A T2T-CHM13 recombination map and globally diverse haplotype reference panel improves phasing and imputation

Joseph L. Lalli,[1] Andrew N. Bortvin,[2] Rajiv C. McCoy,[2,3,*] and Donna M. Werling[1,3,*]
[1]Laboratory of Genetics, University of Wisconsin-Madison, Madison, WI, United States
[2]Department of Biology, Johns Hopkins University, Baltimore, MD, United States
[3]These authors jointly supervised this work.
*Correspondence: dwerling@wisc.edu
*Correspondence: rajiv.mccoy@jhu.edu

Summary

The T2T-CHM13 complete human reference genome contains ~200 Mb of newly resolved sequence, improving read mapping and variant calling compared to GRCh38. However, the benefits of using complete reference genomes in other contexts are unclear. Here, we present a reference T2T-CHM13 recombination map and phased haplotype panel derived from 3202 samples from the 1000 Genomes Project (1KGP). Using published long-read based assemblies as a reference-neutral ground truth, we compared our T2T-CHM13 1KGP panel to the previously released GRCh38 1KGP phased callset. We find that alignment to T2T-CHM13 resulted in 38% fewer assembly-discordant genotypes and 16% fewer switch errors. The largest gains in panel accuracy are observed on chromosome X and in the regions flanking disease-causing CNVs. Simons Genome Diversity Project samples were more accurately imputed when using the T2T-CHM13 panel. Our study demonstrates that use of a T2T-native phased haplotype panel improves statistical phasing and imputation for samples from diverse human populations.

Keywords

1000 Genomes Project; population genetics; haplotype phasing; genotype imputation; reference imputation panel; CNVs; reference genomes; T2T-CHM13; GRCh38

Introduction

The T2T-CHM13 reference genome corrects thousands of errors in previous reference genome assemblies and resolves approximately 200 Mb of human genetic sequence[1]. When compared to the GRCh38 reference genome, use of this telomere-to-telomere (T2T) genome as a reference has been shown to reduce short-read alignment errors and improve variant discovery and genotyping[2–4]. Nevertheless, several practical barriers hinder the widespread adoption of T2T-CHM13 for downstream applications.

For example, statistical phasing and imputation techniques require the use of reference population haplotypes and a genome-wide map of recombination rates. These techniques are critical for fields of genetics as diverse as genome-wide association studies (GWAS)[5–7], clinical genomics[8], admixture analyses[9], allele-specific studies of gene regulation[10–14]. Clinically, phasing genetic variants enables the investigation of compound heterozygosity effects[15] and the identification of disease-causing haplotypes[16]. However, to our knowledge, there is no publicly available panel of T2T-CHM13 haplotypes, and recombination maps have not been generated from T2T-CHM13-aligned data. While GRCh38 resources can be “lifted over” to T2T-CHM13 coordinates[17], these methods are frequently unreliable in highly variable regions of the genome, do not benefit from improved read mapping or variant calling, and preclude the investigation of newly resolved regions[18].

In this study, we present both population-specific and globally-averaged T2T-CHM13-native recombination maps, along with a phased reference haplotype panel from 3202 individuals. Our work seeks to take advantage of the improvements to alignment and genotyping observed with the T2T-CHM13 reference to more accurately infer genome-wide recombination rates and generate chromosome-resolved haplotypes

representing human genetic diversity. We build off the previously released T2T-CHM13 1KGP variant callset, which was produced by aligning Illumina short reads from the phase 3 high coverage release of 1KGP data[19] to the CHM13v2.0 reference genome using the functionally equivalent GATK pipeline[20,21]. This pipeline is widely used in both clinical and research contexts and is designed to produce variant callsets that are comparable between research groups.

Evaluation of genotyping, phasing, and imputation quality all have one challenge in common: a shortage of ground truth datasets to serve as a common measuring stick. In the past few years, near-reference quality full length assemblies of 100 1KGP samples have been released by the Human Pangenome Reference Consortium (HPRC)[22] and the Human Genome Structural Variant Consortium (HGSVC)[23]. Importantly for our work, aligned genome assemblies do not need to be lifted over; genetic variation in shared genomic regions (syntenic regions) can be accurately described in either GRCh38 or T2T-CHM13 coordinates. These assemblies therefore serve as a reference-neutral source of ground truth when assessing genotyping and phasing accuracy.

Using these assemblies, we evaluate the accuracy of both genotyping and phasing in our panel compared to the 1KGP consortium's GRCh38 1KGP reference haplotype panel. We also evaluate the accuracy of reserved HPRC samples and Human Genome Diversity Project (HGDP) samples that have been phased and imputed using either the 1KGP GRCh38 and 1KGP T2T-CHM13 reference haplotype panels, highlighting genomic regions where the greatest improvements are achieved. To facilitate the use of a T2T-CHM13 recombination map and 1KGP T2T-CHM13 panel in downstream applications, we have made both resources openly available for public access.

Results

T2T-CHM13 recombination rate maps are broadly consistent with previous GRCh38 maps

Our panel builds upon data from Rhie et al.[24], which used the functionally equivalent GATK pipeline to discover small variation (SNPs and indels) and genotype 1KGP samples with respect to T2T-CHM13. To reduce false positive variant calls that may interfere with haplotype inference, we applied strict variant quality controls, including filtering out variants with a variant quality score log-odds (VQSLOD) less than zero (see Methods). We chose to retain singleton variants, as recent advances in computational phasing now allow for singleton phasing[25]. Our VQSLOD filter resulted in the exclusion of an additional 14,900,640 variants compared to the NYGC's GRCh38 1KGP variant filtering methods, but the inclusion of singleton alleles added 40,097,103 unique variants to our dataset (Supplemental Figure 1).

We then used this set of high-confidence variant calls to generate linkage disequilibrium (LD)-based recombination maps for T2T-CHM13 autosomes and the X chromosome using the software package pyrho[26,27] (Figure 1). LD-based inference of recombination can be sensitive to demographic history, and the underlying recombination landscape can vary between populations. For this reason, we first inferred recombination maps separately for the 26 subpopulations defined by the 1KGP. A combined global map was then created by averaging all population-specific maps, weighting by respective sample sizes (see Methods). We therefore note that this "global" average is a reflection of the population definitions and sample sizes of the original input data.

We calculated an unadjusted cumulative map length of 2187 cM, corresponding to a genome-wide average recombination rate of 0.72 cM/Mbp. This is notably lower than either the pedigree-based 3615 cM estimate produced by the deCODE consortium or the 2959 cM estimate obtained using cytologic observations of meiotic recombination[28,29]. Most phasing and imputation tools' default settings are tuned using the GRCh38 HapMap2 recombination map. Accordingly, we scaled our recombination maps to the per-chromosome length of the HapMap2 and deCODE recombination maps, consistent with the approach taken in the original pyrho study[27].

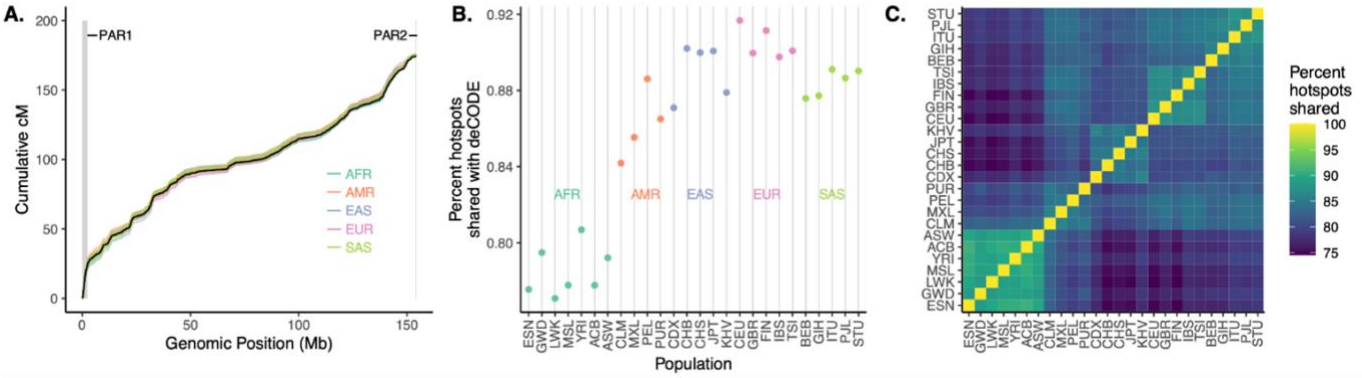

Figure 1. A CHM13v2.0 reference recombination map. a) Genetic and physical distance across chromosome X. Genetic distance is represented in cumulative centimorgans. Vertical dashed lines denote boundaries of pseudoautosomal regions. Colored lines represent individual populations; the black line represents map averaged across populations. b) Proportion of hotspots identified in our study that overlap hotspots in deCODE. Data is stratified by population. c) Proportion of hotspots shared between populations.

To investigate variation in recombination rates, we measured the Spearman correlation between recombination maps for all populations at multiple resolutions (Figure 1c). Hotspot locations, defined as regions with recombination rates over ten times the genome-wide average, were similar across all populations, but this similarity was highest between human populations from the same continental group (Figure 1a). We also compared these recombination maps to previously published LD-based maps generated using pyrho from 1KGP variant calls aligned to GRCh38. We find that Spearman correlation between recombination maps are high across all populations (91.4-97.6%).

We also compared our results to recombination maps generated by deCODE based on detection of crossovers in Icelandic trios[30] (Figure 1b). Given the high correlation in recombination maps between populations[27], we expect the deCODE pedigree-based and 1KGP LD-based recombination maps to be similar, especially for populations with more recent divergence. Consistent with this hypothesis, we observe that for all populations, 74-92% of hotspots detected in our 1KGP LD-based maps are also present in the deCODE map. When we restricted our analysis to hotspots that are shared between all 26 subpopulations, we found that 95% were also identified by deCODE.

While whole-genome studies of diverse cohorts are likely best served by averaged maps, which also benefit from the larger sample size, locus-specific analyses may benefit from the use of population-specific recombination maps. This may be particularly relevant for regions of the genome where features of the recombination landscape and haplotype structure differ between populations. To this end, we quantified variation in recombination rates between populations at various resolutions and report the genomic regions with highest variance (Supplemental Data S1).

A panel of 6404 statistically phased T2T-CHM13 haplotypes improves phasing and imputation in all regions of the genome

Pangenomic reference assemblies provide new source of ground truth for haplotype panel evaluation

We then sought to create a T2T-CHM13-native, haplotype-resolved panel of 1KGP SNP and indel variation. Our goal was to produce a dataset suitable for use as a reference for statistical imputation and phasing. To that end, we used the statistical phasing package SHAPEIT5[25]. This tool allowed us to pre-phase variants using a pedigree-aware phasing algorithm before performing the classical Li and Stephens HMM. SHAPEIT5's rare variant phasing algorithm is also able to phase singletons variants with moderate accuracy (< 40% SER).

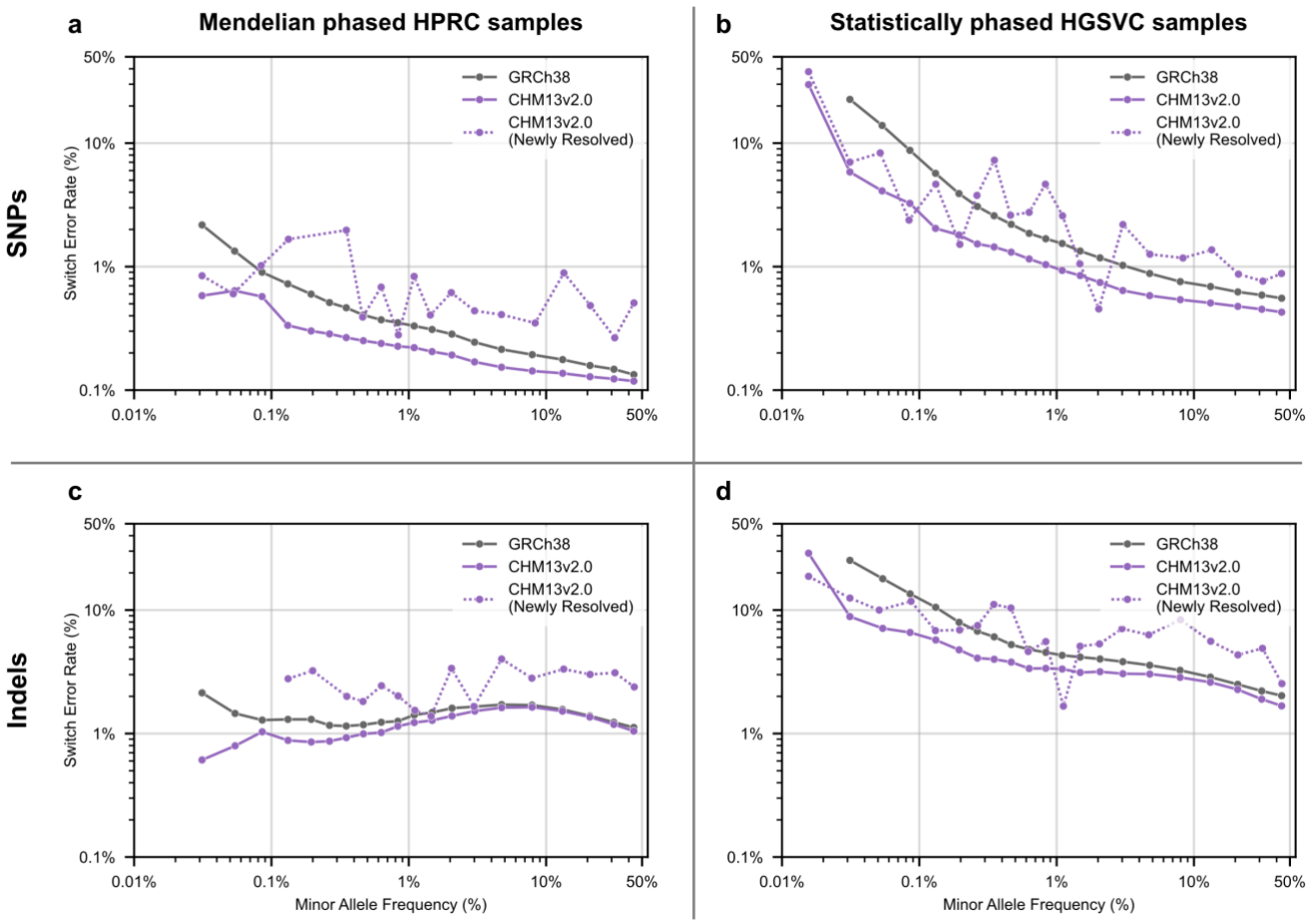

Figure 2. The T2T-CHM13 1KGP haplotype panel is more accurately phased than the GRCh38 1KGP haplotype panel. Of the 3202 participants in the 1KGP, 39 had their genomes assembled by the HPRC. An additional 61 had their genomes assembled by the HGSVC, 19 of which did not undergo Mendelian pre-phasing. a) The switch error rate of haplotype panel SNPs from 39 individuals, using their HPRC-assembled genomes as ground truth. Variants located in regions of T2T-CHM13 genome (CHM13v2.0) that are not present in GRCh38 are included in overall variant bin error rates but are also plotted separately as CHM13v2.0 (Newly Resolved) variants. b) SER of haplotype panel SNPs from 19 individuals, using their HGSVC-assembled genomes as ground truth. Unlike the 39 HPRC samples, these 19 individuals are not part of a 1KGP trio, and therefore were phased without any Mendelian-based error correction. c) Indel SER from the Mendelian-phased HPRC samples. d) Indel SER from the 19 individuals phased without Mendelian error correction. Switch error rates and average minor allele frequency per bin are displayed on a log scale.

The resulting panel contains 6404 haplotypes and 101,152,794 unique variants, including 305,914 high-confidence phased variants in previously unresolved regions of the T2T-CHM13 genome.

The standard measure of phasing accuracy is the switch error rate (SER), defined as the percentage of heterozygous variants which are incorrectly phased with respect to the prior heterozygous variant (Supplemental Figure 2)[31–33]. Most studies do not possess ground truth knowledge of phasing. Instead, incorrectly phased variants are operationally defined as variants where the offspring's phased genotypes switch parental origin from the prior heterozygous site–a phenomenon termed trio discordance.

Unfortunately, this method of identifying switch errors requires withholding parental samples from the panel being evaluated. The 1KGP dataset contains 1803 individuals from 608 families (6 parent-child duos, 602 trios) and 1399 non-duo/trio individuals (Supplemental Figure 3). It is also limited to samples that are part of family trios. Trio concordance measures would not reflect phasing accuracy of the 1399 non-trio 1KGP samples that were only phased using the Li and Stephens HMM method. (We term these samples "statistically phased.")

For these reasons, we instead utilized high quality phased T2T assemblies as a measure of ground truth phasing. These assemblies were generated from long read sequencing of patient derived cell lines from the HPRC and HGSVC[22,23]. The recently published HPRC draft human pangenome contains assemblies from 39 1KGP members (all trio probands), and the HGSVC recently released assemblies of 63 1KGP members (44 trio probands, 19 non-trio samples). Importantly, the presence of non-trio samples in the HGSVC dataset allowed us to separately assess the phasing accuracy of genomes that underwent Mendelian pre-phasing from those that were phased by purely statistical means.

1KGP short-read variant calls are more accurate when aligning to the CHM13v2.0 reference genome

Errors in variant calling are known to affect panel phasing accuracy[31]. Therefore, we began by assessing the concordance of our panel's genotype calls with the HPRC and HGSVC assemblies. One additional benefit of using assemblies as a source of ground-truth phasing is that they are reference-agnostic (i.e., phased variation is available in either GRCh38 or T2T-CHM13 coordinates), offering a common measuring stick to compare the two references. We were therefore able to compare this panel to the 1KGP consortium-released GRCh38 panel[19].

Overall, both panels' variant calls were ~99% concordant with variants derived from aligned assemblies (Table 1). We note that one challenge of using *de novo* assemblies as a ground truth for benchmarking was discrepancy in variant representation. Long-read assemblies and short-read variant calls often represent indels differently, especially in repetitive regions of the genome[34]. Accordingly, we observed higher rates of genotype discordance in indels than in SNPs. Common indels were more frequently erroneous than rare indels (Table 1, Supplemental Figure 4). Manual inspection of these errors indicated that discordant common indel genotypes often originate from globally diverse short tandem repeats, and GATK HaplotypeCaller represents these indels differently than the Minigraph-Cactus pipeline used to generate a VCF representation of pangenomic variation. It is unclear whether these discordant sites are due to true errors in variant calling or challenges in harmonizing indel representation between assemblies and short-read-derived variant callsets.

T2T-CHM13 1KGP variation was more concordant with assembled genome sequences than the GRCh38 1KGP variation (Table 1, Supplemental Figure 4). We were surprised to find that genotyping accuracy did not substantially vary by minor allele frequency; SNPs with only two minor alleles present in the panel (i.e., doubletons) were called almost as accurately as the overall panel (GRCh38: 0.51% discordant, T2T-CHM13: 0.45% discordant). While singleton variants were excluded from the GRCh38 panel due to concerns about genotyping accuracy, singleton SNPs in the T2T-CHM13 panel were only 1.41% discordant with assembly genotypes. When examining variants in regions that were syntenic between GRCh38 and T2T-CHM13, SNPs from the 1KGP NYGC GRCh38 panel were 0.42% discordant with assembled genomes, compared to 0.26% of syntenic SNPs from the T2T-CHM13 panel. Indels called in syntenic regions of the genome were also more discordant with assembled genomes in the GRCh38 panel (4.22%, Table 1) than in the T2T-CHM13 panel (4.00%). We found both panels to be more concordant with HPRC assemblies than HGSVC assemblies (Supplemental Table S1).

Prior work with 1KGP variants called against the CHM13v1.0 assembly had suggested that ~25% of non-reference SNV calls in nonsyntenic regions of the genome are false positives[2]. However, this estimate compared all GATK-passing variants against long-read HiFi sequences as ground truth. In our hands, comparing strictly filtered variants with reference agnostic assemblies resulted in a set of nonsyntenic T2T-CHM13 SNVs which only differed from assembled genomes at 4.60% of sites. Indels called in nonsyntenic regions of T2T-CHM13 had a discordant genotyping rate of 9.56% (Table 1). This is lower than the prior estimated false positive rates, but still two to six times the discordance rates observed in syntenic regions of the genome.

The 1KGP T2T-CHM13 haplotype panel is more accurately phased than the 1KGP GRCh38 panel in both previously unresolved and syntenic regions of the genome

We next sought to quantify the phasing accuracy of each panel. To maximize phasing accuracy, both panels have taken Mendelian inheritance rules into account when phasing trio samples. The NYGC GRCh38 1KGP

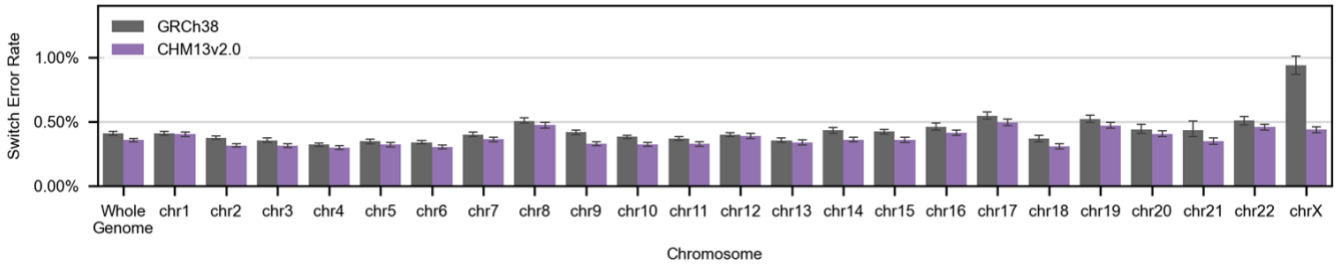

Figure 3. Whole genome and per-chromosome switch error rates from all 100 samples with HPRC or HGSVC assemblies. Ground truth phasing is defined by HPRC or HGSVC assembly. Error bars show +/- 95% confidence interval. T2T-CHM13 genome is noted as CHM13v2.0.

panel performed a post-phasing error correction on trio proband samples to ensure consistency between parental and child haplotypes. Similarly, our T2T-CHM13 1KGP panel was pre-phased according to Mendelian inheritance logic where possible, with the resulting haplotypes used as a scaffold for statistical phasing[25]. Unlike the GRCh38 panel, the phasing of both trio probands and trio parents was performed with trio-based information. In both panels, non-trio samples were only phased using the Li and Stephens statistical phasing algorithm[35]. As a result, we expect panel probands (19% of 1KGP samples) and parents (37% of 1KGP) to be more accurately phased than samples that are not part of a trio (44%).

Unlike the 1KGP dataset, all HPRC assemblies come from trio probands. Given the fact that the HPRC assemblies were slightly more concordant with our short-read variant calls (Supplemental Figure 4), we used those assemblies as a ground truth measure of the accuracy of the phasing of the probands in our dataset.

Phasing accuracy of the two panels broadly mirrored genotyping accuracy. For all sample and variant subsets, syntenic T2T-CHM13 variants were phased more accurately than GRCh38 variants (Figure 2). The most common SNPs were phased roughly two orders of magnitude more accurately than the least common SNPs (comparing minor allele frequency bins of 0.025% and 50%) for both Mendelian-phased (Figure 2a) and statistically-phased samples (Figure 2b). Indel SERs were substantially higher in both Mendelian- and statistically-phased samples, especially for common variants (Figure 2c, d). Unlike SNP SERs, indel SERs in both panels were stable across minor allele frequencies; for Mendelian pre-phased samples, common indels were phased slightly less accurately than rare indels. This result mirrors the genotyping accuracy result described previously.

We observed that samples that underwent Mendelian pre-phasing (Figure 2a, c) had lower SERs than samples that underwent statistical phasing (Figure 2b, d, Supplemental Table S1). SHAPEIT5 was able to phase singleton variants in statistically phased samples to accuracies that were consistent with previous reports[25] (Figure 2b, d; 29.77% for SNPs, and 28.85% for indels). In the GRCh38 panel, parental samples were phased roughly as accurately as statistically phased samples. In contrast, parental SERs in the T2T-CHM13 panel were closer to proband error rates (Supplemental Figure 5, Supplemental Table S1). This difference likely reflects the difference in phasing pipelines applied to the two datasets. In the GRCh38 panel, only trio probands underwent post-phasing Mendelian error correction.

Trio SER differed modestly by continental group. Trio parents in the Admixed American superpopulation showed disproportionately higher SER in the GRCh38 panel compared to individuals of other continental groups. This effect was not observed in the T2T-CHM13 panel. Instead, higher SER were observed in genomes from individuals from the African continental group in the T2T-CHM13 panel compared to individuals of other continental groups (Supplemental Figure 6). Taking the weighted average of proband error rates, parental error rates, and statistically phased error rates allowed us to estimate a panel-wide SER of 0.90% for the GRCh38 panel and 0.74% for the T2T-CHM13 panel.

As expected, SERs were substantially higher in regions of T2T-CHM13 that were previously unresolved in GRCh38 (Figure 2; Table 2; overall SER of 1.30% in T2T-CHM13 panel vs 0.35% in GRCh38 panel), reflecting the difficulties of genotyping in these complex and repetitive loci (Supplemental Figure 5). However, the difference was less substantial than anticipated. In statistically phased samples, SNPs in the previously

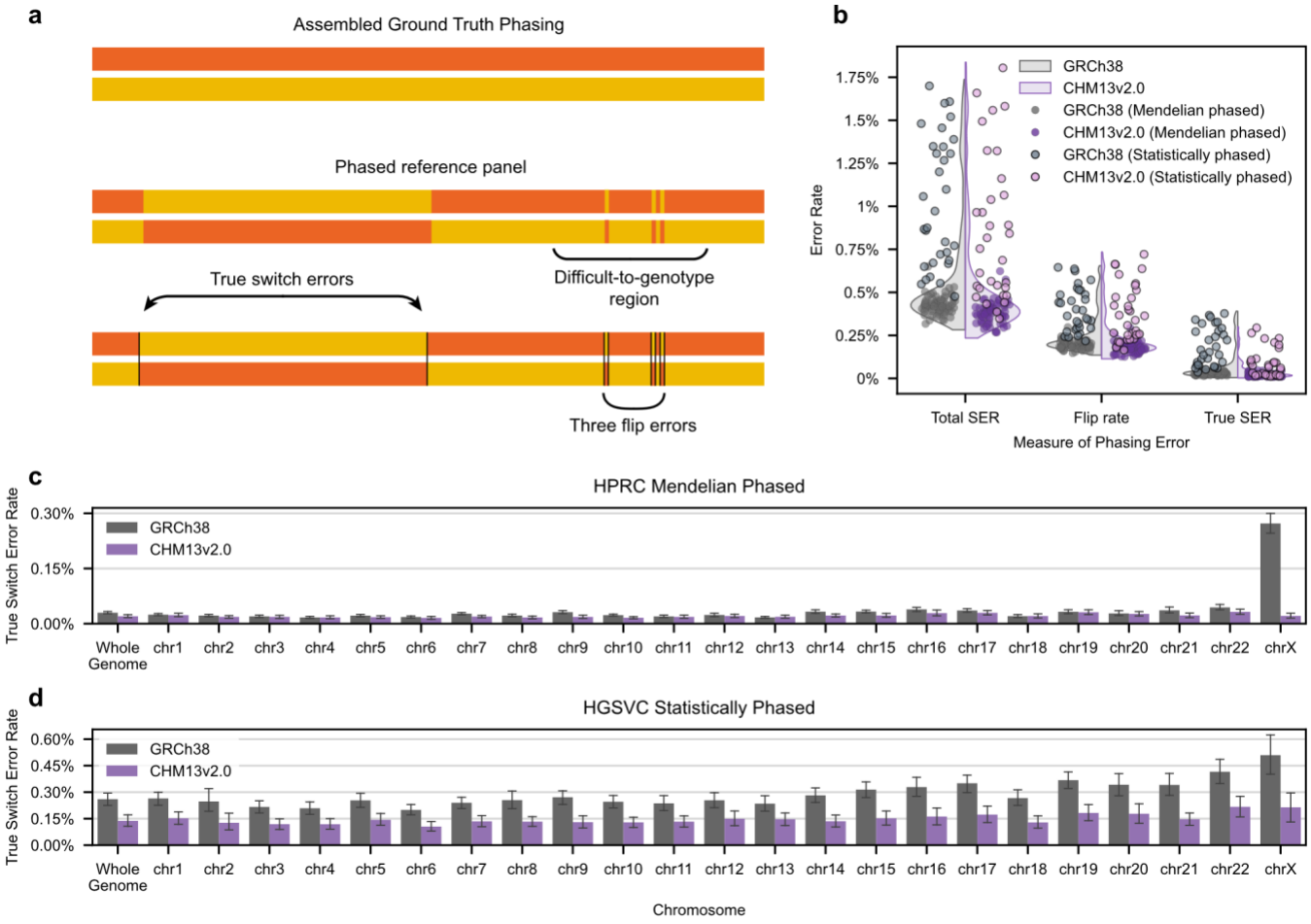

Figure 4. True switch error rates are lower in T2T-CHM13 phased 1KGP haplotypes compared to GRCh38 phased 1KGP haplotypes. a) Throughout this study, we have compared our panel's phasing with ground truth reference assemblies to identify switch errors, defined as sites at which the maternal and paternal haplotypes switch strands. However, regions which are difficult to sequence can often have genotyping errors that result in flip errors, defined as two consecutive switch errors. Unlike true switch errors, flip errors do not impact the phasing accuracy of nearby variants. Eight switch errors are in the illustrated haplotypes, but only two of these are true switch errors - the other six switch errors are part of three flip errors. b) The SER, flip error rate, and true SER rate of all 100 samples with assembly-based ground truths. Data from the 19 HGSVC individuals that were phased without mendelian error correction are highlighted. Data points are plotted on top of a violin plot showing the overall distribution of error rates. c) Whole genome and per-chromosome true SERs from 39 HPRC-assembled samples. d) Whole genome and per-chromosome true SERs from 19 HGSVC-assembled samples phased without Mendelian error correction. T2T-CHM13 genome is noted as CHM13v2.0.

unresolved regions of T2T-CHM13 were phased roughly as accurately as syntenic GRCh38 SNPs (1.62% vs 0.98%) (Figure 2a, b).

SERs were lower in the T2T-CHM13 panel by roughly the same proportion for all chromosomes except for chromosome X, which was phased twice as accurately in the T2T-CHM13 panel compared to GRCh38 (0.91% vs 0.42%) (Figure 3). Most of the improvement in chromosome X SERs were observed in the PAR2 and non-PAR regions of chromosome X (Supplemental Figure 5). This likely reflects updates to the software packages used to phase these panels. Specifically, native support for mixed diploid/haploid phasing was first introduced in SHAPEIT4, after the release of the 1KGP GRCh38 panel. The release of this feature brings chromosome X SER in line with the rest of the genome.

Most switch errors likely reflect genotyping errors instead of errors in phasing

Switch errors can be the result of inaccurate phasing or inaccurate genotyping. Inaccurate phasing tends to result in scattered haplotype switches. Inaccurate genotyping, however, frequently results in two pseudo-switch errors occurring immediately around the erroneous genotype. These paired, back-to-back haplotype switches are called 'flips' (Figure 4a, Supplemental Figure 2)[31]. In addition to calculating the SER, we also determined the true SER, defined as the percentage of heterozygous calls that are non-flip associated switches. We discovered that the majority of switch errors in both panels were in fact associated with flip events, especially for samples that underwent Mendelian phasing. In line with the theory that flip errors arise from genotyping error, we observed that SERs and flip error rates were correlated with genotyping error rates (SER r^2 = 0.31, flip error rate r^2 = 0.35, Pearson correlation), while true SERs were only loosely correlated with genotyping error rates (r^2 = 0.006) (Supplemental Figure 8). The measured flip error rate for HPRC probands from the 1KGP NYGC GRCh38 panel was 0.186%, while the true SER was 0.029%. HPRC probands from the T2T-CHM13 panel had a flip error rate of 0.166% and a true SER of 0.018%. (Figure 4b). As expected, SERs were higher in statistically phased samples compared to Mendelian phased samples. However, after excluding flip errors, the true SER of some statistically phased T2T-CHM13 samples were phased almost as well as Mendelian phased samples (Figure 4b).

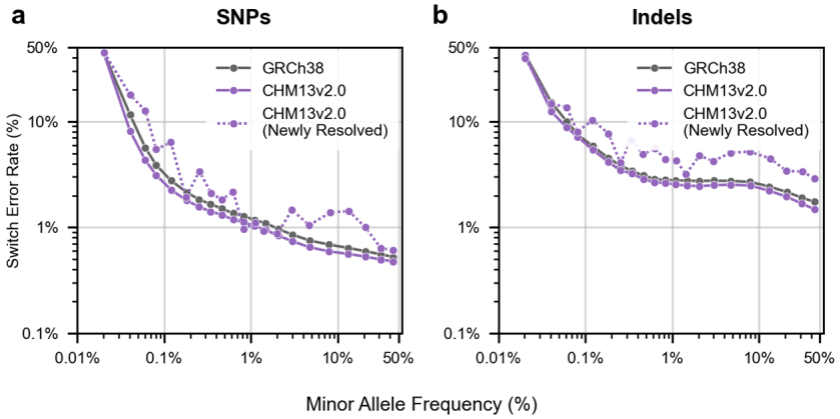

Figure 5: Accuracy of out-of-panel variant phasing when using a GRCh38 or T2T-CHM13 1KGP haplotype panel as reference. All HPRC-assembled samples and their relatives were removed from the GRCh38 1KGP and the T2T-CHM13 1KGP (CHM13v2.0) panels of 2504 unrelated samples, yielding 'non-HPRC' panels containing data from 2426 individuals. Filtered short-read derived SNPs from 39 HPRC-assembled individuals were phased, using either the GRCh38 or T2T-CHM13 2426 member panel as a source of reference haplotypes for SHAPEIT5. a) SNP and b) Indel SERs were measured by comparison with ground truth pangenomic assemblies, and variants were binned by reference panel minor allele frequency (MAF). SERs and average MAF per bin are displayed on a log scale.

True switch error rates in non-trio T2T-CHM13 haplotypes are reduced by 50% compared to GRCh38 haplotypes

When stratified by chromosome, we observed a ten-fold reduction in true SERs on the T2T-CHM13 Chromosome X (GRCh38: 0.26% true SER; T2T-CHM13: 0.016% true SER). True SERs on acrocentric chromosomes and Chromosome 9 particularly benefited from the phasing workflow used to generate the T2T-CHM13 panel (Figure 4c, d). A large proportion of these chromosomes' sequences are newly resolved in the T2T-CHM13 reference genome, and so we hypothesized that resolving previously unknown sequences allowed for improved phasing of variation on these chromosomes. Supporting that hypothesis, we found that the per-chromosome improvement in true SER was associated with the percentage of chromosome length that was previously unresolved (Supplemental Figure 9, p=1.6e-8 by two-sided t-test).

Use of a T2T-CHM13 1KGP haplotype panel when phasing out-of-panel samples reduces phasing error

We next assessed the accuracy of variant phasing when phasing out-of-panel samples using our T2T-CHM13 haplotype panel as a reference. For these experiments, we considered the draft human pangenome as ground truth[22]. It is best practice to use reference panels that contain unrelated samples when phasing or imputing genomic variation, and so we used the 2504 sample version of the 1KGP NYGC GRCh38 and T2T-CHM13

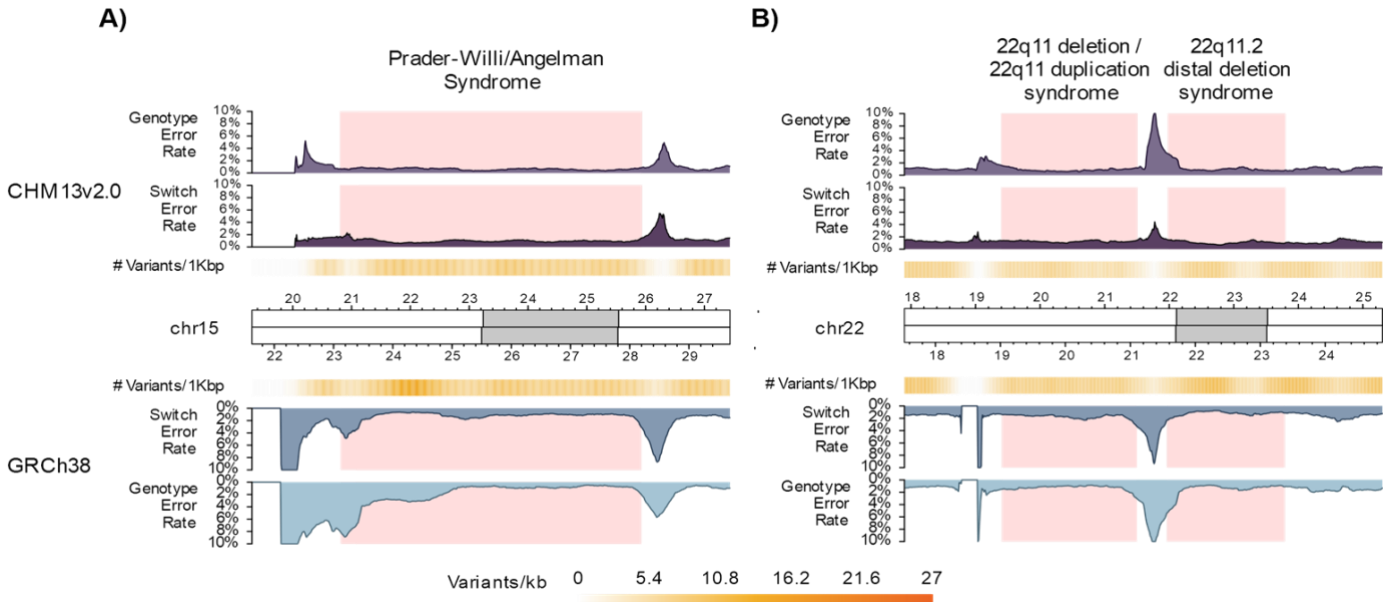

Figure 6: Genotyping and switch error rates spike in areas flanking CNV disorder associated regions. Each panel shows average error rates obtained when phasing short read variant calls from 39 individuals using either the T2T-CHM13 1KGP panel (CHM13v2.0) or the GRCh38 NYGC 1KGP panel as a reference. Data from variants phased using the T2T-CHM13 panel is illustrated in purple at the top of each panel, and data from variants phased using the GRCh38 panel is illustrated in blue at the bottom of each panel. In the center of each panel is an ideogram illustrating the CHM13v2.0 coordinates (top half) or GRCh38 coordinates (bottom half) displayed. In gold is the number of phased variants per 1000 bp; darker shades indicate more variants. DECIPHER-defined CNV disorder regions are highlighted in red. a) Variant density and error rates around the genomic region (+/- 1.5mb) commonly deleted in Prader-Willi/Angelman Syndrome (GRCh38 coordinates: chr15:23123712-28193120, CHM13v2.0 lifted coordinates: chr15:20807475-25935855). b) Variant density and error rates around the genomic regions (+/- 1.5mb) commonly affected in 22q11 syndrome and 22q11.2 distal deletion syndrome (GRCh38 coordinates: chr22:19022279-23380258, CHM13v2.0 lifted coordinates: chr22: 19397653-23803198). Error rates were first calculated in non-contiguous blocks of 10,000 bp. Then, a rolling average error rate was calculated, centered on a 500kb window.

1KGP reference haplotype panels. This version of the 1KGP dataset removes trio probands, leaving only the unrelated trio parents and non-trio individuals. We then removed the parents of pangenomic samples, leaving 2426 unrelated samples in each reference panel.

We initially evaluated each haplotype panel as a reference for variant phasing. To do this, we used SHAPEIT5 to phase the previously described filtered short-read variant calls from the 39 HPRC samples present in the 1KGP dataset. SHAPEIT5's rare variant phasing algorithm does not work with small datasets or reference panels. Therefore we were only able to use its common variation algorithm, preventing us from phasing singleton variants. When phasing the HPRC short read variant genotypes, we observed lower SERs when using 1KGP T2T-CHM13 phased haplotypes as a reference haplotype panel. (Figure 5). The improvement over use of the 1KGP GRCh38 panel was small and consistent across minor allele frequencies. As expected, nonsyntenic variants were phased less accurately than syntenic variants (SNPs: 1.52% SER vs 1.01% SER; Indels: 4.00% SER vs 2.10% SER).

Genomic regions associated with disease-causing CNVs disproportionately benefit from use of a T2T-CHM13 reference panel

Improvements in out-of-sample phasing performance were not uniform throughout the genome. To gain insight into the kinds of genomic regions where phasing was most affected by the choice of reference panel genome, we calculated the SERs of the rephased pangenome samples, stratified by cytogenetic band (Supplemental Table S1). We found that while a minority of regions (12%) were more accurately phased using the 1KGP NYGC GRCh38 reference haplotype panel, most cytoband regions (88%) were phased more accurately when

using a T2T-CHM13 reference haplotype panel. Eight of the fifteen regions which benefited the most from the T2T-CHM13 reference panel (Table 2) are co-located with DECIPHER-defined common copy number variant (CNV) disorders, including 22q11.2 (DiGeorge Syndrome) and 15q11.2 (Angelman Syndrome and Prader-Willi Syndrome). This is a statistically significant enrichment for DECIPHER disorder-associated loci (p=0.0011, weighted GSEA permutation test[36]).

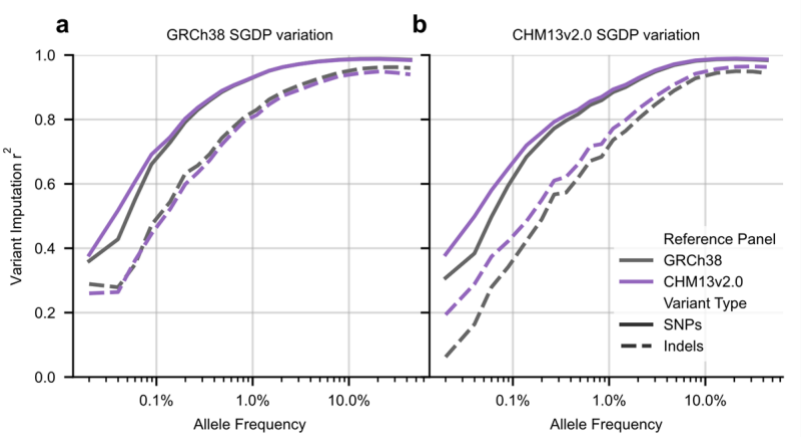

Figure 7: Imputation of genomic variation in 256 non-1KGP Human Genome Diversity Project (HGDP) samples, using 1KGP haplotype panels as references. a) Genotyping array data was simulated by downsampling variation derived from short reads aligned to GRCh38 to those variants present in the Infinium Omni2.5 genotyping array. Variation was imputed with the 1KGP GRCh38 reference panel or the 1KGP T2T-CHM13 reference panel (CHM13v2.0) lifted to GRCh38. b) Downsampled CHM13v2.0 HGDP variation was phased and imputed with the 1KGP GRCh38 reference haplotype panel or the 1KGP T2T-CHM13 reference panel lifted to GRCh38 coordinates. Variants that were not present in both panels were removed from analysis. All imputed variants were binned by the minor allele frequency of the variant in the reference panel used during imputation. r^2 values were calculated per bin. r^2 statistics are stratified by SNPs and Indels. Average minor allele frequency per bin is displayed on a log scale.

To investigate the nature of this improvement, we calculated the panel genotype error rate in non-overlapping 10 kbp intervals across the genome. We also determined the SER of short read variation from pangenomic samples phased using either panel as a reference. As a general trend, we observed that both panel genotype error rates and out-of-panel phased SERs frequently peaked on either side of regions of the genome associated with CNV disorders (Figure 6). Improvements in panel genotyping accuracy and out-of-panel phasing accuracy were observed within CNV disorder regions, but were much more apparent in their flanking regions. This pattern was true for both panels, but was especially pronounced when genetic variation was phased with the GRCh38 panel. Often, these spikes were associated with a drop in variant density, suggesting that these are regions which are difficult to genotype. In some instances, such as the loci flanking the 22q11 region (Figure 6b), the T2T-CHM13 panel was not appreciably genotyped more accurately than the GRCh38 panel, but sample phasing error was still much lower when using the T2T-CHM13 panel as a reference. This suggests that improvements in phasing with the use of a T2T-CHM13 panel are not solely due to improved panel genotyping.

The Prader-Willi/Angelman region on chromosome 15 is a particularly illustrative example of how the T2T-CHM13 genome affects variant calling and phasing in complex genetic regions. The 15q11-q13 locus is rich in segmental duplications and complex mechanisms of gene regulation, including parent-specific imprinting. Coding genes *MKRN3, MAGEL2, NDN* and *SNURF-SNRPN* (along with several different snoRNAs and lncRNAs) are exclusively transcribed from the paternal allele, while *UBE3A* and *ATP10A* are exclusively transcribed from the maternal allele[37]. Therefore, accurate phasing is essential to identify disease-modifying variants[38,39].

The previously published NYGC GRCh38 1KGP panel contains a region of high variant density near the Prader-Willi breakpoint 2 (chr15:24100000-24600000). This region contains numerous segmental duplications and is known to be a region of poor mappability (Supplemental Figure 10). Genotype discordance in this region is substantially higher in the GRCh38 panel than the T2T-CHM13 panel. Importantly, this difference is not due to improvements in reference accuracy, as the GRCh38 sequence at this locus is 99.9% identical with CHM13v2.0. Instead, we suspect that reads from other similar regions of the genome no longer misalign to this

region and therefore do not induce errors in variant calling. This example illustrates how the use of the CHM13v2.0 reference genome can lead to improved variant calling, even in syntenic genomic regions.

Imputation of rare SNPs is improved by use of a T2T-CHM13 panel

We next sought to assess the effect of using a T2T-CHM13 1KGP reference haplotype panel when imputing missing genetic variation. Variant imputation is most commonly used to increase the resolution of microarray or low-coverage sequencing-based genomics studies, where whole genome sequencing can be cost prohibitive. To mimic this use case, we downsampled GRCh38 and T2T-CHM13 whole genome sequencing data from 256 participants in the Simons Genome Diversity Project (SGDP) to the set of variants that are assayed by Illumina Omni 2.5 genotyping array. We chose this dataset because it is highly diverse (containing 60 populations compared to 26 populations in 1KGP) and it is one of the few well-studied, publicly available variation datasets that has been mapped to both reference genomes using the same pipeline[5].

To ensure that imputation accuracy was based on the same set of variants in both builds, we wanted to compare variants imputed using the GRCh38 panel with variants imputed using the T2T-CHM13 panel lifted to GRCh38 coordinates (Figure 7a). However, lifting variant sets between genomic references can introduce erroneous genotypes that may propagate to imputation. To ensure a fair comparison between the two panels, we imputed SGDP variation in both T2T-CHM13 coordinates and GRCh38 coordinates (Figure 7b). In both cases, we only examined variants that were present in both the lifted over and genome-native panel. After observing challenges with indel imputation from lifted over panels, we wrote a liftover script that implemented many of the same improvements to indel liftover that are used in the recently released bcftools liftover plugin[18,40] (see Methods). Our observed indel r^2 when imputing from a lifted-over panel substantially improved after switching liftover techniques, illustrating how liftover methods can substantially affect the accuracy of lifted-over datasets (Supplemental Figure 11).

When imputing SGDP variation from pseudo-array variant calls, the use of a T2T-CHM13 1KGP reference panel resulted in comparable or slightly improved imputation accuracy compared to the use of a GRCh38 reference panel. SNP imputation was consistently enhanced by the use of the T2T-CHM13 panel, even when working with a lifted over panel in GRCh38 coordinates (Figure 7a). As previously observed[25], indel imputation was less accurate than SNP imputation, regardless of the reference panel. Indel imputation was best when working with a panel that had not been lifted between builds. As previously reported, the 1KGP panel struggles to impute genotypes of individuals from African and Oceanic populations to the same level of accuracy as American, European, and Asian populations, likely due to the absence or limited representation of these populations in 1KGP[19]. We observed population-specific differences in imputation accuracy in both panels, though no population's overall imputation accuracy was significantly affected by the use of either panel (Supplemental Figure 12, 13).

Discussion

Use of the T2T-CHM13 reference genome has been shown to improve SNP and indel discovery and genotyping in both short- and long-read datasets[2]. However, researchers and clinicians have held back on using updated reference genomes in the past due to uncertainty around the risks and benefits of such a transition[41].

Here we present recombination maps and a phased imputation panel derived from CHM13v2.0-aligned 1KGP short-read DNA sequencing data. Thoroughly benchmarking our panel against the commonly recommended GRCh38 1KGP panel shows that the use of a T2T-CHM13 reference improves the degree of genotype and phasing concordance between GATK-called variation and long read/Hi-C based genome assemblies. Using modern phasing packages that appropriately handle chromosome X haplotypes resulted in a 10-fold reduction in chromosome X true SER, while applying Mendelian trio phasing before statistical phasing halved the true SER in the statistically phased cohort.

We note that the increase in variant calling accuracy was largely limited to SNPs. Indel variants in both panels were more discordant with assemblies than expected. This class of variation is difficult to benchmark with concordance-based genotype comparison, as differences in variant representation likely inflate our genotyping error estimates.

Population haplotype panels are most commonly used as references when phasing and imputing variation from array-based genotyping calls. In comparison to the GRCh38 1KGP reference panel, we find that using the T2T-CHM13 1KGP panel produces more accurately phased and imputed calls. We must note that using this panel as a reference when imputing genetic variation mainly improves the accuracy of rare imputed SNPs and that imputation of common variation is roughly equivalent between the two panels. When using the panel as a reference during genetic phasing, we observe a small but consistent improvement in phasing accuracy when using the T2T-CHM13 reference panel as compared with GRCh38. Critically, the improvement in phasing accuracy is particularly large when phasing genetic variation around structurally complex regions that are commonly associated with human disease. We hypothesize that this improvement is in part due to the panel's substantially more accurate genotyping in these regions.

There is still room to expand upon these advances in variant calling, phasing, and imputing accuracy. The T2T-CHM13 callset was produced using the GATK short read germline variant calling workflow, which utilizes a set of lifted over references to calibrate variant quality scores. These reference datasets do not include variation from the newly resolved regions of the T2T-CHM13 genome, and therefore quality score-based cutoffs are likely to be less accurate in these regions.

Importantly, this panel was produced using short read-based variant calls. This limitation is common in the field and applies to most available reference haplotype panels. However, most of the newly resolved sequence in T2T-CHM13 is repetitive in nature[2], and thus inaccessible to short read variant calling. Large structural variation, centromeric satellites, telomeres, and surrounding regions experience unique dynamics of mutation, recombination, and selection[42–44]. Future development of a panel that accurately measures these dynamics will be of great use. In our hands, including these regions improves variant calling and phasing in short read-accessible regions of the genome, but accuracy within these newly accessible regions remains poor. We look forward to forthcoming population-scale long-read sequencing data from the 1000 Genomes Project and other diverse population studies. These long-read datasets should soon enable accurate variant calling, phasing, and imputation within these newly accessible regions, unlocking these complex loci for association studies, evolutionary analyses, and other applications[22,45–48]. Until those datasets are available, our short-read-based resources can be paired with previously released genome accessibility masks that define regions of T2T-CHM13 where variant calls are most reliable[24].

We hope that our study will broaden awareness of the importance of using a complete human genomic reference for common genetic applications. In addition to allowing for the measurement of variation in new regions of the genome, the T2T-CHM13 reference allows mapping software to resolve ambiguously mapped reads in structurally complex genomic regions. Our work demonstrates that the benefits of using this reference are particularly large around these disease-associated loci, with potential implications for both fundamental biological research and clinical care. To facilitate these goals, the recombination maps and reference panel are freely available at https://github.com/marbl/CHM13.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Donna Werling (dwerling@wisc.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- T2T-CHM13 population-specific and global averaged recombination maps are available at https://zenodo.org/records/14891074 and https://github.com/JosephLalli/phasing_T2T/tree/main/resources/recombination_maps/t2t_native_scaled_maps
- SHAPEIT5 can be found at https://github.com/odelaneau/shapeit5
- SHAPEIT5 with ability to measure flip events can be found at https://github.com/JosephLalli/shapeit5/tree/main
- 1KGP T2T-CHM13 phased variant calls are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000_Genomes_Project/chm13v2.0/Phased_SHAPEIT5_v1.1
- 1KGP T2T-CHM13 unphased variant calls are available at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/1000_Genomes_Project/chm13v2.0/all_samples_3202/
- 1KGP GRCh38 phased variant calls are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/
- 1KGP unphased variant calls are available at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_raw_GT_with_annot/
- Combined HGSVC3 and HPRC pangenome available in CHM13v2.0 coordinates at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Graph_Genomes/1.0/2024_02_23_minigraph_cactus_hgsvc3_hprc/hgsvc3-hprc-2024-02-23-mc-chm13-vcfbub.a100k.wave.norm.vcf.gz
- Combined HGSVC3 and HPRC pangenome available in GRCh38 coordinates at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Graph_Genomes/1.0/2024_02_23_minigraph_cactus_hgsvc3_hprc/hgsvc3-hprc-2024-02-23-mc-chm13.GRCh38-vcfbub.a100k.wave.norm.vcf.gz
- GATK Resource bundle in CHM13v2.0 coordinates can be found at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/GATK_CHM13v2.0_Resource_Bundle/
- T2T-CHM13 short-read accessibility mask is at https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/accessibility/combined_mask.bed.gz
- Simons Genome Diversity Project variant calls can be found in T2T-CHM13 coordinates at https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/CHM13/assemblies/variants/SGDP/chm13v2.0/ and GRCh38 coordinates in the T2T_ChrY Anvil workspace at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T_CHRY under SGDP_GRCh38_chromosome
- The pedigree of 1KGP samples can be found at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/1kGP.3202_samples.pedigree_info.txt
- Bed file of nonsyntenic CHM13v2.0 regions is at https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/chain/v1_nflo/chm13v2-unique_to_hg38.bed
- CHM13v2.0 cytoband coordinates can be found at https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/annotation/chm13v2.0_cytobands_allchrs.bed
- Chain files to lift from CHM13v2.0 to GRCh38 and vice-versa are at https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/CHM13/assemblies/chain/v1_nflo/grch38-chm13v2.chain
- Code used to lift variation between CHM13v2.0 and GRCh38 is at https://github.com/JosephLalli/LiftoverIndel
- Code used to generate recombination maps is at https://github.com/andrew-bortvin/1kgp_chm13_maps
- Bash scripts to phase panel are available at https://github.com/JosephLalli/phasing_T2T
- Scripts and Jupyter notebooks to generate all figures are available at https://github.com/mccoy-lab/1kgp_chm13_maps and https://github.com/JosephLalli/phasing_T2T. Original code used to generate the figures in this publication has been deposited at Zenodo at https://zenodo.org/record/7612953 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

The authors would like to thank Dr. Oliver Delaneau and the T2T Consortium for their generosity and advice during the development of our benchmarking methods, and to Dr. Samantha Shapiro for her constructive feedback during manuscript drafting. This paper could not have been attempted, let alone completed, without Nils Irland and the generous computing support of the University of Wisconsin-Madison Laboratory of Genetics and the UW-Madison Center for High Throughput Computing. Thank you to the Johns Hopkins Department of Biology and Center for Computational Biology, as well as members of the McCoy laboratory for constructive feedback. We also thank the staff of the Maryland Advanced Research Computing Center for computing support. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award R35GM133747 to RCM, the Brain and Behavior Research Foundation 29815 to DMW, US Department of Agriculture National Institute of Food and Agriculture Hatch Act Formula Fund WIS04078 to DMW, Simons Foundation Autism Research Initiative 606289 to DMW, and the University of Wisconsin-Madison Medical Scientist Training Program T32GM140935 to JLL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

Conceptualization: J.L.L., A.N.B., R.C.M., D.M.W.; Methodology, investigation: J.L.L. and A.N.B. Writing—original draft, J.L.L. and A.N.B.; writing—review & editing, J.L.L., A.N.B., R.C.M., D.M.W.; funding acquisition, R.C.M., D.M.W.; resources, R.C.M., D.M.W.; supervision, R.C.M., D.M.W.

Declaration of Interests

The authors declare no competing interests.

Supplemental Information

Document S1. Figures S1–S13, Supplementary Table S1

Supplementary Data S1. Table of variance in recombination rates in 1MB windows.

Supplementary Data S2. Per-sample variant counts, error counts, switch error rates, and genotyping error rates stratified by panel, method of phasing, and ground truth data source.

Supplementary Data S3. Per-MAF bin variant counts, error counts, switch error rates, and genotyping error rates stratified by panel, method of phasing, syntenic/nonsyntenic status, and ground truth data source.

Supplementary Data S4. Imputation summary statistics by subject, panel, variant category, and syntenic/nonsyntenic status.

Methods

Generating recombination maps

We generated recombination maps for our dataset using pyrho[26,27]. As input, we used unphased VCFs. VCFs were filtered as described in “Phase T2T-CHM13 1000 Genomes Project variant calls”, removing low confidence genotype calls and SNPs. We used previously published estimates of historic population size inferred by smc++[49]. VCFs were split into 26 sets of population-specific VCFs. Population-specific lookup tables were calculated using a Moran population size 150% of the sample size. We selected population-specific hyperparameters from the potential hyperparameters suggested by pyrho by identifying the parameters that minimized the L2 norm for each population.

To generate masked recombination maps, we applied a short-read accessibility mask[24,50] after filtering, but prior to recombination mapping. The same mask was applied to recombination maps immediately after creation. To create a single map for all populations, we first scaled recombination maps for individual populations to deCODE cumulative map lengths. We then computed an average per-base recombination rate across all populations, weighing the average by the sample size of the population within 1KGP.

The X-chromosome was divided into PAR and non-PAR regions. Recombination mapping of PAR regions was done identically to autosomal mapping, using the same smc++ demographic histories and hyperparameters as for autosomes. Non-PAR regions could not be mapped alongside PAR regions and autosomes, as they have a separate demographic history, and all males are haploid. For non-PAR recombination mapping, we rescaled smc++-inferred demographic histories by ¾, reflecting the lower frequency of X-chromosomes. We calculated lookup tables and hyperparameters for these non-PAR-specific regions using this rescaled demographic history. To make all samples in the population diploid, we created “pseudodiploid” males by randomly selecting pairs of haploid males and combining their variant calls. Pyrho was run on pseudodiploid VCFs. PAR and non-PAR recombination maps were combined for each population. As with autosomes, we standardized cumulative X chromosome map lengths prior to averaging recombination calls across populations. For standardization, we used deCODE sex-averaged recombination maps.

Correlation of Maps and Hotspot Comparisons

To measure correlation between recombination maps, we binned our maps into bins. We used 1MB, 100kb, and 10kb bins. Within each bin, we computed the genetic distance in centimorgans. We then measured the Spearman correlation between pairs of binned maps.

Hotspots were identified as in Halldorsson et al 2019[30], with any region with a recombination rate at least tenfold greater than the genome-wide average called as a hotspot. To compare hotspot landscapes between maps, we ran bedtools intersect[51], classifying hotspots that overlap by at least one nucleotide as hotspots that are shared between maps.

Phasing T2T-CHM13 1000 Genomes Project variant calls

In previously published work[24], Illumina-generated short reads from 3,202 samples collected as part of the 1000 Genomes Project were aligned to the CHM13v2.0 reference genome using bwamem, and variants called using the functionally equivalent GATK HaplotypeCaller pipeline. Of note, the “resource bundle” of true positive variants used to calibrate this pipeline was previously created by lifting over the GRCh38 resource bundle using GATK LiftoverVCF. Calibration scores should be approached with caution in previously unresolved genomic regions, as HaplotypeCaller had no ground truth variants as a source of calibration in these regions.

Our group downloaded this variant callset from the T2T Consortium's public AWS bucket. Using a bash script, VCFs containing variants from the autosomal chromosomes and chromosome X were atomized, left-normalized, and multiallelic variant calls were split into biallelic calls using bcftools' norm command. Haploid chromosome X variation was converted to diploid. The number of mendelian violations present at each site was calculated using bcftools mendelian2 plugin. To help ensure equivalent variant representation, the same procedure was applied to the VCF representations of the HPRC pangenome and the combined HPRC-HGSVC3 pangenome[23] (see data availability).

We removed variants from further analysis if they had: 1) more than 5% of alleles uncalled or missing; 2) mendelian errors comprised more than 5% of all non-missing variant calls; 3) no within-superpopulation Hardy-Weinberg equilibria of more than 1e-10; 4) a Variant Quality Score Log-Odds (VQSLOD) of less than 0, indicating that the variant is more likely to be a false call than a true variant under GATK's Variant Quality Score Recalibration model; 4) A "*" alt allele; 5) Any variant with a reported filter value of anything other than "PASS"; 6) Any variant with an minor allele count (MAC) of 0. Because a structural variant caller was not used when generating the T2T-CHM13 variant callset, we additionally removed all variants whose alternative allele and reference allele differed in length by 50 or more base pairs.

The remaining SNVs and indels from all 3202 samples were statistically phased using the two-step method recommended by the creators of SHAPEIT5. First, variants with a panel MAF above 0.1% were phased using SHAPEIT5_common v5.1.1[25]. The 1000 Genomes pedigree was obtained from the pedigree file on the International Genome Sample Resource (IGSR)’s FTP website. The following non-default settings were used

to maximize accuracy: An MCMC iteration scheme of "10b + 1p + 1b + 1p + 1b + 1p + 1b + 1p + 10m" was used, a PBWT depth of 8, and an HMM window of 5cM. SHAPEIT5_common was applied to whole chromosomes, with no chunking strategy employed. Chromosome X was split into the PAR1 region, the PAR2 region, and the central chromosome X region before phasing. The non-PAR chromosome X region was phased in a haploid-aware manner by using the --haploid option and providing SHAPEIT5 with a list of male samples. Testing revealed that phasing accuracy was maximized when using an Ne value of 135000.

After obtaining a panel of phased common variants, the remaining variants were phased using SHAPEIT5_rare v5.1.1. Briefly, SHAPEIT5_rare uses the phased common variant panels produced by the previous step as a scaffold. Rare variants are then sequentially phased into the scaffold, with the most likely haplotype for each variant determined using the Li and Stephens HMM method. Singleton variants are assigned a haplotype using an inheritance by descent approach to determine which haplotype is likely older on an evolutionary scale. Rare variants were phased per chromosome, in 40mb chunks with 1mb overlapping, using default settings. A sample pedigree was also provided at this step to allow for pedigree-informed pre-phasing and mendelian error correction. Variants in PAR regions and non-PAR chromosome X variants were split in the same way as common variants. The chunked rare variant output was ligated with bcftools concat -l. Both SHAPEIT5_common and SHAPEIT5_rare simultaneously imputed any missing variant calls.

Statistical analysis of phased variant panels

The T2T-CHM13 variant panel produced above was compared to the panel of phased variants in GRCh38 coordinates previously produced by the NYGC and published by the 1000 Genomes Consortium[19]. This panel was downloaded from the public 1000 Genomes FTP server, using the 'v2' version of the chromosome X panel. Unphased, unfiltered GRCh38 variant calls were also downloaded from the same server.
To measure the flip error rate and SER of the GRCh38 and T2T-CHM13 phased panels, we used SHAPEIT5_switch v1.1.1. SHAPEIT5 can determine if a heterozygous site is a switch error by either using parental haplotypes as a source of ground truth haplotypes (termed ‘trio concordance’) or comparing a phased panel to a ground truth set of phased variants (termed ‘empirically phased ground truth’). (For more information, see Figure 4a or Supplemental Figure 2.) We utilized both methods of identifying switch errors. SERs calculated via trio concordance are indicated as such in text and figure legends. Unless indicated as such, SERs were calculated using empirically phased ground truths.
For the 39 samples present in both the 1000 Genomes dataset and the HPRC draft human pangenome, the vcf representation of the HPRC v1.1 pangenome (with nested variants of 100,000bp or larger collapsed using vcfbub[52]) was provided to SHAPEIT5_switch as ground truth. For the additional 61 samples that are present in the 1000 Genomes datasets and the HGSVC3 assembly release, the vcfbub[52]-processed version of the vcf representation of the combined HPRC-HGSVC3 pangenome[23] was used as ground truth. GRCh38 panels were compared to GRCh38-referenced pangenome vcfs, and CHM13v.20 panels were compared to T2T-CHM13-referenced pangenome vcfs. Male samples were dropped from analysis when calculating error rates of non-PAR chromosome X regions. SHAPEIT5_switch reports per-variant statistics on the number of switch errors, the number of heterozygous genotypes, the number of discordant variant calls, and the number of variant calls checked for discordancy. In addition, we modified SHAPEIT5_switch to calculate the number of flip errors and switch errors (available at https://github.com/JosephLalli/shapeit5/tree/main).

The reports generated by SHAPEIT5_switch were then gathered into one dataframe using polars in a custom python script[53]. We encountered the same issues with private trio variants that Byrska-Bishop (2022) did[19], and excluded these variants from analysis. For any group of variants discussed (per MAF bin, per chromosome, per-sample, etc.), per-group error rates were calculated via custom python script by first summing the per-group numerator and denominator. For example, the SER per MAF bin was calculated by summing the number of switch errors per MAF bin and dividing that value by the sum of the per-variant heterozygous genotype count in that bin. Where possible, calculated summary error rates were compared to those calculated by SHAPEIT5_switch to ensure the accuracy of our code.

Bedfiles obtained from UCSC Genome Browser were used to identify syntenic and nonsyntenic regions of GRCh38 and T2T-CHM13. We used bcftools annotate to tag each panel variant as syntenic or nonsyntenic, and bcftools query to extract that information into a per-variant tsv file that could be imported into a python script.

Measuring performance in out-of-panel samples

To determine the switch error rate of out-of-panel variants phased with either panel, we removed 78 parents of HPRC samples from the 1KGP 2504-member unrelated panel to produce a 2426 member panel. Additionally, we created unphased variant files that only contained the 39 1KGP participants present in the HPRC panel. We then applied the same method outlined above to phase HPRC members' variation. We were unable to use the SHAPEIT5_rare program to phase rare variation, as SHAPEIT5_rare only processes variants with a MAF less than 0.1%, and even a singleton variant in a 39-member panel has a greater than 0.1% MAF. Phasing statistics were gathered and summarized as per above.

Cytoband and CNV-specific switch error rates

Cytoband switch error and genotype error statistics were calculated by grouping variants by cytoband, and then applying the method described above. CNV disorder coordinates were obtained from the DECIPHER group website and used to subset per-variant switch error and genotyping error statistics. Rolling average SER and genotype error rates were calculated via the 'dynamic group by' function of pandas. Briefly, our variant dataframe was subset to variants originating from the 'out-of-panel' HPRC samples (see above). Both reference genomes were divided into non-overlapping chunks of 10kb. The variants within each chunk were grouped together. For each 10kb region, the switch error rate was calculated as the total number of switch errors observed in within-region variants across all samples divided by the total number of heterozygous observed in within-region variants across all samples. Genotyping error was similarly calculated as the number of incorrect within-region genotypes divided by the number of called genotypes. A genotype was declared incorrect if either allele was called incorrectly. Finally, the average value of SER and genotyping error was calculated for a 500kb window (moving 10kb at a time). The resulting average error rates were assigned to the midpoint of the window. Variant density was calculated as the number of unique biallelic variants in each window divided by 500, yielding the number of variants per 1000 bp. All resulting data was plotted using karyoplotR[54].

Variant imputation and performance evaluation

Variant callsets from 279 participants in the Simons Genome Diversity Project were previously generated against both GRCh38 and T2T-CHM13 using the same GATK germline genotyping pipeline used to produce the 1KGP T2T-CHM13 variant calls[24]. Twenty-three samples are present in both the 1KGP dataset and the SGDP dataset; we dropped these samples from SGDP datasets, leaving 256 samples. We normalized indels and split multiallelic variants into biallelic variants using bcftools norm. To simulate genotyping array data, we used GRCh38 and T2T-CHM13-referenced vcfs of the variants called in the Illumina Omni 2.5 array to downsample the SGDP variant calls to those obtainable via array. There were 1,997,492 variants in the T2T-CHM13 intersection, and 1,984,898 variants in the GRCh38 intersection. We then followed best practices for imputing variation from genotyping arrays, removing any SGDP variants with more than a 5% missing allele rate or a Hardy-Weinberg Equilibrium value of less than 1e-4.

Common (>0.1% MAF) GRCh38 or T2T-CHM13 SGDP variant calls were pre-phased with SHAPEIT5_common. The resulting set of phased genotypes was used as a scaffold onto which rare variants (<0.1% MAF) were phased. The pre-phased SGDP 'array' variants were then imputed using IMPUTE5. We used an estimated population size of 135,000. This value produced the best phasing accuracy when phasing the 1KGP dataset, and the SGDP dataset also contains diverse haplotypes from around the world. GLIMPSE2_concordance[6] was used to evaluate the r2 accuracy of the imputed variant calls in various MAF bins. To determine the r2 accuracy of different classes of variant, we used a custom bash script to obtain separate lists of SNPs and indels stratified by minor allele frequency bin and/or presence in a syntenic/nonsyntenic region of the genome. These lists were used to create custom variant bins that were provided to GLIMPSE2_concordance.

References

1. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
2. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
3. Chen, N.-C. *et al.* Improved sequence mapping using a complete reference genome and lift-over. *Nat. Methods* **21**, 41–49 (2024).
4. Mun, T., Chen, N.-C. & Langmead, B. LevioSAM: fast lift-over of variant-aware reference alignments. *Bioinformatics* **37**, 4243–4245 (2021).
5. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
6. Rubinacci, S., Hofmeister, R. J., Sousa da Mota, B. & Delaneau, O. Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes. *Nat. Genet.* **55**, 1088–1090 (2023).
7. Das, S., Abecasis, G. R. & Browning, B. L. Genotype Imputation from Large Reference Panels. *Annu. Rev. Genomics Hum. Genet.* **19**, 73–96 (2018).
8. Roach, J. C. *et al.* Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science* **328**, 636–639 (2010).
9. Salter-Townshend, M. & Myers, S. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* **212**, 869–889 (2019).
10. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
11. Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
12. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
13. Chen, J. *et al.* A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. *Nat. Commun.* **7**, (2016).

14. Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic variation and environment. *Nat. Methods* **14**, 699–702 (2017).
15. Guo, M. H. *et al.* Inferring compound heterozygosity from large-scale exome sequencing data. *Nat. Genet.* **56**, 152–161 (2024).
16. Sales, R. R. *et al.* Fetal hemoglobin-boosting haplotypes of BCL11A gene and HBS1L-MYB intergenic region in the prediction of clinical and hematological outcomes in a cohort of children with sickle cell anemia. *J. Hum. Genet.* **67**, 701–709 (2022).
17. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
18. Genovese, G. *et al.* BCFtools/liftover: an accurate and comprehensive tool to convert genetic variants across genome assemblies. *Bioinformatics* **40**, btae038 (2024).
19. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
20. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
21. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun. 2018 91* **9**, 1–8 (2018).
22. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
23. Logsdon, G. A. *et al.* Complex genetic variation in nearly complete human genomes. *bioRxiv* 2024.09.24.614721 (2024) doi:10.1101/2024.09.24.614721.
24. Rhie, A. *et al.* The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
25. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
26. Kamm, J. A., Spence, J. P., Chan, J. & Song, Y. S. Two-Locus Likelihoods Under Variable Population Size and Fine-Scale Recombination Rate Estimation. *Genetics* **203**, 1381–1399 (2016).
27. Spence, J. P. & Song, Y. S. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Sci. Adv.* **5**, eaaw9206 (2019).

28. Gruhn, J. R., Rubio, C., Broman, K. W., Hunt, P. A. & Hassold, T. Cytological Studies of Human Meiosis: Sex-Specific Differences in Recombination Originate at, or Prior to, Establishment of Double-Strand Breaks. *PLoS ONE* **8**, e85075 (2013).
29. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
30. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
31. Browning, B. L. & Browning, S. R. Genotype error biases trio-based estimates of haplotype phase accuracy. *Am. J. Hum. Genet.* **109**, 1016–1025 (2022).
32. Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nat. Biotechnol. 2020 393* **39**, 302–308 (2020).
33. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
34. Olson, N. D. *et al.* Variant calling and benchmarking in an era of complete human genome sequences. *Nat. Rev. Genet.* **24**, 464–483 (2023).
35. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
36. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
37. Kalsner, L. & Chamberlain, S. J. Prader-Willi, Angelman, and 15q11-q13 duplication syndromes. *Pediatr. Clin. North Am.* **62**, 587–606 (2015).
38. Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature* **562**, 268–271 (2018).
39. Paschal, C. R. *et al.* Concordance of Whole-Genome Long-Read Sequencing with Standard Clinical Testing for Prader-Willi and Angelman Syndromes. *J. Mol. Diagn.* **0**, (2025).
40. McRae, J. jeremymcrae/liftover. (2025).

41. Lansdon, L. A. *et al.* Factors Affecting Migration to GRCh38 in Laboratories Performing Clinical Next-Generation Sequencing. *J. Mol. Diagn.* **23**, 651–657 (2021).
42. Logsdon, G. A. *et al.* The variation and evolution of complete human centromeres. *Nature* **629**, 136–145 (2024).
43. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
44. Mostovoy, Y. *et al.* Resolution of ring chromosomes, Robertsonian translocations, and complex structural variants from long-read sequencing and telomere-to-telomere assembly. *Am. J. Hum. Genet.* **111**, 2693–2706 (2024).
45. Gustafson, J. A. *et al.* High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.* **34**, 2061–2073 (2024).
46. Schloissnig, S. *et al.* Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. 2024.04.18.590093 Preprint at https://doi.org/10.1101/2024.04.18.590093 (2024).
47. Noyvert, B. *et al.* Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. 2023.12.20.23300308 Preprint at https://doi.org/10.1101/2023.12.20.23300308 (2025).
48. Jeong, H. *et al.* Structural polymorphism and diversity of human segmental duplications. *Nat. Genet.* **57**, 390–401 (2025).
49. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole-genomes. *Nat. Genet.* **49**, 303–309 (2017).
50. Mitchell A. Bekritsky, Camilla Colombo, & Michael A. Eberle. Identifying genomic regions with high-quality single nucleotide variant calling. *Illumina Genomics Res. Hub* (2021).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Releases · pangenome/vcfbub. *GitHub* https://github.com/pangenome/vcfbub/releases.
53. Polars. (2024).

54. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).

55. Starke, H. *et al.* Homologous sequences at human chromosome 9 bands p12 and q13-21.1 are involved in different patterns of pericentric rearrangements. *Eur. J. Hum. Genet.* **10**, 790–800 (2002).

56. Stankiewicz, P. *et al.* Recurrent deletions and reciprocal duplications of 10q11.21q11.23 including CHAT and SLC18A3 are likely mediated by complex low-copy repeats. *Hum. Mutat.* **33**, 165–179 (2012).

57. Phadke, S. R. & Sharda, S. A report of a patient with interstitial deletion of 15q22: Further delineation of a new micro deletion syndrome. *Am. J. Med. Genet. A.* **146A**, 1999–2000 (2008).

58. Borrow, J. *et al.* Diagnosis of acute promyelocytic leukaemia by RT-PCR: detection of PML-RARA and RARA-PML fusion transcripts. *Br. J. Haematol.* **82**, 529–540 (1992).

59. Verhelst, H. *et al.* Anti-NMDA-receptor encephalitis in a 3 year old patient with chromosome 6p21.32 microdeletion including the HLA cluster. *Eur. J. Paediatr. Neurol. EJPN Off. J. Eur. Paediatr. Neurol. Soc.* **15**, 163–166 (2011).

60. Ferlini, A. *et al.* Custom CGH array profiling of copy number variations (CNVs) on chromosome 6p21.32 (HLA locus) in patients with venous malformations associated with multiple sclerosis. *BMC Med. Genet.* **11**, 64 (2010).

61. Writzl, K. & Knegt, A. C. 6p21.3 microdeletion involving the SYNGAP1 gene in a patient with intellectual disability, seizures, and severe speech impairment. *Am. J. Med. Genet. A.* **161A**, 1682–1685 (2013).

62. Angulo, M. A., Butler, M. G. & Cataletto, M. E. Prader-Willi syndrome: a review of clinical, genetic, and endocrine findings. *J. Endocrinol. Invest.* **38**, 1249–1263 (2015).

63. Pizzo, L., Andrieux, J., Amor, D. J. & Girirajan, S. Clinical utility gene card for: 16p12.2 microdeletion. *Eur. J. Hum. Genet.* **25**, 271–271 (2017).

64. Niemi, A.-K., Kwan, A., Hudgins, L., Cherry, A. M. & Manning, M. A. Report of two patients and further characterization of interstitial 9p13 deletion--a rare but recurrent microdeletion syndrome? *Am. J. Med. Genet. A.* **158A**, 2328–2335 (2012).

65. Montenegro, M. M. *et al.* Expanding the Phenotype of 8p23.1 Deletion Syndrome: Eight New Cases Resembling the Clinical Spectrum of 22q11.2 Microdeletion. *J. Pediatr.* **252**, 56-60.e2 (2023).

66. Barber, J. C. K. *et al.* 8p23.1 duplication syndrome; a novel genomic condition with unexpected complexity revealed by array CGH. *Eur. J. Hum. Genet.* **16**, 18–27 (2008).

67. Zahir, F. *et al.* Novel deletions of 14q11.2 associated with developmental delay, cognitive impairment and similar minor anomalies in three children. *J. Med. Genet.* **44**, 556–561 (2007).

68. Bonati, M. T. *et al.* 9q34.3 microduplications lead to neurodevelopmental disorders through EHMT1 overexpression. *Neurogenetics* **20**, 145–154 (2019).

69. Morrow, B. E., McDonald-McGinn, D. M., Emanuel, B. S., Vermeesch, J. R. & Scambler, P. J. Molecular genetics of 22q11.2 deletion syndrome. *Am. J. Med. Genet. A.* **176**, 2070–2081 (2018).

70. Griggs, B. L. *et al.* Identification of ectodysplasin-A receptor gene deletion at 2q12.2 and a potential autosomal MR locus. *Eur. J. Hum. Genet. EJHG* **17**, 30–36 (2009).

71. Firth, H. V. *et al.* DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).

Synteny	Reference Genome	Variant type	SER (%)	Discordance rate (%)	# of Discordant variant calls	Total number of potentially	# of haplotype switches	# of heterozygous sites
All regions	CHM13v2.0	SNPs	0.149	0.257	1,581,451	615,986,668	125,429	83,970,616
		Indels	1.312	4.010	5,285,347	131,789,145	236,515	18,027,890
		SNPs + Indels	0.355	0.918	6,866,798	747,775,813	361,944	101,998,506
	GRCh38	SNPs	0.211	0.425	3,025,689	711,355,997	203,830	96,829,961
		Indels	1.385	4.440	6,480,948	145,982,877	270,802	19,555,065
		SNPs + Indels	0.408	1.109	9,506,637	857,338,874	474,632	116,385,026
Nonsyntenic	CHM13v2.0	SNPs	0.497	4.597	15,469	336,505	133	26,737
		Indels	2.730	9.586	16,586	173,023	410	15,019
		SNPs + Indels	1.300	6.291	32,055	509,528	543	41,756
	GRCh38	SNPs	0.366	2.371	25,015	1,054,999	450	122,930
		Indels	1.123	6.925	10,527	152,017	174	15,490
		SNPs + Indels	0.451	2.945	35,542	1,207,016	624	138,420
Syntenic	CHM13v2.0	SNPs	0.149	0.254	1,565,982	615,650,163	125,296	83,943,879
		Indels	1.311	4.003	5,268,761	131,616,122	236,105	18,012,871
		SNPs + Indels	0.354	0.915	6,834,743	747,266,285	361,401	101,956,750
	GRCh38	SNPs	0.210	0.422	3,000,674	710,300,998	203,380	96,707,031
		Indels	1.385	4.437	6,470,421	145,830,860	270,628	19,539,575
		SNPs + Indels	0.408	1.106	9,471,095	856,131,858	474,008	116,246,606

Table 1. Summary statistics of panel error rates. Switch errors and genotyping errors were determined with reference to HPRC assemblies.

	GRCh38 panel				T2T-CHM13 panel					
Genetic Region	**Num. of alt genotypes**	**Num. of het sites**	**GT error rate (%)**	**Switch error rate (%)**	**Num. of alt genotypes**	**Num. of het sites**	**GT error rate (%)**	**Switch error rate (%)**	**Relative reduction in SER**	**Relevant Medical Conditions**
9p12	95245	7490	16.49%	14.09%	25086	2624	5.09%	2.86%	79.71%	Site for pericentric inversions/SVs of unknown clinical significance[55]
10q11.22	607416	71272	5.54%	2.94%	337548	44594	4.11%	1.64%	44.31%	Developmental delay (case reports)[56]
10p12.32	31781	4911	0.84%	1.14%	29717	4562	0.54%	0.70%	38.49%	ADHD / ASD / Developmental Delay[38]
15q22.1	47813	6320	1.59%	2.23%	44693	6011	1.27%	1.40%	37.36%	Developmental delay (case reports[57]), site of PML 15;17 translocation[58]
6p21.32	928515	204345	1.91%	0.55%	779551	172245	0.73%	0.36%	35.64%	HLA Locus. Many autoimmune disorders[59,60], SYNGAP1 related disorders[61]
15q11.2	1035198	151851	3.54%	1.52%	743920	112070	0.78%	0.99%	34.74%	Start of Prader-Willi / Angelman imprinting region[62]
16p12.2	562123	71676	1.84%	2.45%	460982	63582	1.17%	1.60%	34.41%	16p12.2 deletion syndrome, variable penetrance[63]
15q13.1	604888	87552	1.35%	1.82%	478130	72705	0.79%	1.23%	32.26%	End of Prader-Willi / Angelman imprinting region[37,62]
9p13.1	325530	43786	2.31%	2.16%	251054	35628	0.90%	1.48%	31.54%	Case reports of developmental delay[64]
8p23.1	2149520	318288	1.40%	1.72%	1518594	228134	0.63%	1.21%	29.72%	8p23.1 deletion[65]/duplication[66] syndromes
14q11.2	1553026	236716	4.51%	2.21%	1226938	197103	1.05%	1.57%	28.92%	Developmental delay, CHD8-associated ASD[67]
9q34.12	169252	24221	1.01%	1.82%	154900	22313	0.62%	1.30%	28.70%	Kleefstra syndrome, 9q34 duplication syndrome[68]
22q11.21	1034410	146805	1.62%	1.78%	1191729	173381	1.40%	1.28%	28.12%	DiGeorge Syndrome[69]
15q22.33	57913	9739	0.83%	1.45%	55926	9458	0.58%	1.07%	26.24%	Developmental delay (case reports[57]), site of PML 15;17 translocation[58]
2q12.2	431312	65847	0.79%	1.45%	374370	58929	0.63%	1.07%	25.78%	Ectodermal dysplasias[70]

Table 2. Top 15 chromosomal regions with an improved switch error rate when using a T2T-CHM13 panel to phase short read variant calls. 1KGP short read variant calls from 39 samples with HPRC assemblies were phased using either a panel of GRCh38 reference haplotypes from unrelated individuals, or a T2T-CHM13 reference haplotypes from unrelated individuals. The parents of the 39 HPRC samples were removed from both panels prior to analysis, resulting in reference panels with haplotypes from 2426 individuals. Switch errors were calculated with phased pangenomic variation as a ground truth. Regions were sorted by the relative difference in SER when using the GRCh38 panel or the T2T-CHM13 panel. Regions associated with recurring copy number variant syndromes listed in the DECIPHER[71] database are in bold.