Title: A complete human reference genome improves variant calling for population and clinical genomics

Sergey Aganezov[1,*], Stephanie M. Yan[2*], Daniela C. Soto[3*], Melanie Kirsche[1,*], Samantha Zarate[1,*], Pavel Avdeyev[4], Dylan J. Taylor[2], Kishwar Shafin[5], Alaina Shumate[6], Chunlin Xiao[7], Justin Wagner[8], Jennifer McDaniel[8], Nathan D. Olson[8], Michael E.G. Sauria[2], Mitchell R. Vollger[9], Arang Rhie[4], Melissa Meredith[5], Skylar Martin[10], Sergey Koren[4], Jeffrey A. Rosenfeld[11], Benedict Paten[5], Ryan Layer[10], Chen-Shan Chin[12], Fritz J. Sedlazeck[13], Nancy F. Hansen[14], Danny E. Miller[9,15], Adam M. Phillippy[4], Karen H. Miga[5], Rajiv C. McCoy[2,†], Megan Y. Dennis[3,†], Justin M. Zook[8,†], Michael C. Schatz[1,2,16,†]

Affiliations:

[1] Department of Computer Science, Johns Hopkins University, Baltimore MD, USA
[2] Department of Biology, Johns Hopkins University, Baltimore MD, USA
[3] Biochemistry & Molecular Medicine, Genome Center, MIND Institute, University of California, Davis, Davis, CA, USA
[4] Genome Informatics Section, National Human Genome Research Institute, Bethesda, MD, USA
[5] UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA
[6] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
[7] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[8] National Institute of Standards and Technology, Gaithersburg, MD, USA
[9] Department of Genome Sciences, University of Washington, Seattle, WA, USA.
[10] Department of Computer Science and the Biofrontiers Institute, University of Colorado, Boulder, CO, USA
[11] Cancer Institute of New Jersey, New Brunswick, NJ, USA
[12] DNAnexus, Mountain View, CA, USA
[13] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
[14] Comparative Genomics Analysis Unit, National Human Genome Research Institute, Rockville, MD, USA
[15] Department of Pediatrics, Division of Genetic Medicine, University of Washington and Seattle Children's Hospital, Seattle, WA, USA.
[16] Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

* Equal contribution
† co-corresponding authors: mydennis@ucdavis.edu, rajiv.mccoy@jhu.edu, justin.zook@nist.gov, mschatz@cs.jhu.edu

Abstract:

Compared to its predecessors, the Telomere-to-Telomere CHM13 genome adds nearly 200 Mbp of sequence, corrects thousands of structural errors, and unlocks the most complex regions of the genome to clinical and functional study. Here we demonstrate how the new reference universally improves read mapping and variant calling for 3,202 and 17 globally diverse samples sequenced with short and long reads, respectively. We identify hundreds of thousands of novel variants per sample—a new frontier for evolutionary and biomedical discovery. Simultaneously, the new reference eliminates tens of thousands of spurious variants per sample, including a 12-fold reduction in 269 medically relevant genes. The vast improvement in variant discovery coupled with population and functional genomic resources position CHM13 to replace GRCh38 as the prevailing reference for human genetics.

Please add figures here:

https://docs.google.com/presentation/d/1zK4-x_EN79TTlVl3wUeufPLXSss2irG2ZDLaTE6f_5U/edit#slide=id.gd12e9ba5d8_0_1

One Sentence Summary: The T2T-CHM13 reference genome universally improves read mapping and variant analysis.

Main Text:

Introduction

For the past twenty years, the reference human genome has served as the bedrock of human genetics and genomics [1–3]. One of the central applications of the reference human genome, and reference genomes in general, has been to serve as a substrate for comparative analyses and population genomics. More than one million human genomes have been sequenced to study genetic diversity and clinical relationships, and nearly all of them have been analyzed by aligning the sequencing reads from the donors to the reference genome, e.g. [4–6]. Even when the donor genomes are assembled *de novo* independent of any reference genome, the assembled sequences will nearly always be compared to a reference genome to catalog and characterize variation by leveraging the deep catalog of available annotations [7,8]. Consequently, genomics benefits tremendously from the availability of a high-quality reference genome, ideally without gaps or errors that may obscure important variation and regulatory relationships.

The current human reference genome, GRCh38, is used for countless applications, with rich resources available to visualize and annotate the sequence across cell types and disease states [3,9–12]. However, despite decades of effort to construct and refine the sequence, the reference human genome suffers from several major limitations that hinder comprehensive analysis. Most immediately, GRCh38 contains more than 100 million nucleotides that either remain entirely unresolved ('N' characters), such as the p-arms of the acrocentric chromosomes, or are represented with artificial models, such as the centromeric satellite arrays [13]. Furthermore, GRCh38 possesses 11.5 Mbp of unplaced and unlocalized sequences that are represented outside of the primary chromosomes [3,14]. These sequences are difficult to study, and many genomic analyses exclude these sequences to avoid identifying false variants and false regulatory relationships [6]. Relatedly, reports have persisted of major artifacts that arise when identifying variants relative to GRCh38, such as an apparent imbalance between insertions and deletions [15–17]. Overall, these errors and omissions in GRCh38 introduce biases in genomic analyses, particularly in the case of centromeres, satellites, and other complex regions.

Another major concern is how the reference genome may influence the analysis of variation across large cohorts for population and clinical genomics. Several large studies, such as the 1000 Genomes Project [18] and gnomAD [6], have provided valuable information about the extent of genetic diversity within and between human populations. For example, many analyses of Mendelian diseases and complex diseases use these catalogs of single nucleotide variants (SNVs), small indels, and structural variants (SVs) to rank and prioritize potential causal variants, using allele frequencies and other information to prioritize variants [19–21]. When evaluating these resources, the overall quality and representation of the human reference genome are important, if often overlooked, factors. Any gaps or errors in the sequence will obscure variation, as well as its contribution to the heritability of human phenotypes and disease. In addition to major omissions such as centromeric sequences or acrocentric chromosome arms, the current reference genome possesses errors and biases distributed across the chromosomes, including within genes of known medical relevance [22,23]. Furthermore, GRCh38 was assembled from multiple donors using clone-based sequencing, which creates an excess of artificial haplotype

structures that can subtly bias analyses [1,24]. Over the years, there have been some attempts to replace rare alleles from these haplotypes with more common alleles, but hundreds of thousands of artificial haplotypes and rare alleles remain to this day [25,26]. Increasing the continuity, quality, and representativeness of the reference genome is therefore crucial for improving genetic diagnosis, as well as for gaining a fundamental understanding of the complex relationship between genetic and phenotypic variation.

The recently assembled Telomere-to-Telomere (T2T) CHM13 assembly addresses many of the limitations of the current reference [27]. Specifically, the T2T-CHM13v1.0 assembly adds nearly 200 megabases of novel sequence and corrects many of the errors present in GRCh38. Here we demonstrate the impact of the CHM13 reference on variant discovery and genotyping in a globally-diverse cohort. This includes all 3,202 samples from the recently expanded 1000 Genomes Project sequenced with short reads [28] along with 17 samples from diverse populations sequenced with long reads [8,27,29]. From these data, we document universal improvements in read mapping and variant calling with major implications for population and clinical genomics including more than 2 million additional variants found in non-syntenic regions, XXX megabases of corrections found within syntenic regions, and 306 medically-important genes showing improved variant calls and interpretation. When combined with the other cellular and genomic resources available for CHM13, this genome assembly is poised to replace GRCh38 as the predominant reference for human genetics.

Results

I. Structural comparisons of GRCh38 and CHM13

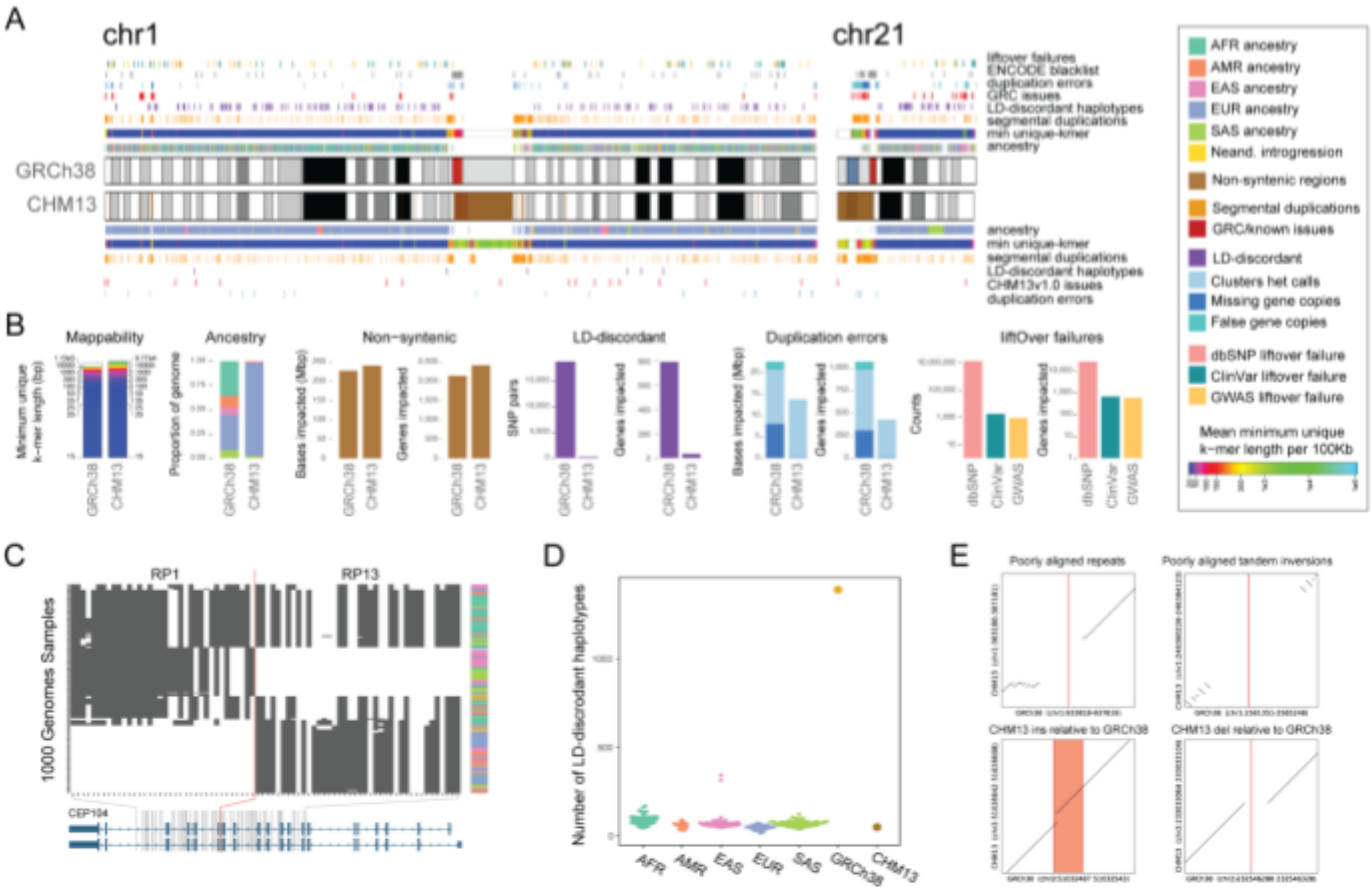

Figure 1. Genomic comparisons of human assemblies GRCh38 and CHM13. (A) Overview of annotations available for GRCh38 and CHM13 chromosomes 1 and 21 with colors indicated in legend. Cytobands are pictured as gray bands with red bands representing centromeric regions within ideograms. Complete annotations of all chromosomes can be found in **Figure S1.1**. **(B)** Summary of the number of bases and/or genes annotated by different features for the assemblies with colors indicated in the legend. **(C)** Example of a rare haplotype region, where variant density abruptly shifts ancestry of the individual at the boundary between clones in GRCh38. **(D)** Tally of rare haplotype regions in a selection of 1000 Genomes samples, as well as GRCh38 and CHM13. **(E)** Examples of variants that cannot be lifted over to CHM13 because of structural differences between the genomes. The position of the reference allele on GRCh38 is shown in red.

Introducing the T2T-CHM13 genome

The T2T-CHM13 reference genome was primarily assembled from Pacific Biosciences (PacBio) High Fidelity (HiFi) reads augmented with Oxford Nanopore Technology (ONT) reads to close gaps and resolve complex repeats [27]. The resulting T2T-CHM13v1.0 assembly was subsequently validated and polished to exceedingly high quality, with a consensus accuracy estimated to be between Phred Q67 and Q73 [27] and with zero known structural defects. This assembly is highly contiguous, with only five unresolved regions from the most highly repetitive ribosomal DNA (rDNA)

arrays, representing only 9.9 Mbp of sequence out of >3.0 Gbp of fully resolved sequence. The 1.0 assembly adds or revises 229 Mbp of sequence compared to GRCh38, defined as regions of the T2T-CHM13 assembly that do not linearly align to GRCh38 over a 1 Mbp interval (i.e., are non-syntenic). Furthermore, 189 Mbp of sequence are not covered by any primary alignments from GRCh38 and, therefore, are completely novel to the CHM13 assembly. A summary diagram of the syntenic/non-syntenic regions and their associated annotations are presented in **Fig. 1A** for chromosomes 1 and 21, and a detailed report of all chromosomes are presented in **Fig. S1.1**. Note that the subsequent T2T-CHM13v1.1 assembly further resolves the rDNA regions using model sequences of the array elements, although for this study we analyze the prior 1.0 assembly, which does not contain these representations.

The bulk of the non-syntenic sequence within T2T-CHM13 comprises centromeric satellites (180 Mbp), segmental duplications (SDs; 68 Mbp), and rDNAs (10 Mbp). These novel sequences could prove challenging for variant analysis, especially variants identified using short-read sequencing, although compared to GRCh38, we report an overall increase of 14.9 Mbp of additional unique sequence considering k-mers of length 50, 23.5 Mbp of unique sequences considering k-mers of length 100, and 39.5 Mbp of unique sequences considering k-mers of length 300, enabling confident mapping for short-read paired-end sequences or longer reads (**Figs. 1B, S1.2, and S1.3**). In terms of highly-repetitive sequence, more than 106 Mbp of additional sequence was identified in CHM13 that require sequences of more than 300 bp to uniquely map compared to GRCh38. Concomitantly, there was a decrease in long, exactly duplicated sequences (≥5 kbp) shared across chromosomes (excluding sequence pairs within centromeres) from GRCh38 to CHM13 (**Figs. S1.4 and1.5**). The former had 28 large shared interchromosomal sequences, primarily consisting of pairs of sub-telomeric sequence, with an additional 42 pairs involving at least one unplaced contig. All of these exactly matching sequence pairs, save one between two subtelomeres, are lost in T2T-CHM13 as small but important differences between repetitive elements are resolved for the first time. T2T-CHM13 contains another 81 sequence pairs of long repeats of which 77 are shared between acrocentric centromeres.

T2T-CHM13 accurately represents the haplotype structure of human genomes

The human reference genome serves as the standard to which other genomes are compared, and is typically perceived as a haploid representation of an arbitrary genome from the population [25]. In contrast with CHM13, which derives from a single hydatidiform mole, the Human Genome Project constructed the current reference genome via the tiling of sequences obtained from bacterial artificial chromosomes (BACs) and other clones with lengths ranging from ~50-150 kbp [24], which derives from multiple donor individuals. GRCh38 and its predecessors thus comprise mosaics of many haplotypes, albeit with a single library (RP11) contributing the majority [24].

To further characterize this aspect of GRCh38 and its implications for population studies, we performed local ancestry inference for both GRCh38 and CHM13 through comparison to haplotypes from the 1000 Genomes Project (see **Methods**; **Fig. 1A; Fig. S1.6**). Continental superpopulation-level ancestry was inferred for 72.9% of GRCh38 clones based on majority votes of nearest-neighbor haplotypes. For the remaining 27.1% of clones, no single superpopulation achieved a majority of nearest neighbors, and ancestry thus remained ambiguous. This ambiguity occurs primarily for short clones with few informative SNPs (**Fig. S1.7**), but also for some longer clones with potential admixed ancestry. In accordance with Green *et al.* [24], we inferred that library RP11, which comprises 72.6% of the genome, is derived from an individual of admixed African-American ancestry, with 56.0% and 28.1% of its component clones assigned to African and European local ancestries, respectively. The second most abundant library, CTD (5.5% of the genome), consists of clones of predominantly (86.3%) East Asian local ancestries, while the remaining libraries are derived from individuals of predominantly European ancestries. In contrast, CHM13 exhibits European ancestries nearly genome-wide (**Fig. S1.8**). In addition, GRCh38 and CHM13 harbor 26.7 Mbp and 51.0 Mbp, respectively, of putative Neanderthal-introgressed

sequences that originate from ancient interbreeding between the hominin groups approximately 60 thousand years ago [24]. The excess of introgressed sequence in CHM13, even when restricting to the same genomic intervals, is consistent with its greater proportion of non-African ancestry.

We hypothesized that the mosaic nature of GRCh38 would generate abnormal haplotype structures at clone boundaries, producing many combinations of alleles that are rare or absent from the human population. Indeed, some previous patches of the reference genome sought to correct abnormal haplotype structures in cases where they were noticed due to their impacts on genes of clinical importance (e.g., *ABO* and *SLC39A4*) [3]. Such artificial haplotypes would mimic rare recombinant haplotypes that are private to any given sample, but at an abundance and genomic scale that is unrepresentative of any living human. To test this hypothesis, we identified pairs of common (minor-allele frequency > 10%) autosomal SNPs that segregate in perfect linkage disequilibrium (LD; $R^2 = 1$) in the 2,504 unrelated individuals of the 1000 Genomes Project and queried their allelic states in both GRCh38 and CHM13 (see **Methods**).

In accordance with our expectations, we identified numerous haplotype transitions in GRCh38 that are absent from the 1000 Genomes samples, with 18,813 pairs of LD-discordant SNP alleles distributed in 1,390 narrow non-overlapping clusters (median length = 3,703 bp) throughout the genome (**Fig. 1C**). Such rare haplotype transitions are comparatively scarce in CHM13, with only 209 pairs of common high-LD SNPs (50 non-overlapping clusters) possessing allelic combinations that are absent from the 1000 Genomes sample (**Fig. 1D**). Based on a leave-one-out analysis, we confirmed that CHM13 possesses a similar number of LD-discordant haplotypes as phased "haploid" samples from the 1000 Genomes Project, whereas GRCh38 vastly exceeds this range (**Fig. 1E**). By intersecting the GRCh38 results with the tiling path of BAC clones, we found that 88.9% (16,733 of 18,813) of discordant SNP pairs straddle the documented boundaries of adjacent clones (**Fig. S1.9**). Of these, 45.9% (7,686 of 16,733) of the clone pairs derived from different BAC libraries, whereas the remainder likely largely reflects random sampling of distinct homologous chromosomes from the same donor individual. Our analysis thus highlights the artificial haplotype structure of GRCh38 and suggests that CHM13 accurately reflects haplotype patterns observed in contemporary human populations.

Identification of systematic errors in GRCh38, including collapsed duplications and falsely-duplicated regions

Genome assemblies often suffer from errors at complex genomic regions such as SDs. In the case of GRCh38, targeted sequencing of BAC clones have been performed to fix these loci [3,16,30–33], but problems persist. To systematically identify errors in GRCh38 that could produce spurious variant calls, we leveraged the fact that CHM13 is a haploid-derived cell line that should produce only homozygous variants when its sequence is aligned to GRCh38; thus, any apparent heterozygous variants can be attributed to (1) mutations accrued in the cell line; (2) sequencing errors; or (3) read mapping errors. In the last case, assembly errors or copy-number polymorphism of SDs could produce a contiguous stretch of heterozygous variants, which would confound paralog-specific variants (PSVs) as polymorphisms. Mapping PacBio HiFi reads from the CHM13 cell line [27] as well as Illumina-like simulated reads (150 bp) obtained from the CHM13 reference to GRCh38, we identified 368,574 heterozygous SNVs within autosomes and chromosome X, of which 56,413 (15.3%) were shared among both datasets, evidencing that each technology is distinctively informative due to mappability differences (**Fig. S1.10, Table S1.1**).

To home in on variants likely deriving from collapsed duplications, we delineated ‘clusters’ of heterozygous calls (see **Methods**) and identified 908 putative problematic regions (541 supported by both technologies) comprising 20.82 Mbp (**Fig. 1**) with a majority intersecting SDs (668/908; 73.6%) and centromere-associated regions (542/908; 59.7%) [34]. Variants flagged as excessively heterozygous in the population by gnomAD [6] were significantly enriched in these regions (10,000 permutations, empirical p-value = 1×10^{-4}) representing 26.7% (78,034/292,514) of our CHM13 heterozygous variants—suggesting that these variants arise in genome screens and represent false positives—whilst

341/908 regions (37.55%) overlap with known GRCh38 issues (**Fig. 1A**, **Fig. S1.1**). We next 'lifted over' 821 of these 908 putative problematic regions to the T2T-CHM13 assembly and used human copy-number estimates (n=268 individuals from the Simons Genome Diversity Project (SGDP)) [35,36] to conservatively identify 203 loci (8.04 Mbp) evidencing erroneous missing copies in GRCh38 (**Fig. S1.10**). These regions impact 308 gene features with 14 of the total 48 protein-coding genes fully contained within a problematic region, indicating that complete gene homologs are hidden from population variation analyses as is the case of *DUSP22*, a gene shown to play a role in immune regulation [37] (**Fig. S1.11**). Additionally, we identified 30 SNPs with known phenotype association (GWAS Catalog) within problematic regions. Finally, we evaluated the status of these regions in the T2T-CHM13 reference by following a similar approach as for GRCh38, obtaining 314,871 heterozygous variants clustered in 442 regions (**Table S1.2**). This revealed that 185 of the 203 problematic regions in GRCh38 no longer exhibit clusters of heterozygous variants in CHM13, enabling accurate variant calling in 46 protein-coding genes. We did identify eight putative collapsed duplications in CHM13 based on the presence of heterozygous variant clusters and reduced copy number in CHM13, but closer inspection revealed that all but one locus represented rare low copy number regions in CHM13. The remaining collapsed duplication localizes to an rDNA array corrected in the most recent version of T2T-CHM13v1.1 [Cite polishing paper].

Conversely, we found that the CHM13 reference fixed regions falsely duplicated in GRCh38. We identified 12 regions affecting 1.2 Mbp and 74 genes (including 22 protein coding genes) with duplications in GRCh38, but not in T2T-CHM13 or in 268 genomes from SGDP [36] (**Fig. S1.11**; **Table S1.3**). In contrast, only four regions affecting 18 kbp and one 142 kbp region have duplications in CHM13 that are not in GRCh38 and rarely (<1%) in SGDP. Inspecting the T2T-CHM13 assembly, we deemed these loci to be in fact rare duplications with strong support from mapped HiFi reads. The five largest duplications in GRCh38, affecting 15 protein-coding genes on the q arm of chromosome 21, have the falsely duplicated sequence in BAC clones misplaced between many gaps on the heterochromatic p-arm of chromosome 21. Of the seven false duplications outside chromosome 21, two are in short contigs between gaps, two are adjacent to a gap, two are on unlocalized "random" contigs, and one is a tandem duplication (**Table S1.4**). A list of falsely duplicated gene pairs in GRCh38 corrected in T2T-CHM13 is in **Table S1.5**. Using T2T-CHM13 as a reference provides an authoritative fix of these false duplications, enabling improved variant calling for short- and long-read technologies, including improved accuracy in medically relevant genes, as we demonstrate below.

Liftover of clinically relevant and trait-associated variation from GRCh38 to CHM13

In transitioning to a new reference genome, it is imperative to document the locations of known genetic variation of biological and clinical relevance with respect to the updated coordinate system. To this end, we sought to lift over 802,674 unique variants in the ClinVar database and 736,178,420 variants from the NCBI dbSNP database (including 151,876 NHGRI-EBI GWAS Catalog variants) from the GRCh38 reference to the CHM13 reference. Unambiguous liftover was successful for 789,909 (98.4%) ClinVar variants, 718,683,909 (97.6%) NCBI dbSNP variants, and 103,879 (68.4%) GWAS Catalog SNPs (**Fig. 1A, Supplementary Table X)**, which we provide as a resource for the scientific community, along with lists of all variants that failed liftover and the associated reasons. Critically, this resource includes 138,300 of 138,927 (99.5%) of Clinvar variants that were annotated as "pathogenic" or "likely pathogenic".

Of the 12,765 ClinVar variants that failed to lift over, 9,794 (76.7%) are cases in which the reference and alternative allele are exchanged in CHM13 versus GRCh38, along with an additional 1,581 liftover failures caused by scenarios where CHM13 possessed a third allele that matched neither the reference nor alternative allele of the respective variant. Of the remaining 1,390 ClinVar variants that failed to lift over, 844 overlap documented insertions or deletions that distinguish the GRCh38 and CHM13 assemblies. The remaining 546 variants (<0.1% of all variants) lie within regions of poor alignment between the GRCh38 and CHM13 assemblies (**Fig. 1F).** It should be noted that 98.5% of the GWAS Catalog variants that failed to liftover were cases in which the position of the variant exists in the CHM13 assembly, but the

reference allele in CHM13 does not match the reference allele in GRCh38. The modes of liftover failure for dbSNP variants and GWAS Catalog variants follow similar distributions and are presented in **Supplementary Table X.**

II. Improvements to Short Read Mapping and Variant Calling

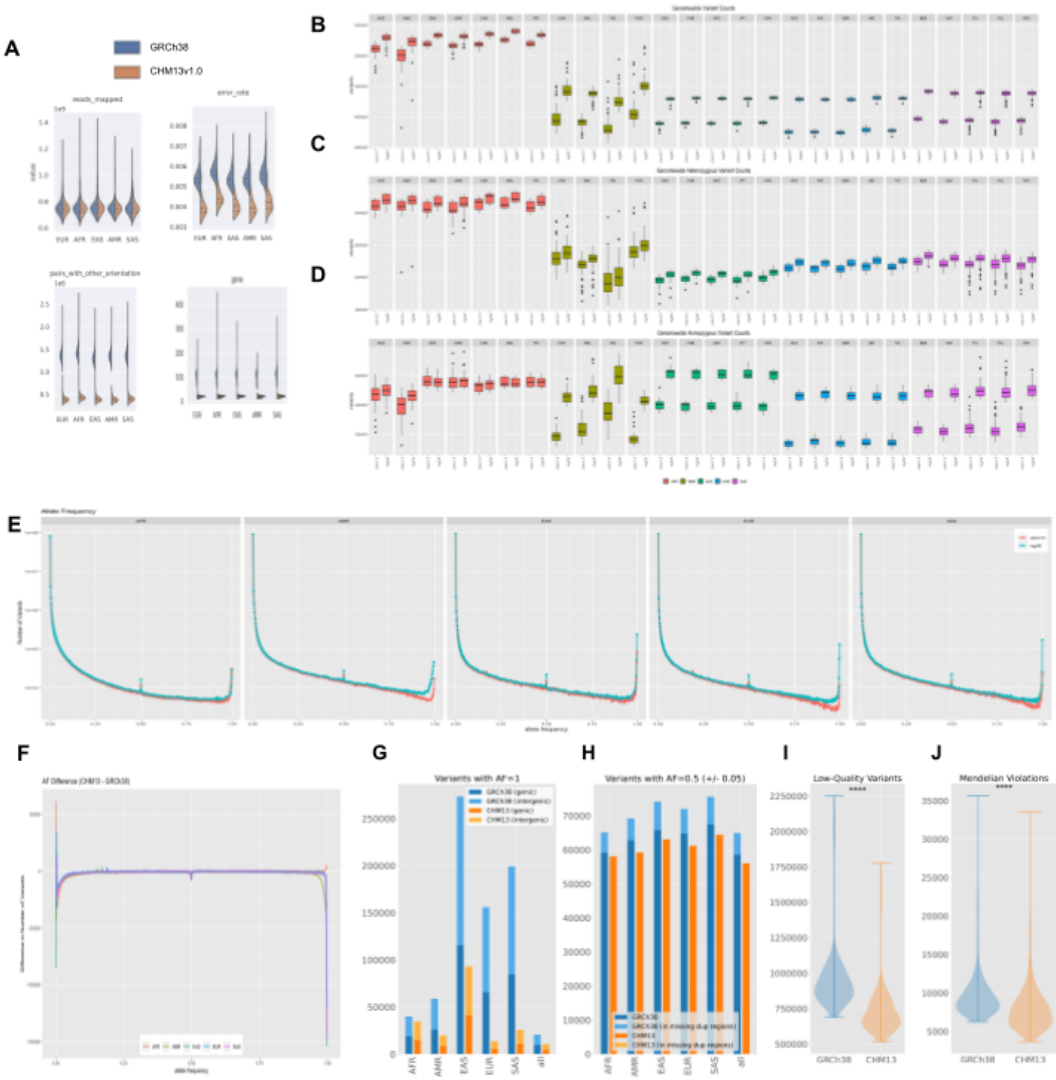

Figure 2. Improvements to Short Read Mapping and Variant Calling. (A) Summary of alignment characteristics aligning to CHM13 instead of GRCh38. **(B)** Boxplot of overall number of variants found in each person across populations. **(C)** Boxplot of the number of heterozygous variants found in each person across populations. **(D)** Boxplot of the number of homozygous variants found in each person across populations. **(E)** Allele frequency distribution of each of the major super populations relative to CHM13 and GRCh38. **(F)** Change in allele frequency distribution. **(G)** Number of variants with allele frequency equal to 100%, both within protein-coding genes and without. **(H)** Number of variants with allele frequency equal to 50%, both within TODO REGIONS and without. **(I)** Violin plot of the number of low-quality variants found when aligning to GRCh38 and CHM13. **(J)** Violin plot of the number of Mendelian violations found when aligning to GRCh38 and CHM13.

Remapping the 1000 Genomes dataset

To investigate how the T2T-CHM13 assembly impacts short-read variant calling, we realigned and reprocessed all 3,202 samples from the recently expanded 1000 Genomes Project cohort [28] using the NHGRI AnVIL Platform (Schatz et al. 2021). In this collection, each sample is sequenced to at least 30x coverage using paired-end Illumina sequencing, with samples from 26 diverse populations across 5 major continental superpopulations (**Fig. S2.1**). Most samples are unrelated, although the expanded collection includes 602 complete trios that we use to estimate the rate of inaccurate (non-Mendelian) variants below. We matched the previously used analysis pipeline for GRCh38 [28] as closely as possible so that any major differences would be because of the change in reference genome rather than because of technical differences in the analysis software used (**Supplemental**).

On average, an additional 7.4×10^6 (0.97%) of properly paired reads map to T2T-CHM13 rather than GRCh38 using bwa mem (Li 2013), even when considering the alt and decoy sequences used by the original analysis project. Interestingly, even though more reads align to T2T-CHM13, the observed per-read mismatch rate after alignment to T2T-CHM13 is 20% to 25% smaller for all continental populations, although African samples continue to present the highest mismatch rate (**Fig. 2a**). This is expected since the observed mismatch rate counts both genuine sequencing errors, which will be largely consistent across all samples, and any true biological differences between the read and the reference genome, which vary substantially by the ancestry of the sample. Relatedly, several other mapping characteristics were also improved using CHM13, such as a decline in the number of mis-oriented pairs in the paired-end sequencing (**Fig. 2a**). Finally, by considering the alignment coverage across 500bp bins across the respective genomes, we observed improved coverage uniformity within every sample using T2T-CHM13 rather than GRCh38. For example, within gene regions, we noted a 4-fold decrease in the standard deviation of the coverage (**Fig. 2a**), and observed similar improvements in other types of genomic regions among all population groups (**Fig. 2X**). Overall, these improvements in error rates, mapping characteristics, and coverage uniformity demonstrate how T2T-CHM13 is a superior reference genome for short read alignment for all populations and superpopulations.

Population-specific variant counts and allele frequencies

From these alignments, we next evaluated SNV and small indel variant calls with the widely-used GATK Haplotype Caller, which uses a joint genotyping approach to optimize accuracy across large populations [38]. Again, we matched the pipeline used in the prior study, albeit with newer versions of some of the tools, to minimize any differences due to software discrepancies so that we could focus on differences due to the change in reference genome. Across all samples, we found 57,315,786 high-quality ("PASS") variants relative to T2T-CHM13 compared to 59,051,733 relative to GRCh38, and noted a decrease in the number of variant calls per individual (**Fig. 2b**). We performed all subsequent analyses using these high-quality variants, as the PASS filter successfully removed a number of spurious variants (**Fig. S2.6**), particularly in complex regions (**Fig S2.7**). As with the improvement to the per-read mismatch rate, we attribute this to improvements in the number of rare alleles, consensus errors, and structural errors in CHM13 versus GRCh38. This conclusion is supported by the observation that the number of heterozygous variants per sample is more similar (**Fig. 2c**) across reference genomes as opposed to homozygous variants (**Fig. 2d**). This discrepancy is especially pronounced in non-African samples, which have substantially more homozygous variants relative to GRCh38 than CHM13. This is likely because ~70% of the GRCh38 sequence comes from an individual with African-American ancestry, which will have more rare and private variants [18].

Further investigating this relationship, we computed the allele frequencies of variants from unrelated samples from each of the 5 continental superpopulations (**Fig. 2e**). The distributions were nearly equivalent over most of the allele frequency spectrum, but substantial differences were observed in rare

variants (AF < 0.05), perfectly heterozygous variants (AF ≅ 0.5), and ultra-common variants (AF > 0.95). The most numerous difference in allele frequency was within the ultra-common variants, in which all non-African superpopulations showed an excess of ~150,000 variants in GRCh38, while the African superpopulation showed an excess of 2,364 variants in T2T-CHM13 (**Fig. 2f**). This category of variants is driven by a striking difference in the number of variants observed to have 100% allele frequency on GRCh38 compared to T2T-CHM13 (**Fig. 2g**). Such variants are representative of positions where the reference genome itself is the only sample observed to contain the corresponding allele, so that 100% of other samples show an alternative allele. As a result, these "variants" will not be reported when the same reads are mapped to a different genome that does not have this private allele. The increase in the number of rare variants relative to T2T-CHM13 is also explained by the haplotype structure of GRCh38: other than variants with AF equal to 1.0, variants observed as ultra-common in GRCh38 will most often present as rare variants in T2T-CHM13 with an allele frequency of 1.0 minus the rate observed in GRCh38. Moreover, the reduction in perfectly heterozygous variants is largely explained by the corrections to collapsed segmental duplications, as these regions are highly enriched for heterozygous paralog-specific variants (PSVs) caused by the false pileup of reads from the duplicated regions to a single location (**Fig. 2h**). Collectively, the decrease in variants with AF=1 and AF≅ 0.5 largely explains the decrease in the overall number of variants observed per sample and across the entire population.

Mendelian violations

As an additional quality control for the variant calls, we performed a Mendelian concordance analysis using the 602 trios represented in the 1000 Genomes Project cohort. From this, we found a statistically significant decrease in both the number of low-quality variants (median=890,701 (GRCh38) vs 682,609.5 (T2T-CHM13), p-value < 2.225×10^{-308}, paired t-test) (**Figure 2i**) and the number of Mendelian violations (median=8,879 (GRCh38) vs 7,484.5 (T2T-CHM13), p-value = 1.402×10^{-223}, paired t-test) (**Figure 2j**) when aligned to T2T-CHM13 as compared to GRCh38. In addition to providing an estimate of the error rate of the variant calls in this callset, this improvement will have broad implications for clinical genetics focused on *de novo* or somatic mutation analysis, such as in the analysis of autism spectrum disorders [39] or many forms of cancer [40].

Visibility and concordance of SNVs when using different reference genomes for variant discovery

Finally, to determine the impact of using the T2T-CHM13 reference for variant calling and then annotating the results using resources available for GRCh38, we used the Picard LiftoverVCF tool to lift single-sample (HG002) and multi-sample (1000 Genomes Project samples, or “1KG”) T2T-CHM13-based SNV calls over to the GRCh38 reference, and then compared the lifted variants to variant calls that were directly made using GRCh38. When lifted to GRCh38, the T2T-derived single-sample callset for HG002 did not include 30.4% (1,051,004) of the 3,460,148 GRCh38-derived SNV calls made using the GRCh38 reference. This was largely due to the 946,126 sites in HG002 that are homozygous for the CHM13 allele, and therefore are called only when using GRCh38 as the reference. Likewise, the set of SNVs called against T2T-CHM13 and then lifted to GRCh38 contained an additional 721,547 SNVs, which were not called against GRCh38, and this was due largely to the 672,341 SNVs for which HG002 is homozygous for the GRCh38 allele. When we compare only SNV sites for which HG002 has alleles, which are variant with respect to both reference genomes, the set of SNVs detected against T2T-CHM13 and then lifted to GRCh38 contains 95.8% of the SNV sites detected using GRCh38, and 98.0% of the SNV sites lifted from CHM13 are contained within the GRCh38 set.

While the difference in visibility of a single sample’s variants with respect to the two reference sequences is significant, variant calls for larger multi-sample sets are less affected since large sets of samples are more likely to contain detectable

variants with respect to both references in places where the two references exhibit different alleles. The 1KG multi-sample callset lacks far fewer of the GRCh38 calls due to the absence of an alternate allele with respect to T2T-CHM13: only 13,001 SNVs of the 106,581,271 GRCh38 1KG SNV calls were not detected with the T2T-CHM13 reference for this reason. Conversely, 12,124 1KG SNVs that were originally called using T2T-CHM13 were not detected when using GRCh38 due to a lack of 1KG samples with an allele that differs from the GRCh38 allele.

III. Improvements to Long Read Mapping and Variant Calling

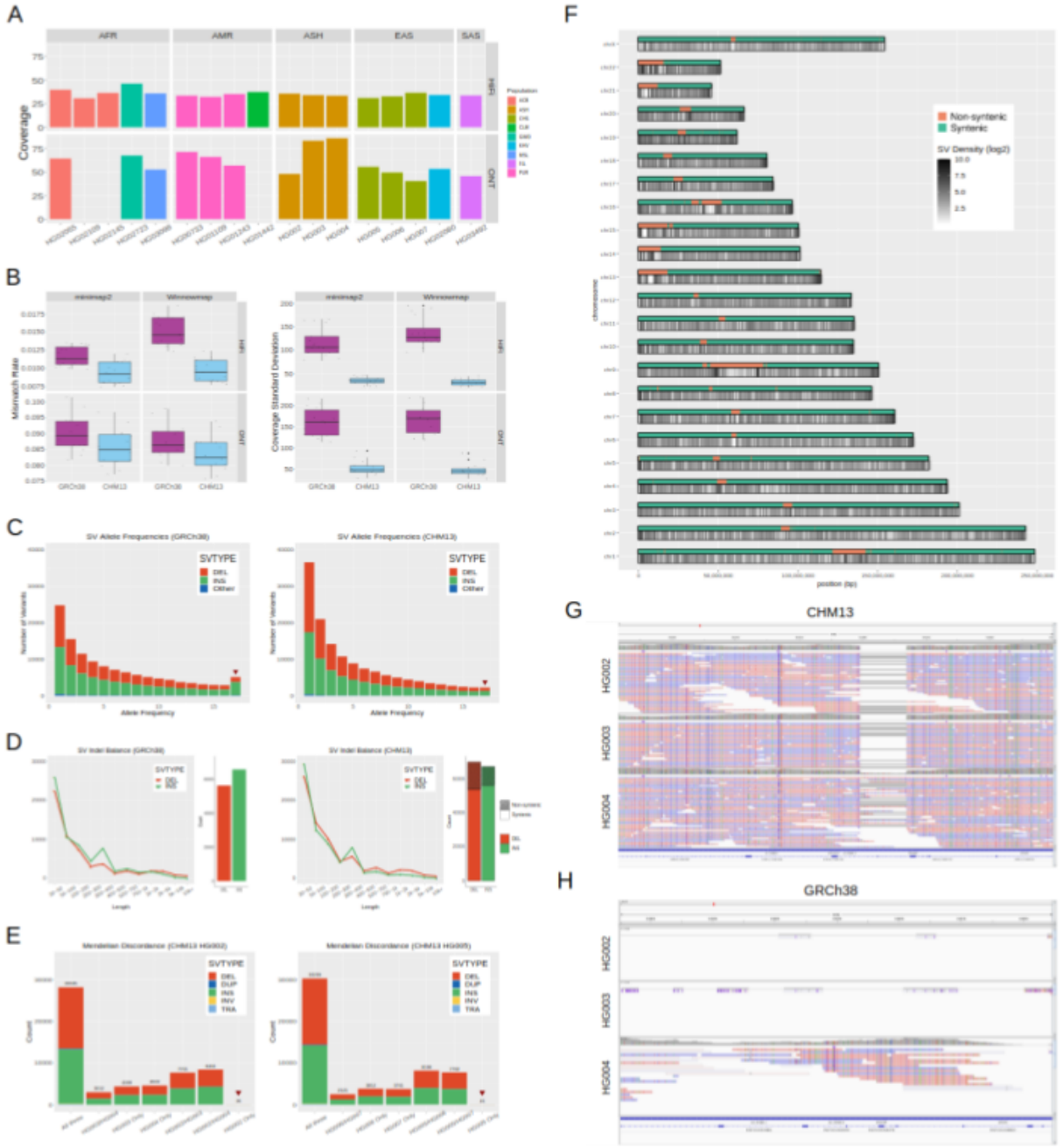

Figure 3. Improvements to Long-Read Alignment and SV Calling in CHM13

(A) The coverage, ancestry, and sequencing platforms available for the 17 samples sequenced with long reads. **(B)** The genome-wide mapping error rate and the standard deviation of the coverage for CHM13 and GRCh38. The standard deviation was computed across each 500bp bin of the genome. **(C)** The allele frequency of SVs derived from HiFi data in CHM13 and GRCh38 among the 17-sample cohort. The red arrows indicate uniform (100% frequency) variants. **(D)** The balance of insertions vs. deletion calls in the 17-sample cohort in CHM13 and GRCh38. Variants in CHM13 are stratified by whether or not they intersect regions which are non-syntenic with GRCh38. **(E)** The SV calls in CHM13 for two trios: a trio of Ashkenazi ancestry (child HG002, and parents HG003 (46XY), and HG004 (46XX)), and a trio of Han Chinese ancestry (child HG005, and parents HG006 (46XY) and HG007 (46XX)). The red arrows indicate child-only, or candidate *de novo*, variants. **(F)** The density of SVs called from HiFi data in the 17-sample cohort across CHM13. **(G)** Alignments of HiFi reads in the HG002 trio to CHM13 showing a deletion spanning an exon of the transcript AC134980.2. **(H)** Alignments of HiFi reads in the HG002 trio to the same region of GRCh38 as shown in (g), showing much poorer mapping to GRCh38 than to CHM13.

Improvements to long-read mapping

Next, we investigated the effects of using T2T-CHM13 as a reference genome for alignment and large structural variant (SV) calling from genomic long reads from both PacBio HiFi and ONT long reads. To study this, we aligned reads and called SVs from long reads in seventeen samples of diverse ancestry from the Human Pangenome Reference Consortium (HPRC+) [27] and the Genome-in-a-Bottle Consortium (GIAB) [29], including two trios (**Table 3.1**). All of these samples had HiFi data available, and fourteen of them had also been sequenced with ONT (**Fig. 3a**), with mean read lengths of 18.1kbp and 21.9kbp, respectively (**Fig. S3.1**).

We first aligned all long reads to both T2T-CHM13 and GRCh38 with both Winnowmap [41] and minimap2 [42] so that we could compare alignment statistics and use these alignments for subsequent SV calling. Similar to the short read results, the number of reads mapped did not substantially change because most of the novel sequence in T2T-CHM13 had some available model sequences in GRCh38 (**Fig. S3.2**). However, aligning to T2T-CHM13 reduced the observed per-read mapping error rates across by 5% to 40% across the four combinations of sequencing technology and aligner because of the correction of errors in GRCh38 and the addition of improved mapping targets, similar to what we observe for short reads (**Fig. 3b**). Next, to study coverage uniformity, we measured the average coverage across each 500bp bin on a per-sample basis, and computed the standard deviation of the coverage among all bins. Across all aligners and technologies, the median standard deviation of the per-bin coverage was reduced by more than a factor of three, indicating more stable mapping to T2T-CHM13 (**Fig. 3b**). This difference in coverage uniformity was particularly present in satellite repeats and other regions of GRCh38 not syntenic with T2T-CHM13 (**Fig. S3.3**).

Improvements to SV calling

We next used our optimized SV-calling pipeline including Sniffles [43], Iris, and Jasmine [44] to call SVs in all 17 samples and consolidate them into a cohort-level callset in each reference from HiFi data. From these results, we observe a reduction from 5,147 to 2,260 SVs with 100% allele frequency when calling variants relative to T2T-CHM13 instead of GRCh38 (**Fig. 3c**). Previous studies [16,17] have noted the excess of such SV calls when using GRCh38 as a reference and attributed them to structural errors, so the use of a complete and accurate reference genome sequence reduces the number of such variants. In addition, the number of insertions and deletions are more balanced when calling against T2T-CHM13, whereas GRCh38 had a bias towards insertions caused by missing or incomplete sequence (**Fig. 3d**). In T2T-CHM13, there is a small bias towards deletions, which is especially prominent in highly repetitive regions such as centromeres and satellite repeats (**Fig. S3.8**). The variants we observe relative to T2T-CHM13 are enriched in the centromeres and telomeres (**Fig. 3f**), likely because of a combination of repetitive sequence which is difficult to align to and greater recombination rates [17]. We observe similar trends among SVs unique to single samples (**Fig. S3.11**)

Relatedly, we similarly observe better insertion/deletion balance for large SVs >500 bp detected by Bionano Optical Mapping data in HG002 against the T2T-CHM13 reference, with an increase in deletions from 1,199 to 1,379 and a decrease in insertions from 2,771 to 1,431 on GRCh38 and T2T-CHM13, respectively (**Fig. S3.14**). Using the CHM13 reference for Bionano also enables improved SV calling around gaps in GRCh38 that are closed in CHM13 (**Fig. S3.15**).

De novo SV analysis

To investigate our ability to detect *de novo* variants in CHM13 and to quantify our false positive rate, we called SVs in both of our trio datasets using a combination of HiFi and ONT data and identified SVs only present in the child of the trio and supported by both technologies, about 40 variants per trio (**Fig. 3e**). We then manually inspected the variants, and found a few strongly-supported variants in each trio with consistent coverage and alignment breakpoints, while the other

candidates had less reliable alignments as seen in previous reports [44]. In HG002, we detected 6 strongly-supported candidate *de novo* SVs which were previously reported [29,44]. In HG005, we detected a novel strongly-supported candidate *de novo* SV relative to both T2T-CHM13 and GRCh38, a 1571bp deletion at chr17:49401990 in T2T-CHM13 (**Fig. S3.9**).

Novel SVs

The improved accuracy and completeness of the T2T-CHM13 genome enables us to better resolve complex regions where reads are unable to align to GRCh38. Within non-syntenic regions, we find a total of 27,055 SVs (**Fig. 3d**), the majority of which were insertions (10,912) and deletions (15,998). The allele frequency and size distribution for these variants follow the characteristics of the syntenic regions, with rare variants (**Fig. S3.10**) and small indels (**Fig. S3.7**) being the most abundant. However, we also note some non-syntenic regions with relatively few or no SVs identified, especially within the interior of the newly resolved p-arms of the acrocentric chromosomes (**Fig. 3f**). We attribute the reduction in variant density largely to the mappability of the reads, which are difficult to place in the most complex regions of the genome. We expect this will improve in time as read lengths and alignment algorithms continue to improve.

Furthermore, within syntenic regions we also observe notable improvements to alignment and variant calling accuracy. For example, in T2T-CHM13 we observe a deletion in all of the samples in the HG002 trio of an exon of the olfactory receptor gene AC134980.2 (**Fig. 3g**), while the reads from those samples are largely unable to align to the surrounding region in GRCh38 (**Figure 3h**). At the same time, reads from that region in African samples (**Fig. S3.13**) align to both references. The difference in alignment among different samples is likely due to the region being highly polymorphic for copy number variation; GRCh38 contains a fair representation of that region in African samples, while that region in T2T-CHM13 more closely resembles European samples (**Fig. S3.12**). This highlights the need for reference genomes consisting of individuals with similar ancestry to samples being studied, and motivates efforts to produce similarly complete genome assemblies for many diverse individuals.

IV. Detailed analysis of novel regions / Evolutionary Genomics

A Table here: Number of variants in novel regions

	Total span (bp) [Excluding Ns]	Unique span (100mers)	Protein-coding genes	Short-read SNVs + indels [Within genes]	Long-read SNVs [Within genes]	Short-read SNVs confirmed by long reads	Long-read SNVs identified in short reads	SNVs concordant between long reads	High-confidence region bases	Short-read SNVs confirmed by long reads in high-confidence regions	Long-read SNVs identified in short reads in high-confidence regions	SNVs concordant between long reads in high-confidence regions
Non-syntenic	240,044,315 [228.57]	65,471,196	207	2,050,129 [67,492]		73-78%	4-5%	41-43%	13,683,528	95-96%	60-63%	91-95%
Novel sequences	189036735 []	40,205,401	20777	1,445,506 [35,392]		64-69%	3%	38-40%	2,987,935	84-88%	39-46%	81-90%

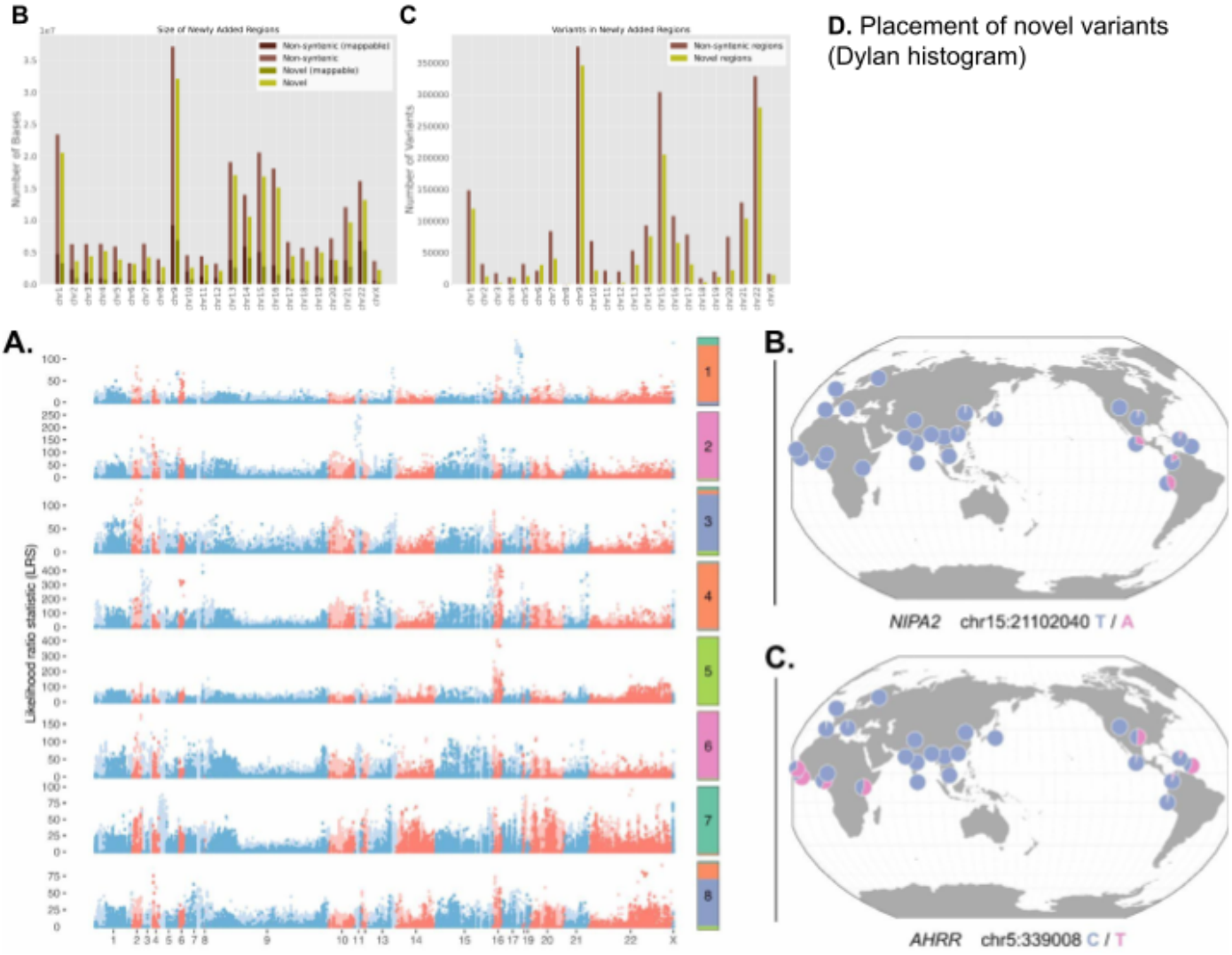

Figure 4. Characterization of novel short read variants
(A) Overview of non-syntenic regions and their variant counts **(B)** Number of bases added in non-syntenic and mappable regions by chromosome, along with how many variants for each respective region are mappable (have contiguous unique 100mers). **(C)** Number of variants in these newly added regions by chromosome. **(D)** TODO: placement of novel regions **(E)** Scan for variants in non-syntenic regions that exhibit extreme patterns of allele frequency differentiation. Allele frequency outliers were identified for each of eight ancestry components, colored by the superpopulation membership of corresponding 1000 Genomes samples. Large values of the likelihood ratio statistic (LRS) denote variants for which variance among populations exceeds that of a null model based on genome-wide covariances in allele frequencies.

Variants within newly accessible and corrected regions

The T2T-CHM13 genome contains 229 Mbp of sequence that is non-syntenic to GRCh38, which contains 207 genes (**Fig. 4a**). Within these regions, we report a total of 2,050,129 PASS variants across all 1000 Genomes samples from short reads. 73-78% of the SNVs in each sample match variants found in each sample's respective PacBio HiFi long read data using the leading PEPPER-Margin-DeepVariant algorithm (51,306-74,122 matching SNVs and genotypes per sample) [45]. Long reads discover over 10 times more SNVs per sample than short reads in these regions, with 447,742-615,085 (41-43%) of SNVs matching between HiFi and ONT. We further define higher-confidence regions by excluding regions with abnormal coverage in any long-read sample, which effectively excludes difficult to map regions and copy number variable regions. Within these regions, 14 Mbp remain, and SNVs from HiFi and ONT long reads are 91-95% concordant (21,835-28,237 variants). 95-96% (14,575-18,949) of short-read SNVs are found in HiFi long-read calls, though 37-40% of HiFi SNVs are still missing from the short-read calls (**Supplemental Table S4.1**).

As these broadly-defined non-syntenic regions include inversions and other large structural changes between GRCh38 and T2T-CHM13 that won't necessarily alter many of the variants contained within, we also considered novel sequences, defined as segments of the T2T-CHM13 genome that do not align to GRCh38 using Winnowmap [41]. Within these novel sequences, we report a total of 1,445,506 PASS variants in 1000 Genomes samples from short reads across 189 Mbp (**Fig. 4a**). Because these novel sequences are more enriched for highly repetitive sequences, concordance is slightly lower, such that 64-69% of the SNVs in each sample match variants found in each sample's respective PacBio HiFi long read data (24,371-36,501 matching SNVs and genotypes per sample), and 339,783-473,074 (38-40%) of SNVs matching between HiFi and ONT. When removing difficult to map and copy number variable regions as above, 3 Mbp remain, and 84-88% of the SNVs in each sample match variants found in each sample's PacBio HiFi long-read data (2,938-3,811 matching SNVs and genotypes per sample), and 5,544-8,298 (81-90%) of SNVs match between HiFi and ONT (**Supplemental Table S4.1**).

We noted that homologs of GRCh38 collapsed duplications (137 regions and 6.8 Mbp) are associated with these T2T-CHM13 non-syntenic and/or novel regions, representing potentially corrected sequences in this new assembly. To assess if these corrected complex regions in T2T-CHM13 improve our ability to identify variation, we evaluated the concordance of variants generated from sequencing and variant calling from short-read Illumina and PacBio HiFi sequencing datasets of two trios from the GIAB and Personal Genome Project. Within corrected false duplications (1.2 Mbp), we saw dramatically improved recall for Illumina data compared with HiFi in T2T-CHM13 (57.4-68.3%) versus GRCh38 (1.1-1.8%), as well as improved precision in T2T-CHM13 (98.5-99.3%) versus GRCh38 (76.5-95.8%). We also compared variant calling in a subset of the GRCh38 collapsed duplications (CN<10; ~910 kbp) and observed similar recall for Illumina data in T2T-CHM13 (20.1-28.3%) and GRCh38 (21.5-25.4%), but with significantly improved precision in the variants identified (98.1-99.7% in T2T-CHM13 vs. 64.3-67.3% in GRCh38). This is likely due to 5-9x more variants discovered at these loci in GRCh38 that represent PSVs or misassigned alleles from the missing paralog (**Supplemental Table S4.2**). This shows that, even though we have reduced ability to discover variants in these SDs with short read Illumina data, variants that can be identified are now reliably genotyped in this new assembly.

Allele frequency differentiation across populations in newly accessible regions

The CHM13 assembly adds more than 100 Mb of accurately resolved sequence not represented in CHM13 — equivalent to an additional chromosome. Although some of this new sequence remains difficult to access with short reads (e.g., rDNA arrays, highly repetitive regions), much of it represents new sequences available for variant calling. Interestingly, many of the short-read variants are within XXX kbp of known clinically relevant genes (**Figure 4c**)

Using short-read variant genotypes generated with the 1000 Genomes cohort, we applied a frequency differentiation-based method to identify signatures of positive selection within novel regions of the CHM13 assembly. This method, Ohana [46], models individuals as possessing ancestry from *k* ancestry components and tests whether individual variant allele frequencies are better explained by this genome-wide null model, or by an alternative model where frequencies are allowed to vary in one population by consequence of positive selection or demographic forces.

Using this method to investigate continental-scale patterns ($k = 8$), we identified 5,154 unique SNPs and indels across all ancestry components that exhibited strong deviation from genome-wide patterns of allele frequency differentiation (99.9th percentile of distribution for each ancestry component; see **Methods**; **Supplementary Fig. X**). These included 814 variants overlapping with annotated genes, and 151 that intersected with annotated exons. Some of the most highly differentiated variants include <SNPs overlapping with olfactory receptor genes on chromosome 7; explanation of why this locus was previously inaccessible>. These variants reach intermediate allele frequencies in the ancestry component corresponding to the Peruvian (PEL), Mexican (MXL), and Colombian (CLM) populations, but are nearly absent in other populations (allele frequencies: 0.29 in PEL; 0.005 in South Asia; absent elsewhere; **Fig. 4A**).

V. Impact of CHM13 on Clinical Genomics

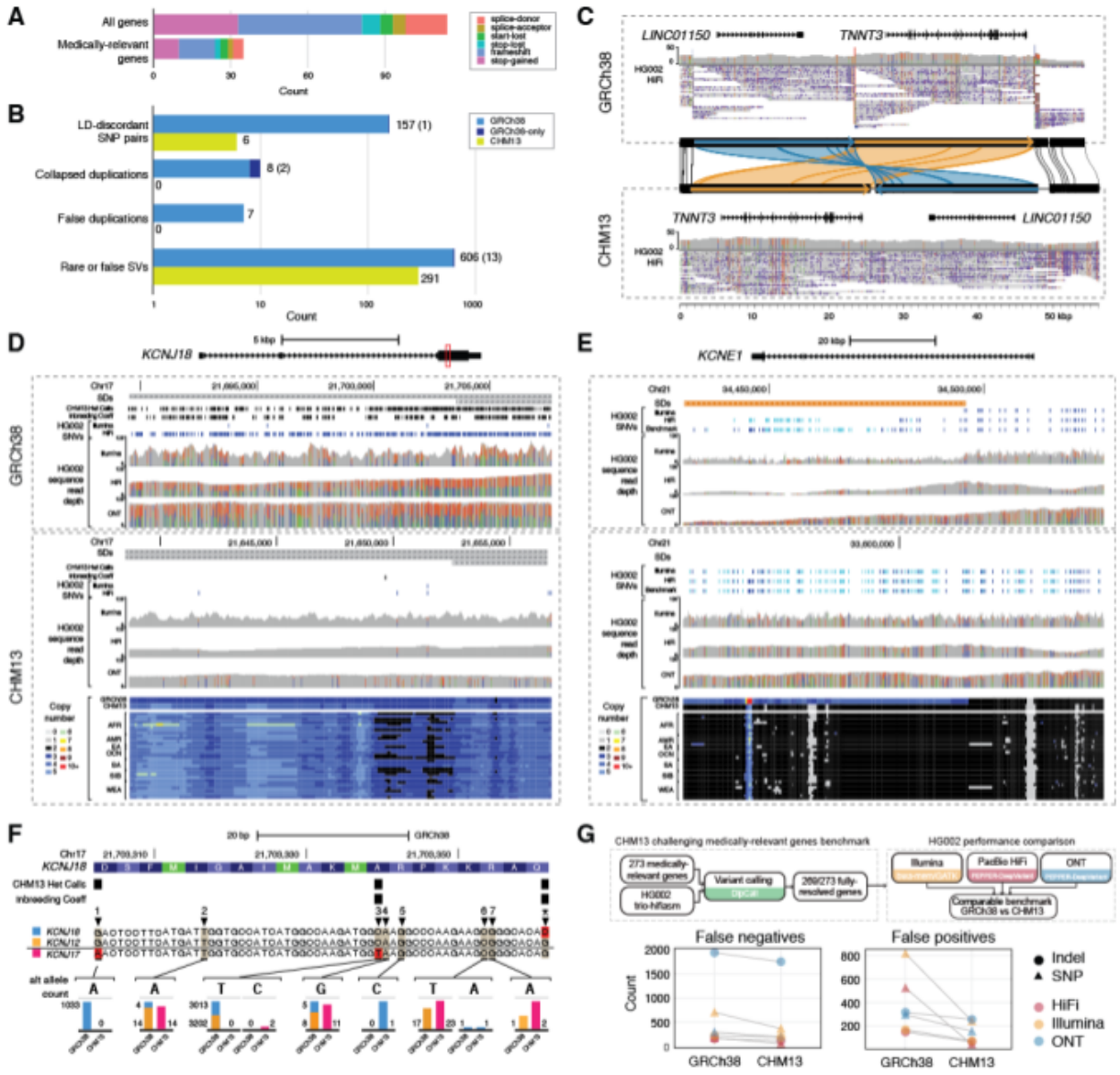

Figure 5. CHM13 Improves Clinical Genomics Variant Calling. (A) Potential loss-of-function mutations (displayed as different colors) exist within genes in the CHM13 assembly at the same scale as found in other human genomes (**Table S5.1**). **(B)** The counts of medically-relevant genes impacted by genomic features and variation in GRCh38 (blue) and CHM13 (green) are depicted in box plots on logarithmic scale. Dark blue indicates genes impacted in GRCh38 where homologous genes were not identified in CHM13 due to inability to liftover, with counts included in parentheses. **(C)** An example complex SV found as a deletion and inversion in all individuals in GRCh38 that has been corrected in CHM13 impacting *TNNT3* and *LINC01150*, displayed by sequence comparison using miropeats with homologous regions colored in orange and blue, respectively. HG002 PacBio HiFi data is displayed showing read coverages and mappings, which are discordant in GRCh38 at the SV breakpoints that are remedied in CHM13. **(D)** Collapsed duplication in GRCh38 corrected in CHM13 impacting *KCNJ18*, causing many reads derived from the missing duplication (*KCNJ17*, not pictured) from all technologies to map incorrectly, leading to spurious variant calls. **(E)** A false duplication in GRCh38 impacting most of *KCNE1*, causing many reads from all technologies to map incorrectly to the false gene copy *KCNE1B* on GRCh38 (not pictured). **(F)** An example coding region of *KCNJ18*, showing alignments of *KCNJ18* (blue), *KCNJ12* (orange), and *KCNJ17* (red) along with allele counts of variants in each gene identified on GRCh38 and CHM13, with examples 1-7 explained in the Results. **(G)** Schematic depicts a new benchmark for 269 challenging medically relevant genes for HG002. The number of variant calling errors from three sequencing technologies on each reference are plotted.

Potentially clinically-relevant variants in CHM13

A deleterious variant in a reference genome can mislead the interpretation of a clinical variant identified in a patient because it may not be flagged as such using standard analysis tools. The GRCh38 reference genome is known to contain such variants that likely affect gene expression, protein structure, or protein function [25], though systematic efforts have sought to identify and remove these alleles [3]. To determine loss-of-function of variants exist in T2T-CHM13, we aligned the assembly to GRCh38 using dipcall [47] to identify and functionally annotate nucleotide differences [48] (**Fig. 5A**). This identified 210 putative loss-of-function variants (defined as variants impacting protein-coding regions and predicted splice sites) impacting 189 genes, 31 of which 31 are clinically relevant [23]. These results are in line with previous work showing that the average diploid human genome contains ~450 putative loss-of-function variants impacting ~200-300 genes using low-coverage Illumina sequencing (before stringent filtering) [49]. Of these 210 variants, 158 have been identified in at least one individual from the 1000 Genomes Project with most variants relatively common in human populations (median allele frequency of 0.47), suggesting they are functionally tolerated. The remaining variants not found in 1000 Genomes individuals comprise larger indels, which are more difficult to identify using 1000 Genomes Illumina data, and CHM13-rare or -unique alleles. We curated those impacting medically-relevant genes and found seven that likely derived from duplicate paralogs; one 100-bp insertion also found in long reads of HG002; a stop gain in a final exon in one gnomAD sample; and an insertion in a homopolymer in a VNTR in *CEL, which* may be an error in the assembly. Understanding that the published T2T-CHM13 assembly represents a human genome harboring potentially functional or rare variants that in turn would affect the ability to call variants at those sites, we have made available this list of variants to aid in clinical interpretation of sequencing results **(Table S5.1)**.

Improvements in CHM13 impacting medically relevant genes

We sought to understand how changing from GRCh38 to the CHM13 reference might impact variants identified in a previously compiled [23] set of 4,964 medically relevant genes residing on human autosomes and chromosome X (representing 4,924 genes in T2T-CHM13 via liftover; **Table S5.2**). Of these genes, 28 map to newly-accessible regions in CHM13 (novel and/or non-syntenic). Focusing on genes impacted by rare or erroneous structural alleles that could affect variant discovery or interpretation (e.g., collapsed duplications, false duplications, and rare haplotype boundaries), we found over half as many medically relevant genes at these loci in CHM13 (n=306) versus GRCh38 (n=756 including 14 with no CHM13 liftover) (**Fig. 5B**). Of these genes, 28 fall within regions flagged as erroneous in GRCh38 by the Genome Reference Consortium (GRC). The majority of impacted clinically relevant genes overlap SVs that exist in all 13 HiFi-sequenced individuals, likely representing rare alleles or errors in the reference (see above), including 13 of the 14 GRCh38 genes with no CHM13 liftover. One interesting example, *TNNT3*, which encodes Troponin T3, fast skeletal type implicated in forms of arthrogryposis [50], was impacted by both a 24-kbp inversion and 22-kbp upstream deletion that also ablates *LINC01150* in all individuals using GRCh38 (**Fig. 5C**). The GRC determined that a problem existed with the GRCh38 reference in this region (GRC issue HG-28). Comparison of this region in CHM13 instead shows a complex rearrangement with the 22-kbp region upstream of *TNNT3* inversely transposed in the new assembly to the proximal side of the gene. Besides potentially affecting interpretations of gene regulation, this structural correction of the reference places *TNNT3* >20 kbp closer to its genetically-linked partner *TNNI2* [51]. Other genes have VNTRs that are collapsed in GRCh38, such as one expanded by 17 kbp in most individuals in the medically relevant gene *GPI*. *MUC3A* was also flagged with a whole-gene amplification in all individuals, which we identified as residing within a falsely-collapsed SD in GRCh38, further evidencing that finding (**Fig. 1A**).

When intersecting erroneous duplicated and collapsed regions in GRCh38 (**Table S1.1**), we find 17 medically relevant genes residing in these loci, including *KCNE1* (false duplication) and *KCNJ18* (collapsed duplication) (**Fig. 5D and E**). Examining the falsely-duplicated *KCNE1*, we find that coverage is much lower than normal on GRCh38 for short and

long reads and that most variants are missed because many reads incorrectly map to the false duplication (*KCNE1B* on the p arm of chromosome 21). The kmer-based copy number of this region in all 266 SGDP genomes supports the CHM13 copy number, and this region was not falsely duplicated in GRCh37 either [23]. As for *KCNJ18*, which resides within a GRCh38 collapsed duplication at chromosome 17p11.2 and includes nine "pathogenic" or "likely pathogenic" variants in ClinVar related to thyrotoxic periodic paralysis [52], we find increased coverage across HG002 Illumina, PacBio HiFi, and ONT sequences in GRCh38 relative to CHM13. Across the 49 kbp SD containing *KCNJ18*, we discovered considerably fewer variants in CHM13 (n=2566) vs. GRCh38 (n=3476) in each 1000 Genomes Illumina dataset, likely due to false heterozygous calls from PSVs (**Fig. S5.1 and S5.2**).

To verify this, we compared the distributions of minor-allele frequencies and observed a shift in variant proportions, with a relative decrease in intermediate and a relative increase in rare alleles for *KCNJ18* and *KCNJ12* (another collapsed duplication residing distally at chromosome 17p11.2) in CHM13 compared with GRCh38 (p=0.08885 and 0.03102, respectively; Mann-Whitney U test with continuity correction). We used liftover to define precisely alternative assignment of variants across homologous sites at each loci, including the missing paralog *KCNJ17* (previously-missing copy located in a centromere-associated region on chromosome 17p in CHM13 denoted *KCNJ18-1*) (**Fig. 5F**) and observed that even true variants (i.e., non-PSVs) had discordant allele counts in *KCNJ18* and *KCNJ12* between the two references. Examples of such variants impacting the CDS of these genes are highlighted in **Fig. 5F** and labeled as 1-7, representing different categories: (1) false positives in *KCNJ12* and/or *KCNJ18* on GRCh38 resulting from PSVs in *KCNJ17*, which are corrected in CHM13; (2) and (4) variants erroneously detected in *KCNJ12* and *KCNJ18* on GRCh38 that actually are variants in *KCNJ17*; (3) false positives in *KCNJ12* and/or *KCNJ18* on GRCh38 resulting from PSVs in *KCNJ17* that also hide true variants in *KCNJ17*, *KCNJ12*, and/or *KCNJ18*; (5) variants not detected in *KCNJ12* or *KCNJ18* on GRCh38, likely due to skewed allele balance from mis-mapped reads; (6) and (7) variants erroneously detected in *KCNJ12* and/or *KCNJ18* on GRCh38 that actually are variants in *KCNJ17* with the allele detected in fewer haplotypes on GRCh38, likely due to skewed allele balance from mis-mapped reads. There are also more complex sites, such as the last base in **Fig. 5F** labeled with an asterisk, in which the reference base for *KCNJ18* changes to C in CHM13 in addition to novel alleles identified in KCNJ17. In summary, T2T-CHM13's improved representation of this gene and other collapsed duplications not only eliminates false positives but also improves detection and genotyping of true variants.

Clinical gene benchmark

Finally, to determine how the T2T-CHM13 genome improved our ability to assay variation broadly, we used a curated diploid assembly to develop a new benchmark for 269 challenging medically relevant genes in GIAB Ashkenazi son HG002 [23], with comparable benchmark regions on GRCh38 and T2T-CHM13. We tested three short- and long-read variant callsets against this benchmark: Illumina-BWA-MEM-GATK, HiFi-PEPPER-DeepVariant, and ONT-PEPPER-DeepVariant.

Both numbers of false positives and false negatives substantially decrease for all three callsets when using CHM13 as a reference instead of GRCh38 (**Fig. 5G**). False positives for HiFi decrease by a factor of 10 in these genes due primarily to adding missing sequences similar to KMT2C and removal of false duplications of CBS, CRYAA, KCNE1, and H19 (**Supplementary Fig IGV?**). As demonstrated above, these genes and others are better represented in CHM13 than GRCh38 for a diverse set of individuals, so performance should be higher across diverse ancestries. The number of true positives also decreases by a much smaller amount than the errors, about xx %, due to a smaller number of homozygous variants, as seen in individuals of European ancestry in **Fig. 2**. This benchmarking demonstrates concrete performance gains in specific medically relevant genes resulting from the highly accurate assembly of a single individual.

Discussion

Persistent errors in the human reference genome, ranging from collapsed duplications to missing sequences, have silently plagued human genetics for decades. The strong and arguably misguided assumption that most genomic analyses make about the correctness of the reference genome has led to spurious clinical findings and mistaken disease-associated variants. Here, we show that the T2T-CHM13 reference genome universally improves genome analyses for all populations by correcting major structural defects and adding missing sequences to GRCh38. In particular, we show that the T2T-CH13 assembly (1) revealed millions of additional variants and the existence of new copies of medically relevant genes (e.g., KCNJ17) within the 240 Mbp and 189 Mbp of non-syntenic and novel sequence in T2T-CHM13; (2) eliminated tens of thousands of spurious variants and incorrect genotypes per samples, including within medically-relevant genes (e.g., KCNJ18) by expanding 185 loci (6.84 Mbp) that were falsely collapsed in GRCh38; (3) improved genotyping by eliminating 12 loci (1.2 Mbp) that were falsely duplicated in GRCh38 (4) yielded more comprehensive SV calling genomewide, with a greatly improved insertions/deletion balance by correcting collapsed tandem repeats. Overall, the T2T-CH13 assembly reduced false positive and false negative SNVs from short and long reads by as much as 12-fold in challenging, medically relevant genes. The T2T-CHM13 reference also accurately represents the haplotype structure of human genomes, eliminating 1390 artificial recombinant haplotypes in GRCh38 that occurred as artifacts of BAC clone boundaries. These improvements will broadly enable new discoveries and refined analysis across all of human genetics and genomics.

Given these advances, we anticipate a rapid transition to the T2T-CHM13 genome. On a practical level, improvements to large genome segments, such as entire p-arms of the acrocentric chromosomes, and discovering new clinically relevant genes and diseases causing variants justify the labor and cost required to incorporate T2T-CHM13 into basic science and clinical genomic studies. On a technical level, T2T-CHM13 simplifies genome analysis and interpretation because it consists of 23 complete linear sequences and is free of "patch" or alternate sequences. In contrast, GRCH38 has 195 different contigs (with names such as chr14_GL000009v2_random and chrUn_GL000195v1) that are often excluded since it is challenging how to interpret results in these regions. The reduced contig set of T2T-CHM13 also facilitates interpretation and is directly compatible with essentially all of the most commonly used analysis tools (e.g., bwa, GATK, bedtools, samtools, bcftools, etc). To promote this transition, we provide variant calls and several other annotations for CHM13 within the UCSC Genome Browser and the NHGRI AnVIL as a resource for the human genomics and medical communities.

Finally, our work also underscores the need for additional T2T genomes. Most urgently, the CHM13 genome is karyotypically female and lacks a Y chromosome, so our analysis relied on the draft representation of chrY from GRCh38. We expect a T2T representation of chrY within the very near future, at which time we will study how including this sequence improves mapping and variant analysis of karyotypically male samples. Furthermore, many of the new regions in T2T-CHM13 are present in all human genomes and enable variant calling with traditional methods from short and/or long reads. However, many novel regions exhibit substantial variation within and between populations, including satellite DNA [34] and segmental duplications that are polymorphic in copy number and structure [35]. Relatedly, the expanded reference provides a basis for calling millions of new variants, but many of these variants are challenging to accurately resolve using current sequencing technology and analysis algorithms. Robust variant calling in these regions will require many, diverse haplotype-resolved T2T assemblies to construct a pangenome reference, in turn motivating the development of new methods for discovering, representing, comparing, and interpreting complex variation, as well as benchmarks to evaluate their performance [53,54].

Through our detailed assessment of variant calling at known and novel loci across global population samples, our study showcases CHM13-T2T as the preeminent reference for human genetics. The annotation resources provided herein will

help facilitate this transition, together expanding knowledge of human genetic diversity by revealing functional variation that was hidden from previous view.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921. doi:10.1038/35057062

2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431: 931–945. doi:10.1038/nature03001

3. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27: 849–864. doi:10.1101/gr.213611.116

4. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13: e1002195. doi:10.1371/journal.pbio.1002195

5. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526: 75–81. doi:10.1038/nature15394

6. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581: 434–443. doi:10.1038/s41586-020-2308-7

7. Seo J-S, Rhie A, Kim J, Lee S, Sohn M-H, Kim C-U, et al. De novo assembly and phasing of a Korean human genome. Nature. 2016;538: 243–247. doi:10.1038/nature20098

8. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020;38: 1044–1053. doi:10.1038/s41587-020-0503-6

9. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. 2021;49: D1046–D1057. doi:10.1093/nar/gkaa1070

10. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583: 699–710. doi:10.1038/s41586-020-2493-4

11. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369: 1318–1330. doi:10.1126/science.aaz1776

12. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature. 2021;590: 290–299. doi:10.1038/s41586-021-03205-y

13. Miga KH, Newton Y, Jain M, Altemose N, Willard HF, Kent WJ. Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. 2014;24: 697–707. doi:10.1101/gr.159624.113

14. Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin C-S, et al. Extending reference assembly models. Genome Biol. 2015;16: 13. doi:10.1186/s13059-015-0587-3

15. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10: 1784. doi:10.1038/s41467-018-08148-z

16. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517: 608–611. doi:10.1038/nature13907

17. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. Cell. 2019;176: 663–675.e19. doi:10.1016/j.cell.2018.12.019

18. Consortium T 1000 GP, The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015. pp. 68–74. doi:10.1038/nature15393

19. Yandell M, Huff C, Hu H, Singleton M, Moore B, Xing J, et al. A probabilistic disease-gene finder for personal genomes. Genome Res. 2011;21: 1529–1542. doi:10.1101/gr.123158.111

20. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315. doi:10.1038/ng.2892

21. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet. 2015;47: 276–283. doi:10.1038/ng.3196

22. Miller CA, Walker JR, Jensen TL, Hooper WF, Fulton RS, Painter JS, et al. Failure to detect mutations in U2AF1 due to changes in the GRCh38 reference sequence. bioRxiv. 2021. p. 2021.05.07.442430. doi:10.1101/2021.05.07.442430

23. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Towards a Comprehensive Variation Benchmark for Challenging Medically-Relevant Autosomal Genes. doi:10.1101/2021.06.07.444885

24. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328: 710–722. doi:10.1126/science.1188021

25. Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biol. 2019;20: 159. doi:10.1186/s13059-019-1774-4

26. E pluribus unum. Nature methods. 2010. p. 331. doi:10.1038/nmeth0510-331

27. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. bioRxiv. 2021. p. 2021.05.26.445798. doi:10.1101/2021.05.26.445798

28. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cold Spring Harbor Laboratory. 2021. p. 2021.02.06.430068. doi:10.1101/2021.02.06.430068

29. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. Nat Biotechnol. 2020;38: 1347–1355. doi:10.1038/s41587-020-0538-8

30. Steinberg KM, Schneider VA, Graves-Lindsay TA, Fulton RS, Agarwala R, Huddleston J, et al. Single haplotype assembly of the human genome from a hydatidiform mole. Genome Res. 2014;24: 2066–2076. doi:10.1101/gr.180893.114

31. Huddleston J, Ranade S, Malig M, Antonacci F, Chaisson M, Hon L, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 2014;24: 688–696. doi:10.1101/gr.168450.113

32. Dennis MY, Harshman L, Nelson BJ, Penn O, Cantsilieris S, Huddleston J, et al. The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol. 2017;1: 69. doi:10.1038/s41559-016-0069

33. O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak AC, et al. Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics. 2014;15: 387. doi:10.1186/1471-

2164-15-387

34. Nicolas Altemose, Glennis A. Logsdon, Andrey V. Bzikadze, Pragya Sidhwani, Sasha A. Langley, Gina V. Caldas. Complete genomic and epigenetic maps of human centromeres. bioRxiv. 2021.

35. Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, et al. Segmental duplications and their variation in a complete human genome. bioRxiv. 2021. p. 2021.05.26.445678. doi:10.1101/2021.05.26.445678

36. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538: 201–206. doi:10.1038/nature18964

37. Li J-P, Yang C-Y, Chuang H-C, Lan J-L, Chen D-Y, Chen Y-M, et al. The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. Nat Commun. 2014;5: 3618. doi:10.1038/ncomms4618

38. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018. p. 201178. doi:10.1101/201178

39. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515: 216–221. doi:10.1038/nature13908

40. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421. doi:10.1038/nature12477

41. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, et al. Weighted minimizer sampling improves long read mapping. Bioinformatics. 2020;36: i111–i118. doi:10.1093/bioinformatics/btaa435

42. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

43. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15: 461–468. doi:10.1038/s41592-018-0001-7

44. Kirsche M, Prabhu G, Sherman R, Ni B, Aganezov S, Schatz MC. Jasmine: Population-scale structural variant comparison and analysis. bioRxiv. 2021. p. 2021.05.27.445886. doi:10.1101/2021.05.27.445886

45. Shafin K, Pesout T, Chang P-C, Nattestad M, Kolesnikov A, Goel S, et al. Haplotype-aware variant calling enables high accuracy in nanopore long-reads using deep neural networks. bioRxiv. 2021. p. 2021.03.04.433952. doi:10.1101/2021.03.04.433952

46. Cheng JY, Racimo F, Nielsen R. Ohana: detecting selection in multiple populations by modelling ancestral admixture components. bioRxiv. 2019. p. 546408. doi:10.1101/546408

47. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat Methods. 2018;15: 595–597. doi:10.1038/s41592-018-0054-7

48. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17: 122. doi:10.1186/s13059-016-0974-4

49. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012;335: 823–828. doi:10.1126/science.1215040

50. Sung SS, Brassington A-ME, Krakowiak PA, Carey JC, Jorde LB, Bamshad M. Mutations in TNNT3 cause multiple congenital contractures: a second locus for distal arthrogryposis type 2B. Am J Hum Genet. 2003;73: 212–214.

doi:10.1086/376418

51. Sheng J-J, Jin J-P. TNNI1, TNNI2 and TNNI3: Evolution, regulation, and protein structure-function relationships. Gene. 2016;576: 385–394. doi:10.1016/j.gene.2015.10.052

52. Ryan DP, da Silva MRD, Soong TW, Fontaine B, Donaldson MR, Kung AWC, et al. Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. Cell. 2010;140: 88–98. doi:10.1016/j.cell.2009.12.024

53. Eizenga JM, Novak AM, Sibbesen JA, Heumos S, Ghaffaari A, Hickey G, et al. Pangenome Graphs. Annu Rev Genomics Hum Genet. 2020;21: 139–162. doi:10.1146/annurev-genom-120219-080406

54. Pritt J, Chen N-C, Langmead B. FORGe: prioritizing variants for graph genomes. Genome Biol. 2018;19: 220. doi:10.1186/s13059-018-1595-x

55. Kirsche M, Das A, Schatz MC. Sapling: accelerating suffix array queries with learned data models. Bioinformatics. 2021;37: 744–749. doi:10.1093/bioinformatics/btaa911

56. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81: 559–575. doi:10.1086/519795

57. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013;93: 278–288. doi:10.1016/j.ajhg.2013.06.020

58. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449: 851–861. doi:10.1038/nature06258

59. Langley SA, Miga KH, Karpen GH, Langley CH. Haplotypes spanning centromeric regions reveal persistence of large blocks of archaic DNA. Elife. 2019;8. doi:10.7554/eLife.42989

60. Chen L, Wolf AB, Fu W, Li L, Akey JM. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. Cell. 2020;180: 677–687.e16. doi:10.1016/j.cell.2020.01.012

61. Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science. 2017;358: 655–658. doi:10.1126/science.aao1887

62. Schatz MC, Philippakis AA, Afgan E, Banks E, Carey VJ, Carroll RJ, et al. Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL). bioRxiv. 2021. p. 2021.04.22.436044. doi:10.1101/2021.04.22.436044

63. Regier AA, Farjoun Y, Larson D, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. doi:10.1101/269316

64. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10. doi:10.1093/gigascience/giab008

65. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN]. 2013. Available: http://arxiv.org/abs/1303.3997

66. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. "O'Reilly Media, Inc."; 2020. Available: https://play.google.com/store/books/details?id=vsXaDwAAQBAJ

67. Cheng JY, Mailund T, Nielsen R. Fast admixture analysis and population tree estimation for SNP and NGS data. Bioinformatics. 2017;33: 2148–2155. doi:10.1093/bioinformatics/btx098

68. Ilardo MA, Moltke I, Korneliussen TS, Cheng J, Stern AJ, Racimo F, et al. Physiological and Genetic Adaptations to Diving in Sea Nomads. Cell. 2018;173: 569–580.e15. doi:10.1016/j.cell.2018.03.054

69. Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, et al. Local adaptation and archaic introgression shape global diversity at human structural variant loci. bioRxiv. 2021. p. 2021.01.26.428314. doi:10.1101/2021.01.26.428314

70. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13: 31. doi:10.1186/s13073-021-00835-9

71. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176: 535–548.e24. doi:10.1016/j.cell.2018.12.015

72. Parsons JD. Miropeats: graphical DNA sequence comparisons. Comput Appl Biosci. 1995;11: 615–619. Available: https://www.ncbi.nlm.nih.gov/pubmed/8808577

73. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297: 1003–1007. doi:10.1126/science.1072047

74. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001;11: 1005–1017. doi:10.1101/gr.187101

75. Numanagic I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. Fast characterization of segmental duplications in genome assemblies. Bioinformatics. 2018;34: i706–i714. doi:10.1093/bioinformatics/bty586

76. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18: 170–175. doi:10.1038/s41592-020-01056-5

77. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls in human genomes. Nat Biotechnol. 2019;37: 555–560. doi:10.1038/s41587-019-0054-x

Acknowledgements:
We would like to thank Steven Salzberg, Ben Langmead, Alexis Battle and several of their lab members for helpful discussions. 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.

Funding:
This work is partially supported by the National Institutes of Health (NIH) grant R35GM133747 to R.C.M, DP2OD025824 to M.Y.D., The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions:

Competing interests:

Data and materials availability:
Aligned reads, variant calls, and other summarizations are available within the NHGRI AnVIL Platform, along with Jupyter notebooks, R scripts, and WDL scripts for computing the analysis.

Supplementary Materials

Section I. Structural Comparisons of GRCh38 and CHM13

Mappability Analysis

We calculated minimum unique k-mer (MUK) lengths for both CHM13 and GRCh38. MUKs represent the minimum distance from a position in the genome needed to identify a unique sequence, either upstream (right-anchored) or downstream (left-anchored). To compute these values, all chromosomes of the target genome were concatenated, followed by their reverse complements. A suffix array (SA) and longest common prefix (LCP) table were calculated from this single sequence using the algorithm and script adapted from Sapling [55]. For each position in the SA, the LCP value plus one represents the MUK at that position. If the SA value indicated that the sequence was a reverse complement, the unique sequence was right-anchored; otherwise, the sequence was left-anchored. Any minimum unique sequences containing an N or overlapping a chromosome end were removed and those positions were marked as having no minimum unique sequence. All scripts for recapitulating this analysis can be found at https://github.com/msauria/T2T_MUK_Analysis.

Examining the tails of the MUK distribution, e.g. the longest perfect repeats in these genomes, we found that GRCh38 contained 70 repeated sequences larger than 5 kb while CHM13 had 81 such sequence pairs. Within GRCh38, more than half of these pairs (42) occurred between an unplaced contig and centromeric (or immediately adjacent), subtelomeric, or another unplaced contig's sequence, accounting for 29, 7, and 6 pairs, respectively. Another 18 sequence pairs in GRCh38 occurred between subtelomeric regions. Interestingly, 6 sequence pairs represented subtelomeric-centromeric sequences. There were two instances of sequences shared between centromeric regions, another between a subtelomeric region and the middle of chromosome 6, and a 12 kb sequence occurring on both chromosomes X and 8 in an intron of the gene *RBPMS*.

All but one sequence pair of long perfect repeats were removed in the CHM13 genome assembly as we were able to capture minute differences between these sequences using the long-read technology. Of the subtelomeric/centromeric mixed pairs, 3 were resolved to the subtelomeric side, 1 to the centromeric side, and 2 had no corresponding mapping in CHM13. The centromeric/other pair resolved to chromosome 6. Finally, the chromosome X/6 sequence pair resolved to chromosome 6. CHM13 contains 81 long sequence pairs, 77 of which occur between acrocentric centromere pairs, Of the remaining 4, 3 overlap a single subtelomeric pairing from GRCh38 and the last pair occurs in the middle of chromosomes 2 and 22, exactly encompassing an L1HS LINE element.

Generating variant calls with assembly-assembly alignment of CHM13 to GRCh38

We generated variant calls with dipcall (https://github.com/lh3/dipcall) [47] using the same CHM13-v1.0 assembly FASTA as haplotype 1 and haplotype 2 with GRCh38 as a reference, as in the NIST assembly benchmarking pipeline (https://github.com/usnistgov/giab-asm-benchmarking). Dipcall generates variant calls using any non-reference support in regions greater than or equal to 50 kb with contigs having mapping quality greater than or equal to 5. Dipcall also produces a BED which denotes confident regions covered by an alignment greater than or equal to 50 kb with contigs having mapping quality greater than or equal to 5 and with no other greater than 10 kb alignments.

Identification of missing gene copies in GRCh38

We simulated Illumina-like reads (400 million PE 150 bp reads) from T2T CHM13 reference v1.0 including the GRCh38 Y chromosome using Mason (https://github.com/seqan/seqan/tree/master/apps/mason2) and aligned them to GRCh38 (no alt or decoys contigs) and CHM13 v1.0 including the GRCh38 Y chromosome using BWA mem. Likewise, previously published CHM13 PacBio HiFi reads (~24X) (Ref.) were aligned to GRCh38 using Minimap2 with the -ax map-pb setting. We called SNVs in both datasets with GATK v4.1.8.1[38] using minimum MAPQ 30 and default parameters. Only PASS variants were used for downstream analyses. Heterozygous variants called by each platform were merged into one multi-sample VCF file with bcftools merge, and the number of heterozygous variants per kbp was calculated using bedtools coverage.

For both references, we first defined problematic regions as regions ≥2 kbp with ≥2 heterozygous calls in the CHM13 sample. From this, we connected regions separated by ≥5 kbp, and then filtered for regions ≥5 kbp in size.

Focusing on GRCh38-derived problematic regions, we intersected them with previously published RepeatMasker and SDs annotations obtained from UCSC Table Browser, as well as known GRC issues obtained from XXX. Additionally, we obtained variants flagged with excess of heterozygosity by the 1000 Genomes Project by filtering for InbreedingCoeff flag in the 1000 Genomes variants file (Ref). Empirical enrichment of variants with excessive heterozygosity within problematic regions was obtained by calculating the number of variants in 10,000 randomly sampled regions of the genome using bedtools shuffle. The empirical *p*-value was calculated as (M+1)/(N+1), where M is the number of iterations yielding a number of features greater than observed and N is the number of iterations.

We lifted problematic regions in GRCh38 to CHM13 using the following approach: (1) coordinates obtained by UCSC LiftOver with available liftover were lifted over to CHM13 usingain (Ref) if lifted coordinates size was within 80-120% of the original size, (2) Minimap2 longest hit if lifted coordinates size was within 80-120% of the original size, and closest hit when more than two options were available, and (3) manual selection and curation of remaining coordinates. To assess functional impact, these likely problematic regions were intersected with all gene features in Gencode v36, as well as a curated list of medically relevant genes [23].

Using available read-depth copy-number estimates in CHM13 reference based [35], we obtained the overall copy-number of the lifted regions as the median window copy-number for a "*k*-merized" version of GRCh38 and CHM13 references, as well as 268 individuals from the HGDP dataset (excluding sample LP6005442-DNA_A08). Regions here GRCh38 copy-number proved to be lower ≤ 1 individuals of the SGDP cohort, were considered likely true collapsed duplications in the GRCh38 reference. Additionally, we intersected lifted coordinates with CHM13 SDs [35] and centromere annotations [34].

Identifying false duplications in GRCh38

To identify falsely duplicated regions in GRCh38, we compared copy number estimates of GRCh38 to copy number estimates of 268 genomes from the HGDP dataset using short reads. We first averaged the copy number estimates for each genome across 1 kbp windows. For each 1 kbp region, we flagged it as a potential false duplication if the copy number in GRCh38 was greater than the copy number in 99% of the other genomes. Flagged regions were assigned a value of 1 and unflagged regions were assigned a value of 0. To filter the flagged regions, we used a median filter approach with a window size of 3 kbp, where the binary value of each 1 kbp region was replaced with the median value of the complete window. We then merged all adjacent flagged regions and reported the start and end coordinates with respect to CHM13. To find the corresponding locations of the duplications on GRCh38, we used minimap2 [42] version 2.17-r941 with parameter -p 0.25. Some regions mapped to more than two locations on GRCh38 due to true segmental

duplications in the genome. We curated these regions with more than two alignments, and identified the incorrect region as the region that did not have an assembly-assembly alignment from CHM13 or the HG002 haplotype. We identified the affected, correct region as the region that aligned most closely to the CHM13 region, which also had reduced HG002 read coverage. Upon curation of the regions with only two alignments on GRCh38, we selected as correct the region that was on the same chromosome arm as the corresponding CHM13 region. When both regions were on the same chromosome arm, we selected as correct the region that was not adjacent to or between gaps in GRCh38. One false duplication was a tandem duplication, and we arbitrarily selected one copy as correct. Upon curation, we also removed one small 8 kb region (chr19:14359000-14367000 on CHM13) that was incorrectly identified as falsely duplicated.

Quantifying artificial haplotype structure of GRCh38

We used PLINK [56] to extract all autosomal SNP pairs with perfect LD ($R^2 = 1$) in the entire 1000 Genomes sample set using recently published genotype data from the 30x sequencing by NYGC. By definition, GRCh38 carries the reference allele at all sites, so we computed the frequency of REF-REF haplotypes at each of these SNP pairs among the 1000 Genomes samples. We then used HTSlib to query CHM13 alleles at each corresponding site, again computing the frequency of the observed CHM13 haplotypes among the 1000 Genomes sample set. This analysis of CHM13 alleles on GRCh38 coordinates was based on the aforementioned variant calls produced by dipcall, restricting to portions of the alignment involving only a single contig (i.e., the region defined by the dip.bed output file) and excluding regions within 10 bp of an indel that distinguishes the two assemblies. All queried genomes were subjected to these same filtering criteria to ensure a common baseline for comparison. For the GRCh38 haplotypes that were absent (i.e. frequency = 0) among the 1000 Genomes sample set, we intersected the intervals defined by the SNP pairs with the clone tiling path used for the construction of the GRCh38 reference genome (https://hgdownload-test.gi.ucsc.edu/goldenPath/hg38/bigZips/hg38.agp.gz), thereby quantifying the number of rare allelic combinations explained by transitions between adjacent clones.

To compare these results to samples from the 1000 Genomes Project, we performed a leave-one-out analysis in which for each iteration, we randomly selected a single phased "haploid" genome from the 1000 Genomes sample set and re-computed LD for the full dataset minus that individual. For all SNP pairs in perfect LD ($R^2 = 1$) in the new sample (N = 2503 individuals), we queried the allelic state of the selected haploid sample in the manner previously described. We repeated this analysis for 100 randomly sampled individuals, generating superpopulation-specific distributions of LD-discordant SNP pairs per haploid genome. We again applied the same filtering criteria as were applied to CHM13 and GRCh38 to ensure a fair comparison.

Inference of local ancestry and Neanderthal introgression for GRCh38 and CHM13

We used RFMix v2.03-r0 [57] (https://github.com/slowkoni/rfmix) to infer the local ancestry of the CHM13 genome. As a set of reference samples for ancestry, we used the 2,504 individuals sequenced in the Phase 3 release of the 1000 Genomes Project [18]. Phased SNP genotypes for this reference set were generated by the 1000 Genomes Consortium through alignment to GRCh38 and calling variants against the GRCh38 reference, and included only biallelic SNVs (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/). We assumed that CHM13 carried the reference allele at sites in the 1000 Genomes dataset that were not called as CHM13 variants by dipcall (described above). All CHM13 genotypes were represented as phased homozygous diploid genotypes (either reference/reference or alt/alt), which resulted in two sets of identical inferred local ancestry segments for each haplotype. We obtained a genetic map for GRCh38, originally generated on build 35 of the human reference genome by

the HapMap Consortium [58] from the Beagle resource bundle (http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/).

We ran RFMix with default parameters (conditional random field spacing = 50 SNPs; random forest window size = 5 SNPs), grouping the 1000 Genomes reference panel into superpopulations (African, Ad Mixed American, East Asian, European, Southeast Asian). Because males are haploid for variants on the X chromosome, we included only female individuals in the reference panel for local ancestry inference for the X chromosome. We excluded from the results regions that overlapped with centromeres (UCSC Genome Browser; https://genome.ucsc.edu/cgi-bin/hgTables) or were marked as inaccessible for the 1000 Genomes Project (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38/). The non-pseudoautosomal regions (PARs) of the X chromosome were also inaccessible for this analysis due to the absence of variant calls outside the PARs.

Inspired by related approaches [24,59], we used an alternative strategy for local ancestry inference of GRCh38. By restricting analysis to individual clones, our approach avoids strong assumptions about the genetic map, which could produce misleading results given the artificial haplotype that we previously described. Specifically, for each clone reported in the tiling path, we extracted phased haplotypes of the 1000 Genomes samples from the corresponding genomic interval. We then computed the pairwise Hamming distances between each of these reference haplotypes and the reference GRCh38 haplotype (represented as an equal-length vector of zeros). We identified the nearest neighbor reference haplotypes to GRCh38 as those with the smallest Hamming distance, recording the superpopulation membership of each corresponding reference individual. Ancestry of the clone was then assigned as the majority superpopulation of the nearest neighbor haplotypes, normalizing by the relative abundances of superpopulations within the entire dataset. Clones for which no superpopulation achieved the majority among the nearest neighbors were designated as ambiguous, as were clones without SNPs in the 1000 Genomes sample (i.e., occurring within inaccessible regions of the genome).

For identification of Neanderthal-introgressed haplotypes, we used IBDmix (https://github.com/PrincetonUniversity/IBDmix), which detects introgressed archaic sequences in a set of individuals by identifying haplotypes that are identical-by-descent to a reference archaic sample [60]. Of the three Neanderthal individuals with available high-coverage sequences, we chose the Vindija Neanderthal as our reference archaic individual due to its inferred closest relation to the Neanderthal population that admixed with modern humans [61]. Because IBDmix identifies introgressed sequences in a set of modern samples, assumed to be from the same population, we additionally provided samples from the 1000 Genomes Project as a background set of modern individuals. These samples were sequenced to 30x coverage by the New York Genome Center, and variants were called on the GRCh38 reference (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/) [28].

Because local ancestry analysis indicated that the majority of the CHM13 genome is composed of European ancestries, we used European samples as the background set of modern individuals for CHM13. For GRCh38, which comprises haplotypes from multiple populations, we identified introgressed segments by pairing it with each of the 1000 Genomes superpopulations. In a given local ancestry tract, we only retained introgression calls from the background population matching the local ancestry in that region.

We required Neanderthal-introgressed haplotypes to have a LOD score of at least 4 and a length of at least 50 kb. As with the local ancestry segments, we removed introgressed segments within regions deemed inaccessible by the 1000 Genomes Project.

Section II. Improvements to Short Read Mapping and Variant Calling

Short read alignment, coverage analysis, and variant calling of the expanded 1000 Genomes collection

In order to perform large-scale alignment and variant calling, we used the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL) [62], which allowed us to analyze nearly 100 terabytes of raw sequence data and perform thousands of analyses in just a few days of wall-clock time. Using the AnVIL/Terra platform, we aligned the 3,202 samples from the 1000 Genomes Project [28] to the CHM13 reference genome using the functional equivalence pipeline standard established by the Centers for Common Disease Genomics project [63] as a framework. For CHM13 we used assembled v1.0 (20200921) main chromosome sequences with an addition of chrY, chrY_KI270740v1_random, and chrEBV from GRCh38, whereas the prior study of the 1000 Genomes analysis used GRCh38 including alt, decoy, random, unknowns, and HLA sequences. For the analysis, we followed the processing of the previous pipeline, although we updated certain elements of the pipeline in order to improve efficiency while maintaining accuracy. Most notably, we substituted samtools v1.11 [64] for Picard when fixing mate information in the BAM file, merging lane-level BAMs, marking duplicates, and coordinate-sorting. Otherwise, we followed the CCDG pipeline, including all pertinent parameters: lane-level alignment of the 1KGP FASTQ files to CHM13 using BWA-MEM v0.7.17-r1188 [65], fixing mate information in the BAM, coordinate-sorting the BAMs, merging lane-level BAM files to sample-level BAM files, marking duplicates, and converting to CRAM format. Though the original pipeline performed base quality score recalibration at the BAM level, we forewent this step because the subsequent variant inference part of the pipeline utilizes a newer version of GATK framework (4.1.9 for AnVIL vs 3.5.0 for NYGC)

After alignment, we quantified the mapping statistics using samtools stats. In addition to the improvements reported in Figure 2a, several other metrics improved such as a decrease in the standard deviation of the insert size, and an improvement in the paired-end concordance. Lastly, because of the resolved/introduced repeat sequences in CHM13 we observe an increase in the number of reads with a mapping quality value of 0, signalling a repetitive alignment)

We then computed coverage statistics using mosdepth on 500bp non-overlapping genome-wide windows on individual sample bam files. For coverage analysis, we only considered autosomal chromosomes for coverage stability analysis, and removed from consideration 500bp windows that have 250+ Ns in the respective references (GRCh38: 260,251 windows, CHM13: 22,950 windows). We then consider the interquartile range to annotate 500bp windows with outlier/abnormal coverage using a threshold of 1.5x IQR. We then intersected these regions against several annotations to evaluate coverage uniformity. This included intervals defined by GIAB [HC_smallvar, HC_SV, gene_medical], RepeatMasker [r_LINE, r_SINE, r_low_complexity, r_satellite, r_simple], CAT [genic], SD [synteny, synteny_not], superpopulation ancestry [a_AFR, a_AMR, a_EAS, a_EUR, a_SAS, a_neanderthal] we computed overlap of every stratification context with 500bp windows for which the coverage was computed and retained context-specific windows that had 1+/500 bp context overlaps.

To call small variants, we again followed the framework presented by the 1KGP, updating to GATK v4.1.9 [66]. We performed raw variant calls using GATK HaplotypeCaller on a per-chromosome basis for each sample. We then generated GATK GenomicsDB files on each chromosome, across all 3,202 samples, per interval, with 1kb of padding on either side of each interval, performed joint genotyping using GATK GenotypeVCFs, and trimmed the padding on the resulting VCFs to produce interval-level joint VCF files across all 3,202 samples [38]. We then used bcftools v1.9 [64] to concatenate the interval-level joint VCF files to produce chromosome-level VCF files and to obtain statistics on all VCF files. [TODO: more info about recalibration, following the PDF]

To identify non-spurious variants, we first filtered these VCF files by the filter field, using only variants labeled as PASS by GATK. This eliminated a number of variants in highly repetitive regions, such as in satellite regions (chr9 figure).

For most subsequent analysis, we only considered the 2,598 founder (unrelated) samples in the 1000 Genomes Project. Using bcftools v1.9 for most of our analyses, we calculated the number of variants per sample, performed allele frequency analysis on autosomes, and performed local ancestry analysis on the population-level VCF files. To identify variants with allele frequency of 1.0, we used bcftools v1.12 norm to stratify multiallelic variants and bcftools view.

We also identified Mendelian discordance in all 602 trios using the MendelianViolationEvaluator evaluation module of GATK VariantEval.

Comparison of GRCh38/NY Genome Center Pipeline to CHM13/AnVIL Pipeline

The main differences between the original NYGC pipeline for short-read data processing and AnVIL setup with CHM13 reference lie in: (i) tool/tool versions and (ii) the starting organization of the sample-specific read data. In contrast to the NYGC pipeline which started with sequenced flow-cell- and lane-specific fastq files straight from the sequencing instruments, the AnVIL setup started from a composite fastq file downloaded for the European Nucleotide Archive. These files were then decomposed into individual flow-cell- and lane-specific fastq files. For subsequent alignment steps AnVIL setup mimics the NYGC pipeline (using a newer bwa v0.7.17 vs v0.7.15 in NYGC pipeline), but substitute steps that are performed with Picard Tools in the NYGC pipeline (e.g., sorting, merging, deduplication) with the samtools' analogs. AnVIL setup also forgoes alignment recalibration step, as the subsequent variant inference part of the pipeline utilizes a newer version of GATK framework (4.1.9 for AnVIL vs 3.5.0 for NYGC), which utilizes elements of alignment recalibration workflow directly during the variant inference. AnVIL setup utilized segmentation (1Mbp) of the CHM13 reference with reciprocal overlapping (extra "extending" margins of 10k) between consecutive segments for HaplotypeCaller step, followed by joint genotyping step via GenomicsDB setup. Each genotyped segment-based variant file is then trimmed off of the extending 10Kbp extending margins on each segment, and AnVIL pipeline then concatenates segments in a sequential manner into chromosomes, and then full genomes. Variant recalibration steps in AnVIL setup follows the outline of the NYGC pipeline with the resource files lifted-over from GRCh38 reference onto the CHM13, with only unambiguously lifted entries considered.

To evaluate how the alignment part of the pipeline differs between the AnVIL setup and the original NYGC pipeline we have performed GRCh38-based alignment on AnVIL setup for 3 samples. We then computed samtools alignment stats on both the original NYGC cram files and on the AnVIL-generated ones. We observed that the produced alignments are almost identical across alignment setups, with most differences being very minor and observed in the number of MAPQ0 alignments (<0.2% reduction), number of reads marked "duplicated" (<0.01% reduction), insert size standard deviation (<1.2% reduction), pairs on different chromosomes (<1.2% reduction), etc.

We further compared the AnVIL small variant inference with the original NYGC setup and observed genome-wide variant inference concordance of 98.7% (via bcftools stats). We further confirmed that the difference is driven by the GATK v3.5 in NYGC pipeline being updated to GATK 4.1.9 in AnVIL setup: identical versions of GATK ran on alignments produced with NYGC vs AnVIL pipeline yielded >99.9% genome-wide concordance. We also observed that the proposed segmentation-with-margins GATK variant inferences approach utilized in the AnVIL setup yielded absolutely identical results when compared to complete-chromosome based segmentation on 200+ samples with CHM13v1.0 reference.

Analysis of visibility of single- and multi-sample SNVs when using different references

Analyses of the visibility of SNVs for single and multiple samples were performed on the GATK short read VCF files for this project. For single sample results, the calls from HG002 were used, and for the "1KG" calls, the full set of variants for 3,202 samples were used. To count the number of HG002 SNVs called using GRCh38 as a reference which are not callable using the T2T-CHM13 reference, the VCF file of HG002 GRCh38 SNVs was lifted to T2T-CHM13 using Picard v2.25.0's LiftOverVCF tool (Broad Institute, http://broadinstitute.github.io/picard/) with the -RECOVER_SWAPPED_REF_ALT option. Any SNV whose genotype with respect to T2T-CHM13 was homozygous reference, or "0/0", and whose VCF record contained the Picard-assigned "SwappedAlleles" INFO tag, was counted among the 946,126 SNVs that are not callable against T2T-CHM13 due to the lack of a variant allele in HG002. Likewise, to calculate the number of HG002 SNVs called using T2T-CHM13 that lift over to GRCh38 positions for which the reference allele is not represented in HG002, Picard LiftOverVCF was again run on the T2T-CHM13 VCF with the -RECOVER_SWAPPED_REF_ALT option. SNVs for which the lifted SNV's GRCh38 genotype was homozygous reference, or "0/0", and which were assigned the INFO field value "SwappedAlleles", were counted among the 672,341 SNVs that are not callable in GRCh38 due to the lack of a variant allele in HG002. To compute similar metrics for the 1KG SNVs, the VCF files of T2T-CHM13 variant calls without genotypes were lifted from T2T-CHM13 to GRCh38, and the GRCh38 variants were similarly lifted from from GRCh38 to T2T-CHM13. Instead of examining genotypes, the lifted allele frequencies (the "AF" values in the VCF INFO field) and the unadjusted alternate and total allele counts ("AC" and "AN") were examined using a custom perl script to count variants that were 100% fixed for one or the other reference allele.

To compare variants between the called and the lifted sets of variants for a given reference, hap.py version 0.3.9 (https://github.com/Illumina/hap.py) was run on the complete sets of "PASS" variants from both references, as well as VCF files which had been filtered to remove variants that are visible only with respect to the T2T-CHM13 reference, only with respect to the GRCh38 reference, or visible with respect to both.

Section III. Improvements to Long Read Alignment and SV Calling in CHM13

Long read alignment and coverage analysis

To assess the impact of using CHM13 as a reference on long-read alignment and variant calling, we aligned data from two independent sequencing platforms - PacBio High Fidelity (HiFi) Circular Consensus Sequencing and Oxford Nanopore (ONT) - to both GRCh38 and CHM13. For the GRCh38 reference, we excluded alternate and decoy sequences because current long-read mappers are not alt-aware, but included randoms and unknowns. For the CHM13 reference, we used the v1 assembly with the addition of chrY, chrY_KI270740v1_random, and chrEBV from the GRCh38 reference.

We used HiFi sequencing data from 17 samples and ONT data from a subset of 14 of those samples (**Fig. 3a**). The average read length of HiFi data across the 17 samples was 18,130.2 bp, and the average read length of ONT data across the 14 samples for which it was available was 21,912.9 bp (**Fig S3.1**). Alignments were performed with minimap2 [42] and Winnowmap [41], and mapping statistics were compared between the two references using samtools stats (**Fig. S3.2**). The number of reads mapped was similar between the two references across all combinations of technology and aligner. However, compared to GRCh38, we observed a lower mapping error rate in CHM13 due to the more accurate and complete sequence. At the same time, alignments to CHM13 yielded more reads with a mapping quality of zero due to the increased number of resolved repeats which resulted in multiple identical targets for mapping. Finally, there were fewer non-primary alignments to CHM13 because the newly resolved sequence enables better end-to-end mapping of long reads in primary alignments.

In addition, we investigated the effect of CHM13 on coverage anomalies by using mosdepth to compute coverage statistics across non-overlapping 500bp bins in each reference. Only autosomal chromosomes were considered, and bins with 250 or more N's were removed (260,251 bins in GRCh38 and 22,950 bins in CHM13).

We first computed the standard deviation of the per-bin coverage in each combination of reference, technology, aligner, and sample and compared the results between references. We observed a similar mean coverage in CHM13 and GRCh38, but found a significant reduction in the standard deviation when aligning to CHM13 (**Fig. S3.3**), indicating alignments which are distributed more uniformly across the genome.

In addition to these genome-wide results, we compared the mean and standard deviation among bins overlapping (≥1 bp) with a number of genomic contexts (**Fig. S3.3**):

- Satellite Repeats
- Genes
- Non-syntenic regions with respect to the other reference
- Syntenic regions with respect to the other reference
- Abnormal coverage bins, or bins in which the coverage is outside the range [Median - 1.5 * (Median - Q1), Median + 1.5 * (Q3 - Median)] among all bins in the same reference with the same technology, aligner, and sample.

We found that the coverage standard deviation in GRCh38 was particularly elevated relative to CHM13 in satellite repeats, non-syntenic regions, and regions of abnormal coverage, likely due to repetitive sequences which are not properly characterized in GRCh38.

Finally, for those bins with abnormal coverage, we counted the number of distinct samples in which it had abnormal coverage (**Fig. S3.4**). We found that many abnormal coverage regions in CHM13 had such coverage in only a single sample, which could be caused by rare structural variation which disrupts alignments to certain bins. On the other hand, GRCh38 had a higher number of bins which had abnormal coverage in every sample, indicating incorrectly resolved repeats or other errors in the reference.

Long read variant calling analysis

To evaluate the impact of CHM13 on structural variant (SV) calling from long reads, we called SVs in all long-read samples from the Winnowmap and minimap2 alignments described above. Variants were called separately in each unique combination of (sample, reference, technology, aligner) using Sniffles [43] with sensitive parameters - a minimum variant length of 20 bp and a minimum read support of two reads. Variants were marked as high-confidence if they were at least 30bp in length, were supported by min(10, 25% of average coverage) reads, and were annotated with the PRECISE INFO field by Sniffles. Then, insertion calls were refined using Iris and duplicate variant calls within the same callset were removed using Jasmine [44] with a constant breakpoint distance threshold of 200 bp. The alignments and variant calls were computed using the default recommended parameters from our optimized Jasmine-SV pipeline: https://github.com/mkirsche/Jasmine/tree/master/pipeline.

To compare SV calling in a large cohort setting between the two references, we constructed cohort-level SV calls with Jasmine from HiFi reads separately for each reference. For each reference, we merged each sample's SV calls derived from Winnowmap and minimap2 alignments, and only retained calls which were detected from both sets of alignments. Then, we merged the per-sample callsets to obtain a unified cohort-level SV callset. SVs which were not annotated as high-confidence in at least one sample in which they were present were discarded, resulting in final callsets of 124,566 SVs in GRCh38 and 141,193 SVs in CHM13. The cohort-level callset was computing using Jasmine with the default recommended parameters.

Using these cohort-level callsets, we computed the allele frequency distribution in each reference (**Fig. 3c**) as well as the number of SVs present in each sample (**Fig. S3.5 and S3.6**). We also counted the number of insertions and deletions in

each reference to compute indel balance (**Fig. 3d**). To show the effects of SV calls in novel regions on the balance of insertions and deletions, we annotated SV calls that were present in non-syntenic regions of CHM13 with respect to GRCh38. We also looked specifically at the distributions of SV calls in CHM13 which intersected (1+ bp) a number of different genomic contexts, including genes and exons, syntenic and non-syntenic regions (**Fig. S3.7**), as well as centromeres and different classes of repeats (**Fig. S3.8**).

To evaluate the ability to detect *de novo* SVs in CHM13 in trio settings, we constructed trio callsets for two different trios from the Genome in a Bottle Consortium - the HG002 trio of Ashkenazim ancestry and the HG005 trio of Han Chinese ancestry. To construct a callset for each trio, we merged the callsets of the child and both parents derived from all four combinations of aligner (Winnowmap and minimap2) and sequencing technology (HiFi and ONT). Then, we discarded any variants which were not supported by both technologies with Winnowmap in at least one of the samples in which they were present. Similar to the population-level analysis above, we also only retained SVs which were annotated as high-confidence in at least one sample. This yielded final trio callsets of 56,414 variants in the HG002 trio and 56,449 variants in the HG005 trio. We manually inspected all SVs in these sets which were present in only the child of the trio: 36 in HG002 and 45 in HG005. Similar analysis in GRCh38 yielded 40 candidates in HG002 and 29 candidates in HG005. In both references, these candidate sets included a previously unreported potential *de novo* variant in HG005, a 1,571 bp deletion at chr17:49,401,990 in CHM13 (**Fig. S3.9**).

Optical mapping assembly and variant calling

Bionano optical mapping data for HG002/GM24385 was generated using the DLS chemistry and Bionano Saphyr system, available at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/Bionano_haplotype_SV_DLS_06172019/. A haplotype-aware optical assembly of Bionano data for HG002 was generated using default parameters with Bionano Solve3.6. The assembly was aligned against GRCh38 and against the T2T-CHM13 v1.0 reference, and SV calling was performed using default parameters with Bionano Solve3.6. For mapping, the T2T-CHM13 reference was *in silico* digested from t2t-chm13.20200921.withGRCh38chrY.chrEBV.chrYKI270740v1r.fasta.

There were 2865 non-redundant variants called against T2T-CHM13, and the number of insertions and deletions called against T2T-CHM13 are more balanced than that against GRCh38, consistent with the observation that T2T-CHM13 corrects collapses in GRCh38 (**Fig. S3.14**). There are 1431 insertions called against the T2T-CHM13 reference compared to the 2771 insertions called against GRCh38.

Among the variants called against the T2T-CHM13 reference, 306 of them overlap non-syntenic regions, which are novel or improved regions compared to GRCh38. Some of these SVs involve GRCh38 gaps that are now fixed in the T2T-CHM13 reference (**Fig. S3.15**). Although the SVs crossing gaps in GRCh38 can still be called with respect to GRCh38, T2T-CHM13 improves resolution for these SVs by adding markers that can be aligned between the optical assembly and the reference.

Further curation of telomeric region also finds that the T2T-CHM13 reference is more comprehensive than GRCh38 at many telomeres, at least in relation to the HG002 optical assembly, though HG002 is closer to GRCh38 for some telomeres. Future work could include evaluating the utility of calling SVs from Bionano data in many individuals against T2T-CHM13 to assess improvements across ancestry groups.

Section IV. Characterization of non-syntenic regions

Newly added region analysis

We intersected the 1000 Genomes population-wide joint genotyped VCF files with the non-syntenic <CITE> and novel <CITE> regions of the CHM13 assembly, which we used to calculate the number of bases newly added in this assembly and the number of variants in these newly accessible regions. Intersected with gene regions

TODO the long-read analysis

Identifying signatures of selection in novel CHM13 regions

We applied a frequency differentiation-based method for identifying signatures of selection within novel regions of the CHM13 assembly. These novel regions were defined as regions of no synteny between the CHM13 and GRCh38 assemblies, as identified by <synteny method>. Using variant genotypes generated with the 1000 Genomes dataset, we applied Ohana [46,67,68], a maximum likelihood-based method that models individuals as possessing ancestry from combinations of *k* ancestry components. Ohana then tests whether individual variants adhere to this genome-wide null model, or are better explained by an alternative model in which frequencies are allowed to vary in one or more populations. Because most demographic events will affect the genome in its entirety, outlier loci exhibiting extreme levels of frequency differentiation compared to the genome-wide average serve as candidate targets of local adaptation.

We used the inferred admixture relationships generated by [69] for the core 2,504 samples from the 1000 Genomes Project, modeling these individuals' genomes as combinations of 8 ancestry components. This choice of *k* follows the precedent of the 1000 Genomes Project [18] and replicates known patterns of population structure at continental scales, as well as signatures of admixture within the Ad Mixed American and African American populations [69] . In order to generate "selection hypothesis" matrices to search for selection in a specific ancestry component, we modified the neutral covariance matrix by allowing one component at a time to have a greater covariance. A scalar value of 10, representing the furthest possible deviation that a variant could have in a population under selection, was added to elements of the neutral covariance matrix depending on the population of interest.

For each variant of interest, Ohana computes the likelihood of the observed ancestry component-specific allele frequencies under the selection and neutral models, then compares them by computing a likelihood ratio statistic (LRS), which quantifies relative support for the selection hypothesis. We filtered these results to remove extreme outliers in null model log-likelihoods (global log likelihood (LLE) < -1000), which were unremarkable in their patterns of allele frequency and instead indicated a failure of the neutral model to converge for a small subset of variants. We also removed variants that could not be genotyped in more than 20% of samples, as well as variants that were extremely rare in all 1000 Genomes populations (minor allele frequency in all populations < 0.05 or > 0.95).

Section V. CHM13 Clinical Genomics

Annotation of CHM13 variants using Variant Effect Predictor (VEP)

VEP [48] (version 102.0) was used to annotate variants generated by dipcall when aligning chm13_v1.0_plus38Y.fa to hg38.no_alt.fa. Briefly, vcf files were annotated without the --filter_common and --canonical flags. CADD [70] v1.6 and raw SpliceAI [71] scores were added using both the CADD and SpliceAI plugins. Variants were filtered based on predicted HIGH functional impact.

Identification of medically-relevant genes with impacted variant discovery

Reference artefacts and errors were intersected with a previously curated list of medically-relevant genes [23]. GRCh38 coordinates for genes were lifted over using Picard liftover tool to identify locations in T2T-CHM13. For putative errors or rare SVs, we intersected gene coordinates after expanding breakpoints for each variant ± 5 kbp. Genomic sequencing containing *TNNT3* implicated with SVs in GRCh38 were extracted from each reference (chr11:1,892,362-1,946,566, GRCh38; chr11:1,978,257-2,034,355, CHM13v1.0) and homologous regions compared using miropeats -s 400 [72]. SDs containing *KCNJ18* were identified in GRCh38 (chr17:21,687,227-21,736,311; CDS: chr17:21,702,787-21,704,088) and T2T-CHM13v1.0 (chr17:21,636,382-21,685,461; CDS: chr17:21,651,942-21,653,243) and paralogs pinpointed using the UCSC Genome Browser SD annotations [73–75]. With *KCNJ18*, Genome coordinates of SDs containing *KCNJ12* (GRCh38: chr17:21,399,778-21,450,415 (whole) and chr17:21,415,343-21,416,644 (CDS); T2T-CHM13v1.0: chr17:21,348,977-21,399,639 (whole) and chr17:21,364,558-21,365,859 (CDS)) and the novel *KCNJ17* (T2T-CHM13v1.0: chr17:22,634,421-22,683,415 (whole) and chr17:22,666,582-22,667,880 (CDS)) were used to extract variants from both references intersecting each locus. Minor-allele frequencies were plotted and distributions compared using a Mann-Whitney U test (R-command etc). Finally, direct comparison of allele frequencies for variants falling within the CDS were compared using 1000 Genomes datasets from the NYGC (GRCh38), produced here (described above for T2T-CHM13), and the liftover of T2T-CHM13 to GRCh38 (described above).

Creating a difficult medically-relevant gene benchmark for HG002 on CHM13

The Genome in a Bottle Consortium recently created a highly-curated benchmark for 273 challenging and medically-relevant genes with GRCh37 and GRCh38 as references [23]. This benchmark used variant calls generated by dipcall [47] when aligning the paternal and maternal haplotypes assembled using hifiasm with PacBio HiFi reads for HG002 and Illumina k-mers from the parents, HG003 and HG004 [76]. These 273 genes tend to be challenging to call variants in HG002 due to structural variants, segmental duplications, and/or reference errors. The final benchmark regions for these genes excluded a small number of errors in homopolymers longer than 20 bp and errors found during extensive manual curation.

For this manuscript, we create a comparable benchmark for HG002 on the CHM13 reference with the following steps. (1) We find the gene coordinates on CHM13 for the 273 genes using the T2T-CHM13 v3 gene annotations. (2) We generate variant calls with dipcall (https://github.com/lh3/dipcall) using the same HG002 trio-hifiasm assembly with CHM13 as a reference, as is done in the NIST assembly benchmarking pipeline (https://github.com/usnistgov/giab-asm-benchmarking). Dipcall generates variant calls using any non-reference support in regions that are greater than or equal to 50 kb with contigs having mapping quality greater than or equal to 5. Dipcall also produces a BED which denotes confident regions that are covered by an alignment greater than or equal to 50 kb with contigs having mapping quality greater than or equal to 5 and with no other greater than 10 kb alignments. (3) We identify that dipcall/hifiasm fully resolves 269/273 genes on CHM13 (i.e., the gene region and any segmental duplications that overlap the gene region are encompassed by the dip.bed file output by dipcall, so that one and only one contig from each haplotype fully covers the region). We exclude 4 genes (ESPN, IFITM3, FCGR3A, and FCGR2B) due to large structural variation in CHM13 relative to GRCh38 and HG002. (4) We liftover the HG002 medically relevant genes v1.0 small variant benchmark regions from GRCh38 to CHM13 and intersect with T2T gene coordinates. (5) We manually exclude new SVs >49bp in HG002 on CHM13 plus any overlapping tandem repeats. (6) We liftover CHM13 benchmark regions from CHM13 to GRCh38 and intersect with GRCh38 benchmark regions to get equivalent benchmarks on both references. The resulting benchmark regions are <1% different in size (11548799 bp and 11482860 bp on GRCh38 and CHM13, respectively), enabling benchmarking performance metrics to be directly compared between the references.

To compare variant call accuracy when using CHM13 vs. GRCh38 as a reference, we follow best practices for benchmarking from the Global Alliance for Genomics and Health, using hap.py with vcfeval to generate performance metrics [77]. We use 3 callsets for HG002 generated similarly for this work on both GRCh38 and CHM13: Illumina-

bwamem-GATK, HiFi-PEPPER-DeepVariant, and ONT-PEPPER-DeepVariant. For Illumina-bwamem-GATK, we use single-sample variant calls before filtering, because HG002 was not included in the 1000 Genomes Project population analysis for GRCh38.

<<Supplementary figures and tables as separate attachments with captions embedded within>>