examples of the largest CHM13 regions with a copy number in the reference that differs from GRCh38 and most individuals. a, One of the two largest examples of rare collapses in CHM13, where one copy of a common 72-kb tandem duplication is absent in CHM13. **b**, The largest rare duplication in CHM13, a 142-kb tandem duplication in GRCh38 that is rare in the population. CHM13 and HG002 PacBio HiFi coverage tracks are displayed for both references, GRCh38 (top) and CHM13v1.0 (bottom), to demonstrate that CHM13 reads support the CHM13 copy number, but HG002 reads are consistent with the GRCh38 copy number. Read-depth copy number estimates in CHM13 are shown at the bottom for '*k*-merized' versions of GRCh38 and CHM13v1.0 references, CHM13 Illumina reads and Illumina reads from a diverse subset ($n = 34$) of SGDP individuals.

are likely rare copy number variants in CHM13 (for example, CHM13v1.0 had only a single copy of the 72-kb tandem duplication in GRCh38; Fig. 4a). Additionally, we found that CHM13v1.0 had 33× fewer false or rare collapses than GRCh38 (~185 loci covering 6.84 Mb)[6]. We identified five regions (160 kb) with rare duplications in CHM13v1.0. This included a single 142-kb region that appeared to be a true, rare tandem duplication based on HiFi read-depth and Illumina *k*-mer-based copy number estimates (Fig. 4b). Two of the smaller regions appeared to be true, rare tandem duplications, and two other small regions were identified during polishing as heterozygous or mosaic deletions, revealing potential tandem duplications arising during cell line division or immortalization. In summary, we found 7.5× fewer rare or falsely duplicated bases in CHM13v1.0 relative to the 12 likely falsely duplicated regions affecting 1.2 Mb and 74 genes in GRCh38 (ref. [6]), including the medically relevant *CBS*, *CRYAA* and *KCNE1* genes[46].

Toward a completely polished sequence of a human genome. While evaluating CHM13v1.0, the T2T consortium successfully completed the construction of the rDNA models and their surrounding sequences on the p arms of the five acrocentric chromosomes[1]. In parallel, we determined that all telomeric sequences remained unpolished. Specifically, in canonical [TTAGGG] n repeats, we found both HiFi read coverage dropouts and ONT strand bias impeded high-quality variant calling (Extended Data Fig. 6a). For ONT, we observed only negative strands on the p arm and only positive strands on the q arm across all telomeric repeats at chromosomal ends; we suspect the ONT ultra-long transposon-based library preparation prevents reads from starting at chromosome ends, causing reads to only read into the telomere[10,33]. We tailored our PEPPER-based polishing approach and performed targeted telomere polishing to remove these errors remaining in telomeric sequences (Methods). Finally, automated polishing (described below), indicated that the *FAM156B* gene was heterozygous in CHM13v0.9, and CHM13v1.0 represented the rare minor allele (encoding a premature stop codon) at this locus. We replaced this minor allele with the other CHM13 allele encoding a full-length protein sequence. Overall, we made 454 telomere edits, producing longer stretches of maximum perfect matches to the canonical *k*-mer at each position across these telomeres compared to CHM13v1.0 (Extended Data Fig. 6b). Combined with the parallel completion of the five rDNA arrays, our final round of polishing led to an improved QV of Q73.94 for CHM13v1.1.

Again, to ensure updates did not compromise the high accuracy of the assembly and to identify any remaining issues, we carried out an additional round of SV detection and manual curation using HiFi and ONT with an updated Winnowmap2 alignment (Methods and Extended Data Fig. 7), classifying seven loci as remaining issues in CHM13v1.1 (Supplementary Table 2). We excluded CLRs because the lower base accuracy compared to HiFi and ONT and shorter read length compared to ONT were adding no information. Bionano was also excluded as the molecules were lacking coverage in centromeric regions (Extended Data Fig. 8) and did not detect any structural issues beyond the missing telomere and a few heterozygous SVs already identified by HiFi and ONT. Two loci located in the rDNA sequences appear to be a potential discrepancy between the model consensus sequence and actual reads or an artifact of mapping or sequencing bias. Lower consensus quality is indicated at two other loci, one detected with read alignments that were both low in coverage and identity, and one of which contained error

k-mers detected by the hybrid dataset. One locus consisted of multiple insertions (<1 kb) with breakpoints detected in low-complexity sequences associated with heterozygous variants and indicated a possible collapsed repeat (Extended Data Fig. 9) and an additional two loci joined and created an artificial chimeric haplotype (Extended Data Fig. 10). Additionally, we found 218 low-coverage loci using HiFi (Supplementary Table 3), with 81.2% associated with GA-rich (78.0%) regions. The remaining 41 loci had signatures of lower consensus quality and alignment identity, and 30 had error *k*-mers detected from the hybrid *k*-mer dataset. In contrast, we detected one low-coverage locus using ONT that overlapped with the GA-rich model rDNA sequence. We associated most remaining loci, totaling only 544.8 kb or <0.02% of assembled sequence, with lower consensus quality in regions lacking unique markers. Overall, we found 394 heterozygous regions, including regions with clusters of heterozygous variants (https://github.com/mrvollger/nucfreq/), totaling 317 sites (~1.1 Mb).

We manually curated both the breakpoints and alternate sequences associated with 47 heterozygous SVs, including sites previously inspected (CHM13v1.0) for SV-like error detection. We then investigated HiFi read alignment clippings and confirmed an association with clipping to both true heterozygous variant and spurious low-frequency alignments. Additionally, we detected a further heterozygous inversion that went previously undetected.

A comparison to automated assembly polishing. To demonstrate the efficacy of the customized DeepVariant-based approach, we compared our semiautomated polishing approach used to create CHM13v1.0 (Q72.62) to a popular state-of-the-art automated polishing tool, Racon[30]. We iteratively polished CHM13v0.9 (three rounds) using Racon with PacBio HiFi alignments. While the QV improved from Q70.22 to Q70.48 after the first round of Racon polishing, it degraded with the subsequent second (Q70.26) and third (Q70.15) rounds, ultimately diminishing assembly accuracy as a result of overcorrection. We also found that Racon incorporated 7,268 alternate alleles from heterozygous variants identified by DeepVariant, thus potentially causing undesirable haplotype switching in originally haplotype-consistent blocks. To examine how Racon polished large, highly similar repetitive elements, we counted the number of corrections in nonoverlapping 1-Mb windows of the CHM13v0.9 assembly and measured local polishing rates. Unlike CHM13v1.0, Racon polishing showed a clear right tail in the distribution of polishing rates, indicating the presence of polishing 'hotspots', defined here as loci with >60 corrections/Mb (Fig. 5a). The proximal and distal junctions of the rDNA units (masked from CHM13v1.0 polishing) were prevalent among these loci, a finding that reinforced the importance of masking known collapsed but resolved loci to avoid overcorrection. We also found non-rDNA loci that were preferentially polished by Racon, including satellite repeats such as the highly repetitive HSat3 region in chromosome 9. Finally, CHM13v1.0 made two corrections, recovering two protein-coding transcript's ORFs, but Racon did not make these corrections (Supplementary Table 1). While CHM13v1.0 did not induce invalid ORFs in any transcripts, Racon made ten corrections that caused invalid ORFs in 22 transcripts (from nine genes) (Fig. 5b). Most of these corrections occurred at homopolymer repeats, consistent with our previous findings that homopolymer bias in HiFi reads could lead to false expansion or contraction of homopolymers during polishing.

To overcome these relative shortcomings of Racon polishing, we tested polishing the CHM13v0.9 assembly with three iterative rounds of Racon followed by filtering with Merfin (Racon + Merfin). After each round of polishing, Merfin removed proposed Racon edits that incorporated false assembly *k*-mers. As expected, the Racon + Merfin assembly QV monotonically increased from Q70.22 to Q77.34, Q77.99 and Q78.12. However, Racon + Merfin still incorporated 2,274 alternate alleles from heterozygous variants and polishing hotspots were still evident, suggesting that some repeats were overcorrected (Fig. 5a). These overcorrections are not reflected in the QV measurements as *k*-mers from true heterozygous variants are considered 'valid' sequences. Merfin mitigated the ten ORF-invalidating Racon corrections; however, Merfin also failed to correct the two reading frame corrections made in CHM13v1.0 but not Racon (Supplementary Table 1). Overall, when considering only automated polishing, we suggest that Racon and Merfin can be used together as a highly effective strategy for building reference assemblies with minimum false-positive corrections. However, we would like to emphasize that a custom polishing pipeline with manual interventions is still required for preserving haplotype consistency and avoiding repeat overcorrection.

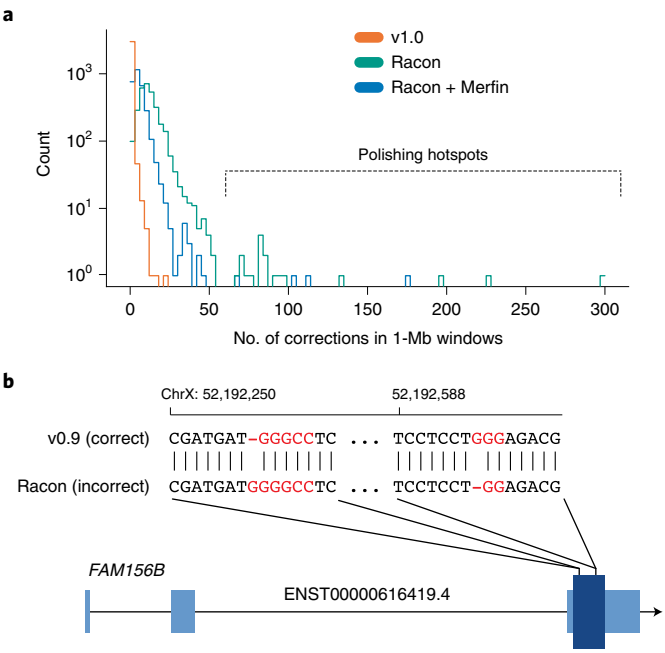


Fig. 5 | Errors made by automated polishing. a, The distribution of the number of polishing edits made in nonoverlapping 1 Mb windows of the CHM13v0.9 assembly. **b**, Two Racon polishing edits causing false frameshift errors in the *FAM156B* gene. Light blue indicates the untranslated region, and dark blue indicates the single coding sequence exon. Highlighted sequence indicates GC-rich homopolymers.

Discussion

The CHM13v0.9 human genome assembly represented a landmark achievement for the genomics community by representing previously unresolved repeats in a locally haplotype-consistent assembly. Although it was imperative to validate and correct this draft assembly, successful polishing faced three major obstacles. First, while repeats are challenging to polish in any draft assembly, the CHM13v0.9 assembly represented hundreds of megabases of exceptionally large and complex repeats genome wide, which could potentially induce false-positive (overcorrection) or false-negative polishing corrections. Secondly, although the CHM13 genome is mostly homozygous, we identified non-negligible levels of interspersed heterozygous variation. Therefore, it was essential to distinguish between heterozygous variants and polishing edits to maintain the original haplotype consistency. Finally, our evaluation of CHM13v0.9 discovered how homopolymer and coverage bias in HiFi reads caused assembly errors genome wide. This analysis also revealed that standard methods for measuring QV can be influenced by technology-specific biases.

These obstacles necessitated a custom and contextualized polishing and evaluation model that capitalized on the wealth of available data to exploit the advantages of each sequencing platform. It also required the use of specialized aligners, hard masking and manual intervention to avoid false polishing corrections within repeats. This polishing approach called for just 1,457 corrections including: p arm of chromosome 18, 454 telomere corrections, 1 large deletion, 2 large insertions, 993 SNPs, 113 small insertions and 880 small deletions. Although the final CHM13v1.1 is highly accurate (Q73.9), we identified 225 loci that were recalcitrant to validation, and we have documented these loci along with 394 heterozygous loci (317 merged loci; https://github.com/marbl/CHM13-issues/).

The high accuracy of CHM13v1.1 showcases the effectiveness of our informed selection and implementation of appropriate repeat-aware aligners[34,36], *k*-mer evaluation and filtration tools, and highly accurate and sensitive variant callers[28,41] while also highlighting the utility of capitalizing on the synergistic nature of multiple sequencing technology platforms. The minimal number of corrections implemented by our approach and uniform coverage (99.86%) exemplifies the high accuracy of the initial graph construction, with sequencing biases being associated with the remaining coverage fluctuations (223 regions were regions of HiFi dropouts, and 77.5% found in GA/TC-rich and AT-rich satellite sequences such as HSat2/3 and HSat1 were associated with HiFi coverage increases and ONT coverage depletion, respectively)[1].

In many respects, the T2T CHM13 genome assembly initiative is not representative of typical assembly projects. The success of the CHM13v1.0 assembly was enabled by the low level of heterozygosity of the CHM13 genome, advancements in sequencing technologies, a combination of sequencing technologies (HiFi, ONT and Illumina), customized assembly algorithms, and a large dedicated team of scientists, yielding results currently not possible with limited resources and automated algorithms[17,18]. However, despite the unique and semiautomated nature of our polishing and evaluation endeavor, recent trends in DNA sequencing and genome assembly algorithms suggest that CHM13v1.1 is just a preview of an imminent wave of high-quality T2T reference genomes in other species[47–49]. It is therefore critical that the lessons outlined here be incorporated into the next generation of automated bioinformatics tools[30,34,36,41]. For immediate projects, combining data types, using phased reads with repeat-ware alignments and carefully filtering polishing edits can improve automated polishing accuracy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01440-3.

Received: 13 July 2021; Accepted: 4 March 2022;
Published online: 31 March 2022

References

1. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, eabj6987 https://doi.org/10.1126/science.abj6987 (2022).
2. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 https://doi.org/10.1126/science.abj6965 (2022).
3. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 https://doi.org/10.1126/science.abj5089 (2022).
4. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eab7117 (2021).
5. Hufford, M. B. et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**, abg5289 https://doi.org/10.1126/science.abg5289 (2021).
6. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 https://doi.org/10.1126/science.abl3533 (2022).
7. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
8. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
9. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
10. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
11. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
12. Baran, N., Lapidot, A. & Manor, H. Formation of DNA triplexes accounts for arrests of DNA synthesis at d(TC)n and d(GA)n tracts. *Proc. Natl Acad. Sci. USA* **88**, 507–511 (1991).
13. Guiblet, W. M. et al. Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res.* **28**, 1767–1778 (2018).
14. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. *PLoS ONE* **8**, e62856 (2013).
15. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
16. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
17. Nurk, S. et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* https://doi.org/10.1038/s41592-020-01056-5 (2021).
19. Zimin, A. V. et al. Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
20. Simpson, J. T. et al. ABySS: a parallel assembler for short-read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
21. Watson, M. Mind the gaps—ignoring errors in long-read assemblies critically affects protein prediction. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-018-0004-z (2019).
22. Salzberg, S. L. et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
23. Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* https://doi.org/10.1038/s41586-021-03451-0 (2021).
24. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* **16**, e1007981 (2020).
25. Pacific Biosciences. GenomicConsensus module. https://github.com/PacificBiosciences/GenomicConsensus (2019).
26. 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).
27. Poplin, R. et al. A universal SNP and small-INDEL variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
28. Shafin, K. et al. Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
29. Oxford Nanopore Technologies. medaka: sequence correction provided by ONT Research https://github.com/nanoporetech/medaka (2018).
30. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
31. Zhang, H., Jain, C. & Aluru, S. A comprehensive evaluation of long-read error correction methods. *BMC Genomics* **21**, 889 (2020).
32. Fu, S., Wang, A. & Au, K. F. A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biol.* **20**, 26 (2019).
33. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
34. Jain, C. et al. Weighted minimizer sampling improves long-read mapping. *Bioinformatics* **36**, i111–i118 (2020).
35. Jain, C., Rhie, A., Hansen, N., Koren, S. & Phillippy, A. M. Long read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* (2022).
36. Mikheenko, A., Bzikadze, A. V., Gurevich, A., Miga, K. H. & Pevzner, P. A. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **36**, i75–i83 (2020).
37. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

38. Fofanov, Y. et al. How independent are the appearances of n-mers in different genomes? *Bioinformatics* **20**, 2421–2428 (2004).
39. Lang, D. et al. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* **9**, giaa123 (2020).
40. Olson, N. D. et al. precisionFDA Truth Challenge V2: calling variants from short- and long-reads in difficult-to-map regions. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.13.380741 (2021).
41. Formenti, G. et al. Merfin: improved variant filtering, assembly evaluation and polishing via *k*-mer validation. *Nat. Methods* (2022). https://doi.org/10.1038/s41592-022-01445-y
42. Zarate, S. et al. Parliament2: accurate structural variant calling at scale. *GigaScience* **9**, giaa145 (2020).
43. Sedlazeck, F. J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
44. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 https://doi.org/10.1126/science.abl4178 (2022).
46. Wagner, J. et al. Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01158-1 (2022).
47. Naish, M. et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. *Science* **374,** abi7489 https://doi.org/10.1126/science.abi7489 (2021).
48. Liu, J. et al. Gapless assembly of maize chromosomes using long-read technologies. *Genome Biol.* **21**, 121 (2020).
49. Du, H. et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **8**, 15324 (2017).

Methods

Evaluating homopolymer concordance. By analyzing the homopolymer length agreement, we assessed sequencing platform-specific biases between reads and the assembly using both Illumina and HiFi reads through the runLengthMatrix submodule of Margin (https://github.com/UCSC-nanopore-cgl/margin/). Here, we used Margin to convert the assembly sequence to a run-length encoded (RLE) sequence. For example, the sequence ACTTG became (ACTG, {1,1,2,1}) where ACTG represented the encoded sequence, and {1,1,2,1} represented the run length for each nucleotide base. While encoding the sequence to run length, Margin created a map of positions in the assembly to the RLE position. Using the position map, Margin converted the raw sequence alignment to run-length alignment by iterating through the matches between the read and the assembly and keeping track of the previous match in RLE space. This way, Margin created a matrix where each row represents a run length of a nucleotide base observed in the reads, and each column represents the run length observed at the corresponding position in the assembly where the read mapped.

Identifying potential polishing edits and heterozygous variants. To find potential polishing edits and heterozygous variants, we aligned a variety of public CHM13 WGS sequencing reads to T2T-CHM13v0.9 (https://github.com/marbl/CHM13/). We refer to these alignments as 'self-alignments' as both the query reads and reference assembly represent the CHM13 genome. Further, we aligned Illumina reads with BWA-MEM (v0.7.15)[50] and removed PCR duplicate-like redundancies using 'biobambam2 bamsormadup' (v2.0.87)[51] with default parameters. Pacific Biosciences CLR and Circular Consensus Sequencing (CCS/HiFi) and ONT reads were aligned using Winnowmap2 (v1.1).

We used both Illumina and HiFi read alignments to call SNPs and INDELs with the 'hybrid' model of DeepVariant (v1.0) but only ONT alignments were used to call SNPs using PEPPER-DeepVariant (v1.0)[28]. To exclude potentially spurious variant calls, we removed variants with low allele fraction support or low genotype quality (VAF ≤ 0.5, GQ ≤ 30 for Illumina/HiFi and GQ ≤ 25 for ONT). We then combined Illumina/HiFi hybrid and ONT variant calls using a custom script (https://github.com/kishwarshafin/T2T_polishing_scripts/blob/master/polishing_merge_script/vcf_merge_t2t.py/). Finally, we filtered small polishing edits using Merfin[41] to ensure all retained edits did not introduce any false 21-mers that were absent from the Illumina or HiFi reads.

Our approach implemented SV inference tools to detect medium-sized polishing edits and structural heterozygosity. For short-read-based SV calling, we used Illumina alignments as input to Parliament2 (v0.1.11)[42] using default settings. For long-read SV calling, we relied on HiFi, CLR and ONT alignments to call SVs with Sniffles[43] (v1.0.12, --s 3 --d 500 --n --1) and we removed all SVs with less than 30% of reads supporting the ALT allele. After this, we generated and refined insertion and deletion sequences with Iris (v1.0.3, using Minimap2 (ref. [52]) and Racon[30] for aligning and polishing, respectively; https://github.com/mkirsche/Iris/). Our approach yielded three independent technology-specific call sets that we merged using Jasmine (v1.0.2, max_dist = 500 min_seq_id = 0.3 spec_reads = 3 --output_genotypes)[53]. Through manual inspection in Integrative Genomics Viewer (v2.6), we validated all long-read variant calls longer than 30 bp supported by at least two technologies and all short-read SV calls[54].

Our approach combined small and SV calls into two distinct VCF files: one for potential polishing edits (homozygous ALT alleles) and one for putative heterozygous variants (heterozygous ALT alleles), and we excluded all edits within known problematic loci, prone to producing false variant calls (rDNA gaps as well as the large HSat3 region on chromosome 9). To generate the CHM13v1.0, we applied 'bcftools consensus' (v1.10.2-140-gc40d090) to incorporate the suggested polishing edits into CHM13v0.9 (ref. [55]) and repeated same previously detailed methods with respect to CHM13v1.0 to ensure that no additional polishing edits were apparent and to call heterozygous loci.

Patching the chromosome 18 p arm telomere. As a result of the string graph simplification process, we found a telomere missing from the graph representing the p arm of chromosome 18. Bionano molecules and assemblies of the molecules[33] were mapped to CHM13v0.9 and CHM13v1.0 using Bionano Solve v3.6 and all scaffolds were manually inspected end to end to search for assembly errors. No issues were detected except this missing telomere. We identified five ONT reads associated with these telomeric sequences using the telomere pipeline developed by the VGP (https://github.com/VGP/vgp-assembly/). Using these reads we ran Medaka (v1.0.3)[29] to generate a consensus sequence and manually patched it into the assembly (https://github.com/malonge/PatchPolish/). We obtained seven matching HiFi reads, not in the assembly graph and confirmed to have telomeric repeats, and used Racon (v1.6.0)[30] to further polish. In total, we added 4,862 bp of telomere sequence to the start of chromosome 18.

Evaluating polishing accuracy. We repeated self-alignment variant calling methods on CHM13v1.0 and confirmed that no additional polishing errors were apparent. In addition to the self-alignments used for polishing and heterozygous variant calling, we derived marker-assisted alignments from previously created HiFi, CLR and ONT Winnowmap2 alignments[35]. For marker-assisted alignment production, we removed Winnowmap2 alignments that did not span 'marker' *k*-mers. We defined marker *k*-mers as any 21-mer present once in CHM13v1.0 and between 42 and 133 times in the Illumina reads[33] and filtered reads using technology-specific length thresholds with HiFi having a 10-kb threshold, CLR a 1-kb threshold and ONT a 25-kb threshold. Our approach relied on both CHM13v1.0 self-alignments and marker-assisted alignments for manual inspection.

We also assessed the genome assembly using Merqury v1.3 QV estimations based on 21-mer databases that we created for both Illumina PCR-free and HiFi reads[37]. Following this, we derived a 'hybrid' Merqury *k*-mer database using Meryl v1.3 by combining Illumina and HiFi *k*-mers that occurred over one time, and adjusting the *k*-mer frequency to match the *k*-mer frequency at 35× in the diploid (two-copy) peak by increasing *k*-mer frequency in HiFi reads by 4 and dividing the Illumina *k*-mer frequency by 3 and combined the *k*-mer databases by taking the union and setting frequency to the maximum observed in the two data types.

To identify regions with rare collapses or rare duplications in CHM13v1.0, we compared copy number estimates of CHM13v1.0 to copy number estimates of 268 human genomes (SGDP) using short reads[2]. We averaged these copy number estimates for each genome across 1-kb windows and flagged a potential false or rare duplication if the copy number in CHM13v1.0 was greater than the copy number in 99% of the other genomes and GRCh38. Moreover, we flagged a potential false or rare collapse if the copy number in CHM13v1.0 was less than the copy number in 99% of the other genomes and GRCh38 and assigned all flagged regions a value of 1 and unflagged regions a value of 0. We included GRCh38 in this analysis to help remove rare technical artifacts where the assembly-based *k*-mer copy number estimate was systematically different from the Illumina read-based *k*-mer estimate. To filter the flagged regions, we used a median filter approach with a window size of 3 kb where the binary value of each 1-kb region was replaced with the median value of the complete window. Finally, we merged all adjacent flagged regions and reported the start and end coordinates with respect to CHM13v1.0, and we curated and removed flagged regions if they overlapped long interspersed nuclear elements (LINEs) as SGDP copy number estimates are less reliable in these high copy number repeats.

Polishing enrichment or depletion within repeats. We performed a permutation test to check if our polishing pipeline suggested significantly more or fewer polishing edits within repeats compared to the rest of the genome. We established two distinct samples of genomic intervals. For the first, we randomly sampled 20,000 100-kb windows from the genome and removed any windows that intersected repeats. For the second, we randomly sampled 20,000 100-kb windows and removed any windows that intersected non-repeats. By measuring the number of polishing edits in each 100-kb window, we established two different random distributions of polishing rates: one within and one without repeats. We utilized SciPy (v1.7.0) stats.ttest_ind using 10,000 permutations to derive the *P* value[56]. SciPy is available in Python package v3.5.

Telomere polishing. We used a targeted polishing of telomeres by retraining PEPPER (v0.4)[28] on HG002 chr20 with all forward strand reads removed to correct for the original model's dependence on having reads from both strands. Using this retrained model, we generated a set of candidate variants in the telomere regions and the coverage depth was calculated using samtools[57] depth (v1.9). Finally, we implemented a custom script (https://github.com/kishwarshafin/T2T_polishing_scripts/blob/master/telomere_variants/generate_telomere_edits.py/) that took these candidate variants and calculated the Levenshtein distance between the canonical telomere *k*-mer and the sequence we derived after the candidate variant had been applied. We selected only those variants as true telomere edits if the candidate had a minimum allele frequency of 0.5, a minimum GQ score of 2 and reduced the Levenshtein distance to the canonical telomere *k*-mer when compared to the existing sequence. Further, we trimmed the consensus sequence where ONT read-depth support was lower than 5.

Coverage supports and excessive clippings. We used SV detection to identify regions with low-coverage support, excessive read clippings, and enriched secondary alleles, and to further ensure that accuracy was not compromised but also to identify and document outstanding issues with CHM13v1.1 (Fig. 1a). On inspection of both Winnowmap2 (ref. [34]) and Minimap2 (ref. [52]) read clippings, artificial alignment breaks were highlighted that caused clipping and coverage drops in regions with highly identical satellite sequences. Notably, we did not identify these breaks in alignments from TandemMapper[36], a more conservative aligner specifically designed for alignment in satellite repeats. On further inspection of clipped reads, we found the chaining algorithm of Winnowmap2 handled lower-confidence alignment blocks incorrectly, and so we updated accordingly (v2.0 to v2.01) for all future evaluations of both CHM13v1.0 and CHM13v1.1 (Extended Data Fig. 7).

Comparison to automated polishing approaches. To evaluate our newly proposed approach to polishing, we compared it to the off-the-shelf tools available for HiFi reads. We performed three rounds of iterative polishing using the Racon consensus tool with each iteration including the following steps: (1) We aligned input HiFi reads to the input target sequences using Winnowmap 1.11 (https://github.com/marbl/Winnowmap/releases/tag/v1.11/; options: '--MD --W bad_mers.txt --ax

map-pb') as used for polishing CHM13v0.9. We used CHM13v0.9 (unpolished) as the first iteration target, while every following iteration used the polished output of the previous stage as the input target. (2) We filtered secondary alignments and alignments with excessive clipping using the 'falconc bam-filter-clipped' tool (available in the 'pbipa' Bioconda package using the options 'falconc bam-filter-clipped --t --F 0×104'). By default, maximum clipping on either the left or the right side of an alignment is set to 100 bp, but this was applied only if the alignment was located at least 25 bp from the target sequence end (to prevent clipping due to contig ends, which could otherwise cause false alignment filtering). (3) Finally, we used Racon (https://github.com/isovic/racon/; branch 'liftover', commit: 73e4311) to polish the target sequences using these filtered alignments. For this work, we extended the 'master' branch of Racon to include two custom features: BED selection of regions for polishing and logging all changes introduced to the input draft assembly to produce the final polished output (in VCF, PAF or SAM format). We then ran Racon with default options with the exception of two new logging options: '--L out_prefix --S' implemented to store the liftover information between the input and output sequences. We used Liftoff (v1.6.0, --chroms --copies --exclude_partial --polish) using GENCODE v35 to annotate each of the polished assemblies[58,59].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data types and assemblies are available on https://github.com/marbl/CHM13/ and under NCBI BioProject PRJNA559484 with the Assembly GenBank accession GCA_009914755. Polishing edits, cataloged remaining issues and known heterozygous regions are available on https://github.com/marbl/CHM13-issues/. All the data in the two GitHub repositories are directly downloadable from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/ with no restrictions. The retrained PEPPER model used for telomere polishing is available to download at https://storage.cloud.google.com/pepper-deepvariant-public/pepper_models/PEPPER_HP_R941_ONT_V4_T2T.pkl. Source data for generating plots in this paper are available on https://github.com/arangrhie/T2T-Polish/tree/master/paper/2022_Mc_Cartney/.

Code availability

To facilitate usability of our evaluation and polishing strategy, we made the up-to-date version of tools that have been used within our workflows openly available on https://github.com/arangrhie/T2T-Polish/. Exact codes used for polishing CHM13v0.9 and CHM13v1.0 are available on https://github.com/marbl/CHM13-issues/. Both GitHub repositories are available through a public domain, and have been deposited to Zenodo[60,61]. Custom scripts used for merging small variants, and generating telomere edits are available at https://github.com/kishwarshafin/T2T_polishing_scripts/ and deposited to Zenodo[62] under an MIT license.

References

50. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
51. Tischler, G. & Leonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code Biol. Med.* **9**, 13 (2014).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
53. Kirsche, M. et al. Jasmine: Population-scale structural variant comparison and analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2021.05.27.445886 (2021).
54. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer: high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
55. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
56. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
57. Danecek, P. et al. Twelve years of SAMtools and BCFtools, *GigaScience* **10**, giab008 (2021).
58. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016 (2020).
59. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
60. Rhie, A., Formenti, G., Shafin, K., Fungtammasan, A., & Jain, C. arangrhie/T2T-Polish: v1.0. https://doi.org/10.5281/zenodo.5649017 (2021).
61. Rhie, A. and Phillippy, A. marbl/CHM13-issues: v1.1. https://doi.org/10.5281/zenodo.5648989 (2021).
62. Shafin, K. kishwarshafin/T2T_polishing_scripts: v0.1 release for zenodo. https://doi.org/10.5281/zenodo.6127865 (2021).

Acknowledgements

This work was supported by the Intramural Research Program of the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH 1ZIAHG200398; to A.M.M., C.J., S.K., A.M.P. and A.R.); National Science Foundation DBI-1350041 and IOS-1732253 (to M.A.); NIH/NHGRI R01HG010485, U41HG010972, U01HG010961, U24HG011853 and OT2OD026682 (to K.S. and B.P.); HHMI (to G.F.); Wellcome WT206194 (to K.H. and J.M.W.); NIGMS F32 GM134558 (to G.A.L.); NIH/NHGRI R01 1R01HG011274-01, NIH/NHGRI R21 1R21HG010548-01 and NIH/NHGRI U01 1U01HG010971 (to K.M.); St. Petersburg State University grant ID PURE73023672 (to A.M.); NIH/NHGRI R01HG006677 (to A.S.); Fulbright Fellowship (to D.C.S.); and Intramural funding at the National Institute of Standards and Technology (to J.Z.). This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov/). Certain commercial equipment, instruments or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments or materials identified are necessarily the best available for the purpose.

Author contributions

A.R. and A.M.P. conceived and supervised the project. A.M.M., K.S., G.F., K.H., J.M.D.W. and A.R. performed the pre-polishing evaluation. K.S., M.A., A.V.B., A.F., C.J., A.M., B.P. and A.R. aligned reads and called variants. A.M.M., K.S., M.A., G.F., A.F., K.H.M., A.M., J.M.Z. and A.R. manually validated variant calls. D.C.S. and J.M.Z. performed the gene collapse and expansion analysis. K.S., M.A., A.V.B., G.A.L., K.H.M., A.M. and A.R. identified and curated heterozygous and 'issues' loci. K.S., M.A., S.K. and B.P. patched and polished the telomeres. A.M.M., M.A., A.S. and I.S. performed automated polishing. A.M.M., K.S., M.A. and A.R. wrote the manuscript, with assistance from all authors. All authors approved the final manuscript.

Competing interests

I.S. is an employee of PacBio. A.F. is an employee of DNAnexus. S.K. has received travel funds to speak at symposia organized by Oxford Nanopore. The remaining authors declare no competing interests.

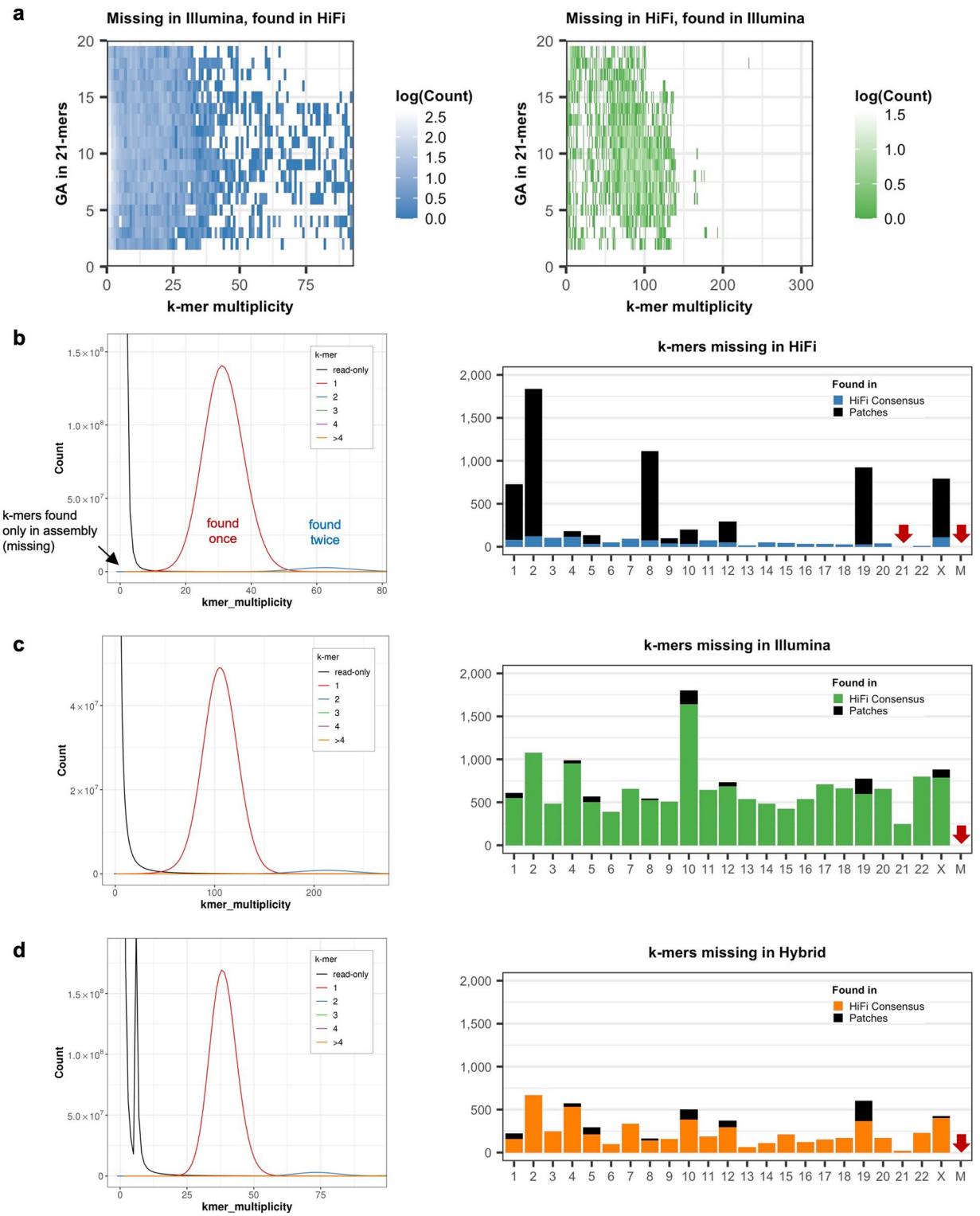
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01440-3.

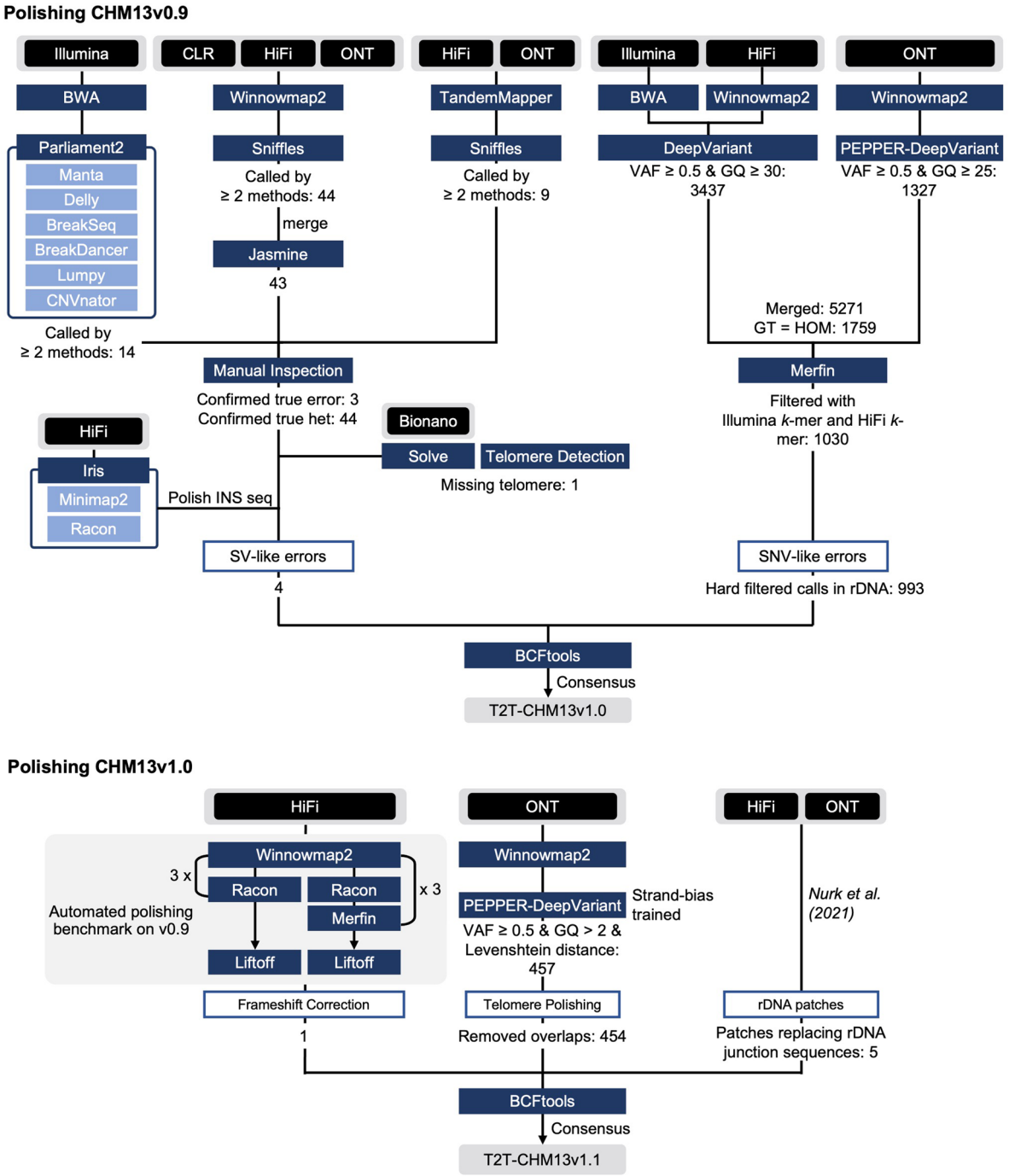
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01440-3.

Correspondence and requests for materials should be addressed to Adam M. Phillippy or Arang Rhie.

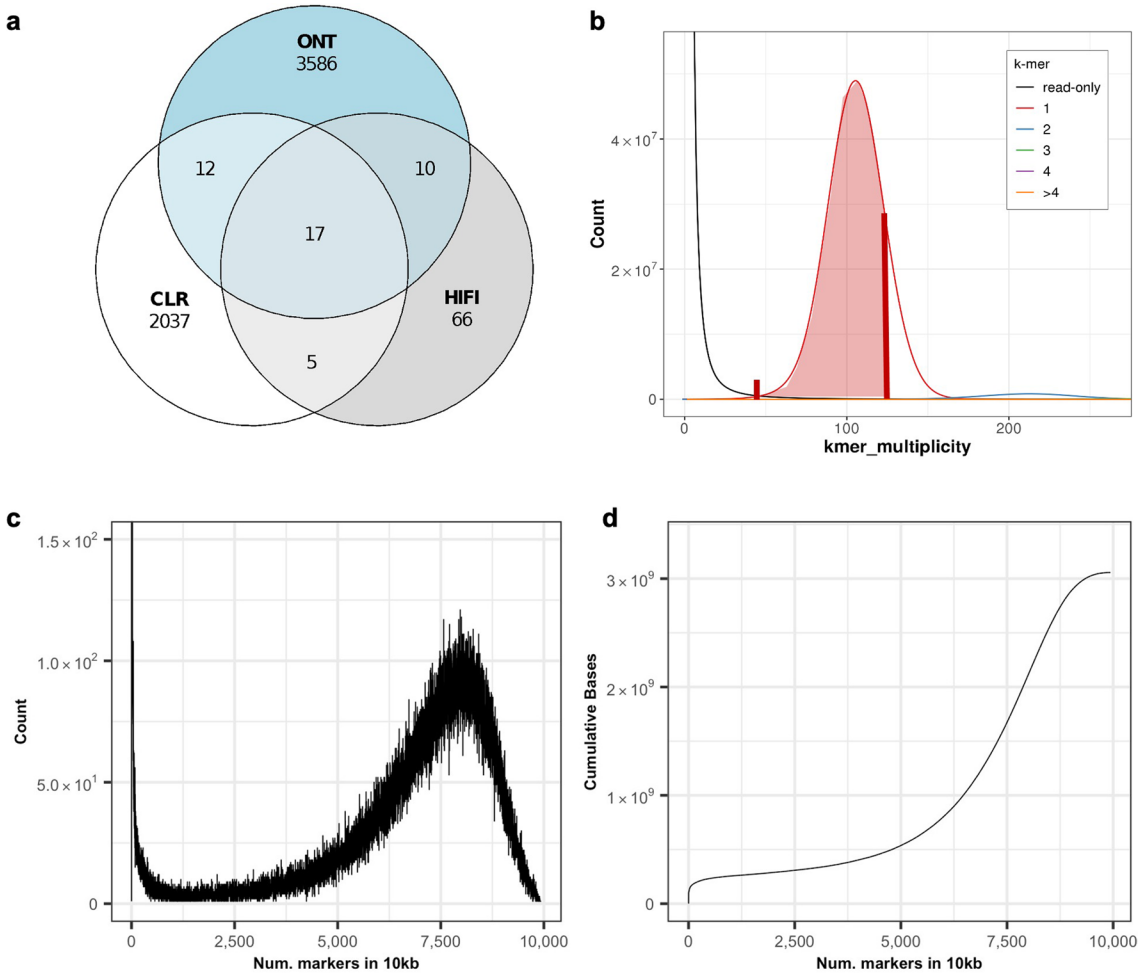
Peer review information *Nature Methods* thanks Kai Wang and Jue Ruan for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.



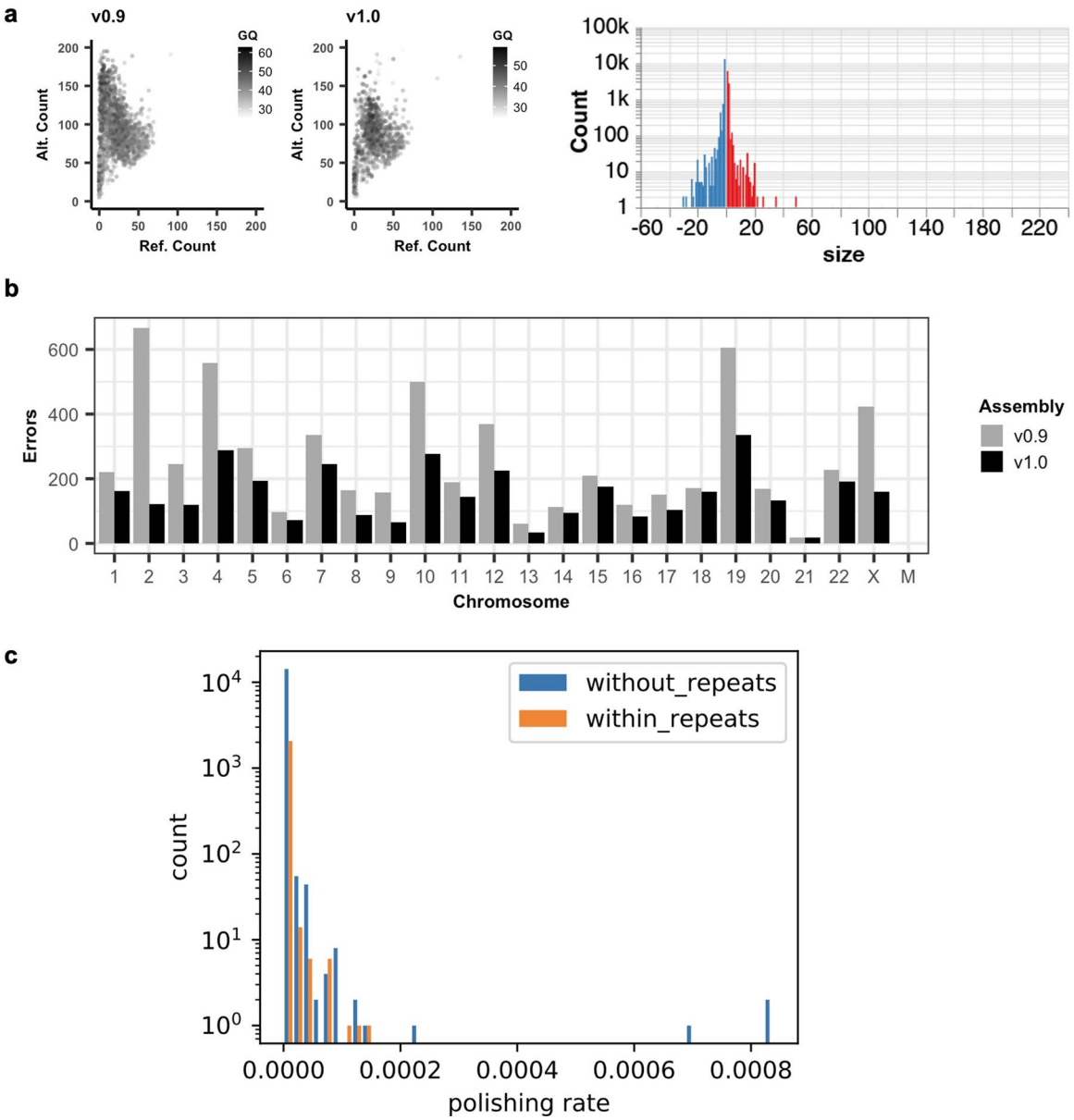
Extended Data Fig. 1 | Sequencing biases observed in missing *k*-mers. a, missing *k*-mers with its GA composition. **b-d**, v0.9 assembly and *k*-mer copy number spectrum from HiFi, Illumina, and hybrid *k*-mer sets (left) and per-chromosome missing (likely error) *k*-mer counts from the HiFi derived consensus or patches (right). Most missing *k*-mers in HiFi overlapped sequences from patched regions. No missing *k*-mer was found on chromosomes indicated with red arrows.



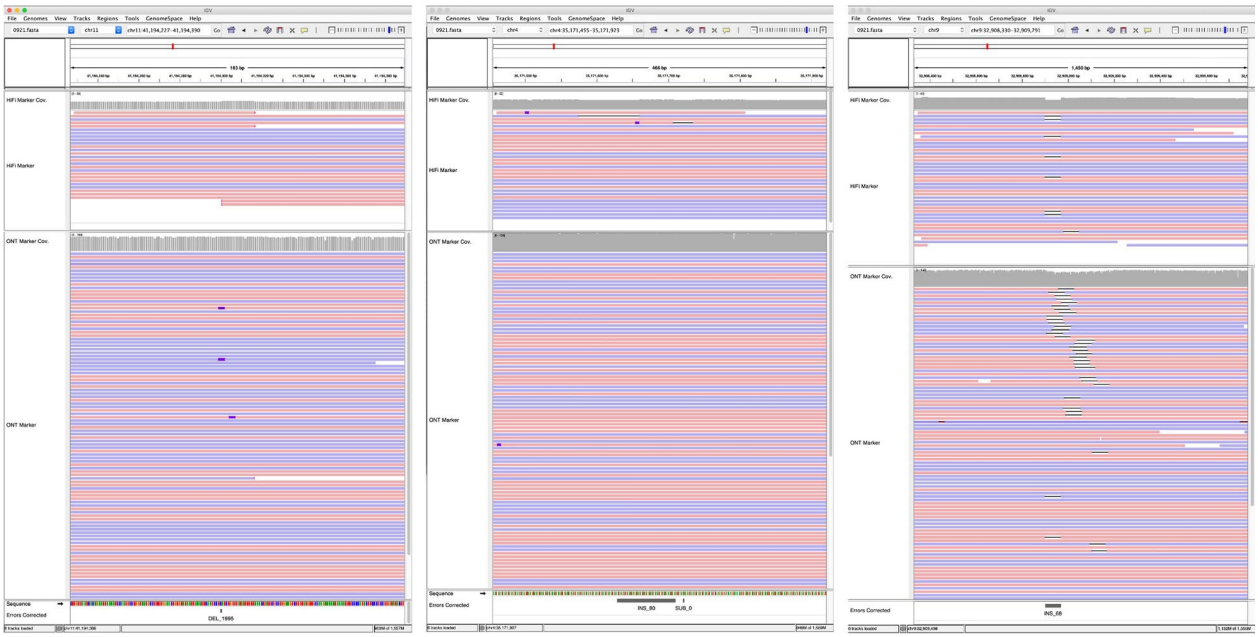
Extended Data Fig. 2 | Error detection and polishing pipeline. A detailed overview of the polishing pipeline along with the number of errors identified and polished at each step. Additionally, data type and polishing tools utilized are highlighted. Illumina, 100X PCR-free library Illumina reads; HiFi, 35x PacBio HiFi reads; ONT, 120x Oxford Nanopore reads.



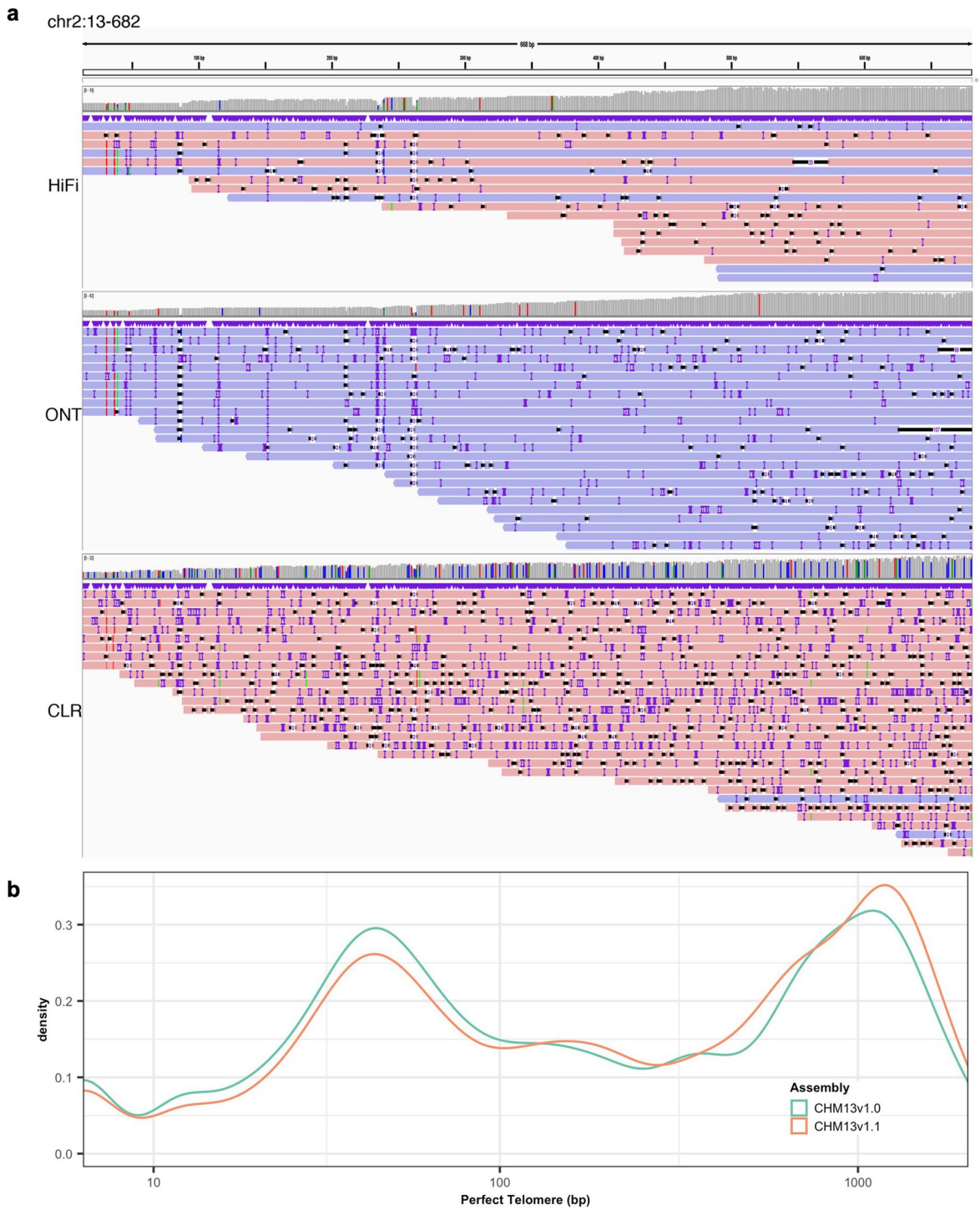
Extended Data Fig. 3 | Number of SV-like errors and globally unique single copy k-mers used for marker assisted alignment. a. Number of SV-like errors called from long-read platforms. **b**. Range of *k*-mer counts defined as 'single-copy' markers from Illumina reads and in the assembly. The cutoffs were chosen to minimize inclusion of low-frequency erroneous *k*-mers and 2-copy *k*-mers. **c**. Number of markers in every 10 kb window. **d**. Cumulative number of bases covered by the number of markers in each 10 kb window.



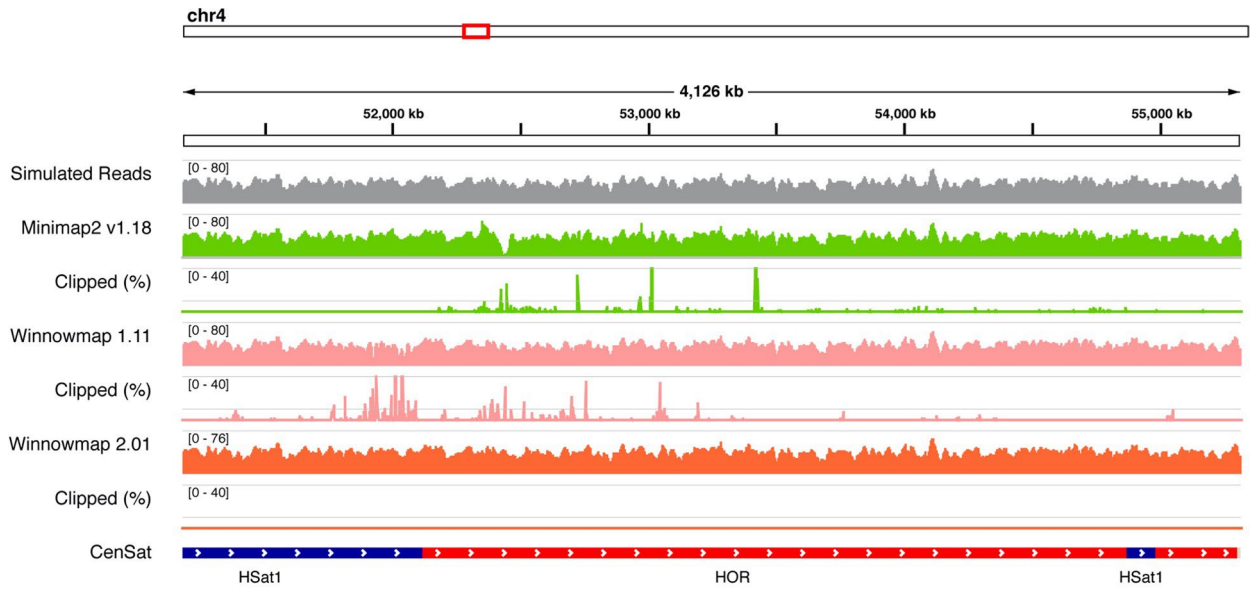
Extended Data Fig. 4 | Post-polishing evaluation. a. Left, genotype quality and number of reads supporting the reference and alternate alleles from the combined Illumina-hifi hybrid and ONT homozygous variant calls, with AF > 0.5. Right, balanced insertion (red) and deletion (blue) length distribution from the Illumina-HiFi hybrid DeepVariant heterozygous calls in CHM13v1.0. **b**. Number of errors detected in each chromosome, before and after polishing. **c**. Polishing inside and outside of repeats. The distribution of CHM13v0.9 polishing rates within and without repeats.



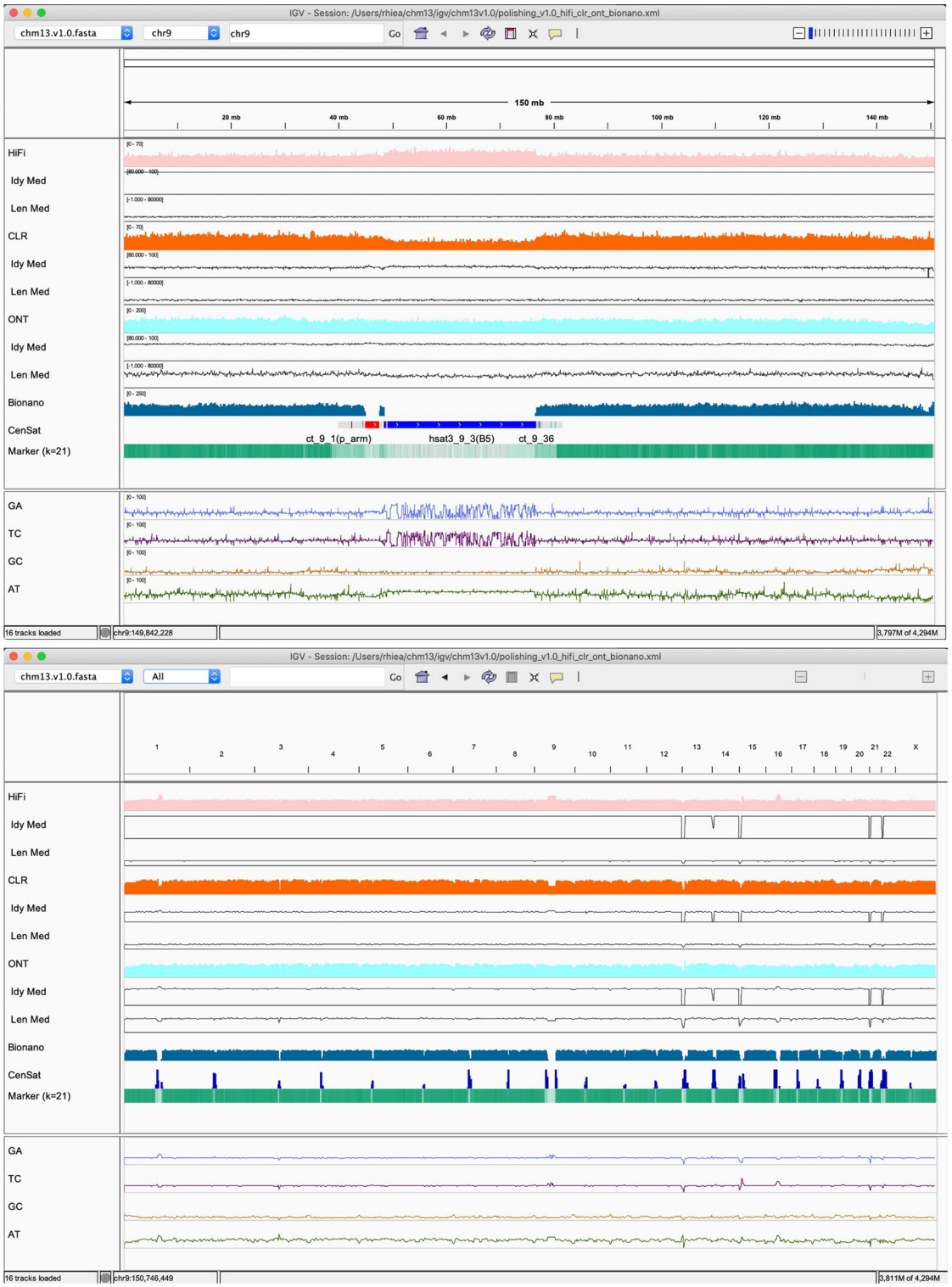
Extended Data Fig. 5 | Three SV-like errors corrected. HiFi and ONT marker assisted alignments, post correction of the 3 large SV-like edits visualized with IGV. HiFi coverage track is shown in data range up to 60, ONT up to 150. Clipped reads are flagged for >100 bp. INDELs smaller than 10 bp are not shown. Reads are colored by strands; positive in red and negative in blue.



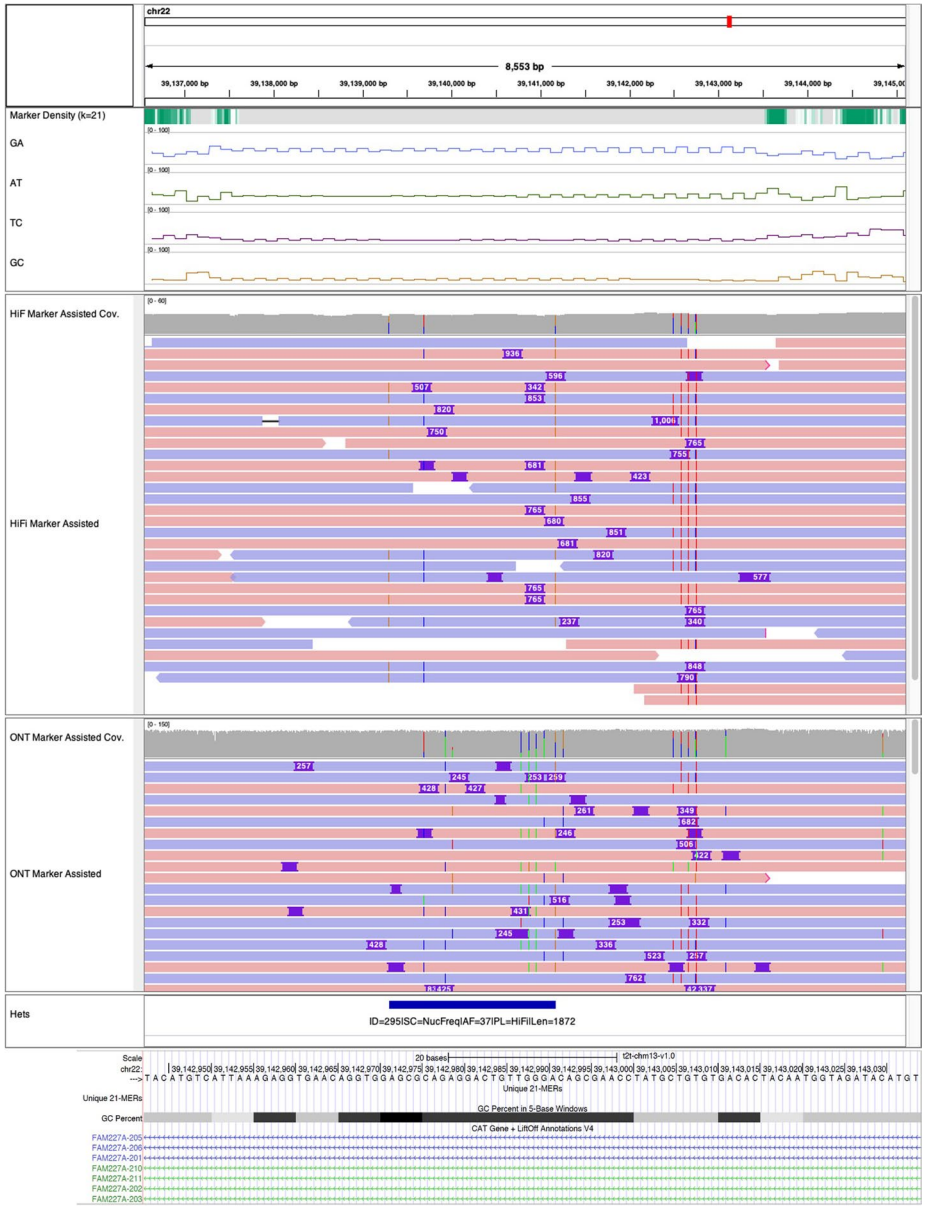
Extended Data Fig. 6 | Telomere polishing. a. An illustration of Chr. 2 telomere sequence reads from HiFi, ONT and CLR platform. **b**. Distribution of maximum perfect match to the canonical *k*-mer observed at each position in the telomere before (CHM13v1.0) and after (CHM13v1.1) polishing the telomeres.



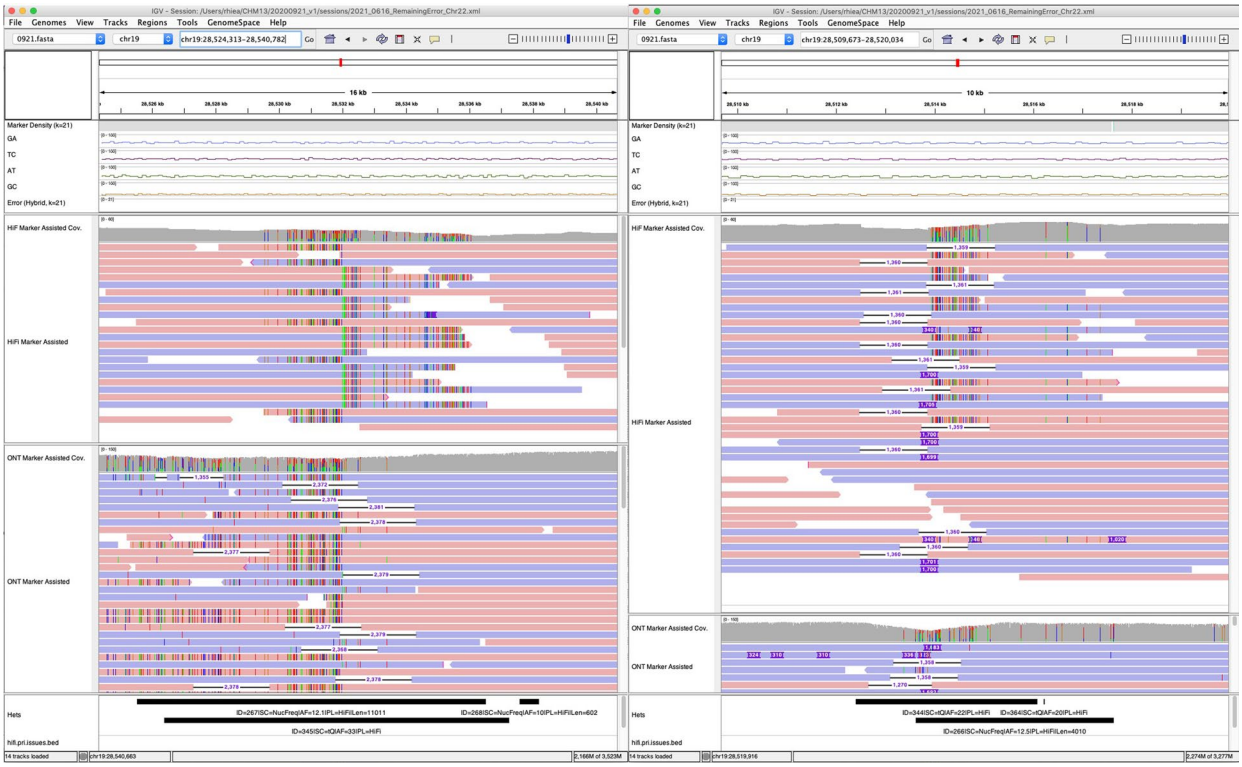
Extended Data Fig. 7 | Mapping biases found and corrected. On simulated HiFi reads, we found excessive clippings in highly identical satellite repeats in Minimap and Winnowmap by the time of evaluation. We have addressed this issue in Winnowmap 2.01+. Clipped (%) indicates the percentage of reads clipped in every 1,024 bp window, shown in 0~40% range with a midline of 10%.



Extended Data Fig. 8 | HiFi, CLR, ONT read coverage, alignment identity, and read length from Winnowmap2 v2.01 alignments and Bionano DLE-1 molecule coverage from Bionano Solve. Upper panel shows a zoomed in region of Chromosome 9, while the upper panel shows the whole-genome alignment view. HiFi, CLR, ONT, and Bionano coverage are shown up to 70x, 70x, 200x, and 250x, respectively. Median read identity in every 1,024 bp is shown in 80-100% range. Median read length in every 1024 bp is shown in 0-100 kb range. Read identity was the worst in CLR, and between HiFi and ONT. Bionano molecules were lacking coverage in most of the centromeric repeats.



Extended Data Fig. 9 | Collapsed simple tandem repeat. The collapse in the Intronic sequences of gene *FAM227A* was undetected, due to the variable insertion breakpoints and insertion length in the HiFi and ONT alignments. The panels above the alignments show marker density and percent microsatellites (GA / AT / TC / GC) in each 64 bp window, which indicates this region is highly repetitive with GA enriched sequences, which later alternates with AT enriched sequences.



Extended Data Fig. 10 | Chimeric junction of two haplotypes. In the shown above regions, both HiFi and ONT reads indicate that the consensus has a chimeric junction of the two haplotypes.

nature research

Corresponding author(s): Adam Phillippy, Arang Rhie

Last updated by author(s): Feb. 24, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The aws command line interface tool (AWS CLI version 1) was used to download data from our hub repository on https://github.com/marbl/CHM13 and SRA toolkit for downloading Illumina data from SRX1009644.
Data analysis	The runLengthMatrix submodule of Margin (https://github.com/UCSC-nanopore-cgl/margin) was used to collect run-length encoded sequence. Illumina reads were aligned with BWA-MEM v0.7.15, and de-duplicated using biobambam2 v2.0.87. Long reads were aligned initially with Winnowmap2 v1.1 for CHM13v0.9, then v2.01 for CHM13v1.0 and CHM13v1.1. DeepVariant v1.0 was used for calling SNV-like errors from Illumina and HiFi, PEPPER-DeepVariant v1.0 for ONT reads. Variants were merged using a custom script (https://github.com/kishwarshafin/T2T_polishing_scripts/blob/master/polishing_merge_script/vcf_merge_t2t.py) and filtered with a developmental version of Merfin, integrated as v1.0, using k-mers obtained with Meryl v1.3. SV-like errors were detected with Parliament2 v0.1.11 and Sniffles v1.0.12. Insertion and deletion sequences were refined with Irys v1.0.3. All SV-errors called from each platforms were merged using Jasmine v1.0.2. Manual inspection of the SV-errors were performed using IGV v2.6. Polished consensus was generated using bcftools v1.10.2. For determining telomeric errors, Bionano molecules and CMAPs were aligned and detected using Solve v3.6. Telomeric regions were identified using telomeric signals from the VGP telomere pipeline (https://github.com/VGP/vgp-assembly). ONT reads overlapping these regions were pulled out and used to patch with Medaka v1.0.3, followed by HiFi polishing and patching using a custom script (https://github.com/malonge/PatchPolish). Additional polishing with HiFi reads were performed with Racon v1.6.0. Assembly base-level quality was measured using Merqury v1.3 using 21-mers obtained with Meryl v1.3. A permutation test was performed to check if our polishing pipeline suggested significantly more or fewer polishing edits within repeats compared to the rest of the genome using SciPy. PEPPER v0.4 was used for telomere polishing. Racon commit: 73e4311 was used for automated polishing benchmarks. To facilitate usability of our evaluation and polishing strategy, we made the up-to-date version of tools that have been used within our workflows openly available on https://github.com/arangrhie/T2T-Polish. Exact codes used for polishing CHM13v0.9 and CHM13v1.0 are available on https://github.com/marbl/CHM13-issues. Both GitHub repositories are available through a public domain, and have been

deposited to Zenodo61,62. Custom scripts used for merging small variants, and generating telomere edits are available at https://github.com/kishwarshafin/T2T_polishing_scripts and deposited to Zenodo63 under a MIT license.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data types and assemblies are available on https://github.com/marbl/CHM13 and under NCBI BioProject PRJNA559484 with the Assembly GenBank accession GCA_009914755. Polishing edits, catalogued remaining issues and known heterozygous regions are available on https://github.com/marbl/CHM13-issues. All the data in the two GitHub repositories are directly downloadable from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/ with no restrictions. The retrained PEPPER model used for telomere polishing is available to download at https://storage.cloud.google.com/pepper-deepvariant-public/pepper_models/PEPPER_HP_R941_ONT_V4_T2T.pkl. Source data for generating plots in this manuscript are available on https://github.com/arangrhie/T2T-Polish/tree/master/paper/2022_Mc_Cartney.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was not considered because this study was designed to evaluate and polish the first and only complete human genome assembly available to date.
Data exclusions	No data were excluded
Replication	All results can be reproduced using the version of softwares used. However, some tools have been updated based on the results we learned from this manuscript, which will produces comparable or better results.
Randomization	Randomization was not performed because this study was designed to polish the first and only complete human genome assembly available to date.
Blinding	Blinding was not performed because this study was designed to polish the first and only complete human genome assembly available to date.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging