T2T-CHM13 improves read mapping and detection of clinically relevant genetic variation in the Swedish population

Daniel Schmitz[1], Adam Ameur[1] and Åsa Johansson[1]

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Box 815, 751 08 Uppsala, Sweden

Corresponding Authors: Daniel Schmitz <Daniel.schmitz@igp.uu.se> and Åsa Johansson <asa.johansson@igp.uu.se>. Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Box 815, 751 08 Uppsala, Sweden

Running title: T2T-CHM13 improves mapping and variant calling

Abstract

The T2T-CHM13 reference genome, released in March 2022, fills in the 8% of the human genome that were not resolved in GRCh38 and reconstructs large parts of the known genome. The more accurate and complete reference genome is expected to improve the quality of read mapping and variant calling. Even though whole genome sequencing (WGS)-based approaches have become the golden standard in medical genetics, the extent of these benefits remains unclear. In this study, we aim to evaluate alignment and variant call performance with T2T-CHM13 as a reference using a cross-sectional Swedish cohort (SweGen) comprising 1000 individuals with short-read Illumina WGS data available. Remapping and variant calling was performed using the nf-core/sarek pipeline. T2T-CHM13 improved a wide range of mapping and variant calling related metrics, including a higher fraction of properly paired reads, lower mismatch rate, and more uniform coverage of coding regions. Moreover, the fraction of ambiguous alignments was higher, reflecting segmental duplications that were incorrectly collapsed in GRCh37 and GRCh38. In comparison to GRCh38, we identified 10 million additional variants in the cohort, including 5.5 million singletons, and observed an increased sensitivity for rare variants. SnpEff assigned impact ratings of moderate or high to 13% more variants in T2T-CHM13 than GRCh38. In summary, we conclude that T2T-CHM13 improves alignment metrics with higher alignment quality, better variant calling performance and confidence, including for rare and deleterious variants. The T2T-CHM13 genome reference thus facilitates enhanced discovery of new disease-causing variation, benefiting, for example, rare-disease diagnostics.

Introduction

The first draft of the human reference genome was published by the Human Genome Sequencing Consortium in 2001 (International Human Genome Sequencing Consortium

2001). Since then, the reference genome has been continuously improved with patch releases, which add missing sequences, correct assembly errors, and augment the genome with alternative haplotypes. In 2022, the Telomere-to-Telomere (T2T) Consortium published the first gapless T2T assembly of a human genome (Nurk et al. 2022), and in 2023, an update which includes Chromosome Y (Rhie et al. 2023). This assembly fills in more than 200 Mbp of missing sequences from GRCh38, which corresponds to ca. 8% of the human genome. Among the regions that are now resolved are centromeres, the short arms of the acrocentric Chromosomes 13, 14, 15, 21 and 22 as well as the majority of Chromosome Y. In addition to resolving these unknown regions, they reported novel findings about the structure of the human genomes. The T2T Consortium discovered 3604 new genes and identified around 50 Mbp of segmental duplications (SDs), which make up half of all resolved gaps. Overall, SDs are a common genomic feature, representing 7% of the whole human genome, which are especially common on the short arms of acrocentric chromosomes (Vollger et al. 2022). Moreover, the T2T consortium benchmarked their assembly's performance for variant calling and found it to allow for the detection of many novel variants, especially in the previously unresolved regions. They also found that their novel reference reduces the number of false-positive calls mainly in protein-coding genes and they highlight the increased sensitivity for detection of rare variants and singletons (Aganezov et al. 2022). The T2T-CHM13 genome has already been seen use in the evaluation of novel tools (Formenti et al. 2022; Mc Cartney et al. 2022), characterization of SVs (Prodanov and Bansal 2022) and epigenetics (Gershman et al. 2022).

During recent years, whole-genome sequencing (WGS) has been introduced as a clinical diagnostic tool, especially for rare disease, which affect between 3.5% and 5.9% of the population (Nguengang Wakap et al. 2019) and are often caused by a single nucleotide variant (SNV) or a structural variant (SV). However, a minority of the rare disease patients

receive a definite diagnosis and often must go through a so-called “diagnostic odyssey” of repeated visits to specialists before a conclusive result can be obtained (Molster et al. 2016; Yan et al. 2020). A recent meta-analysis found WGS to significantly improve diagnostic yield over other established methods, including whole-exome sequencing (WES), but still estimated it to only be around 38.6% (Nurchis et al. 2023). This contributes to the increasing pool of evidence that variants in non-coding regions contribute to rare-disease development (Whiffin et al. 2020; Turro et al. 2020). Consequently, even with WGS as a first-line test, the fraction of patients receiving a diagnosis remains low. Possible explanations for this low yield could be insufficient annotation of non-coding variants, errors in the analysis pipeline, including mapping and variant calling, and misassemblies of the reference genome. These issues could be addressed by applying an improved reference genome but to which extent T2T-CHM13 improves detection of rare and deleterious variants over the established reference genomes GRCh37 and GRCh38 remains to be investigated.

SweGen is a cross-sectional cohort from Sweden providing a resource of genetic variability in the local population (Ameur et al. 2017). It consists of 1000 individuals who have undergone whole-genome sequencing (WGS). Since its inception in 2017, SweGen has facilitated clinical genomics research and diagnostics by providing a representative set of genetic variation in healthy individuals and, for example, by serving as a control cohort in SWEDEGENE, a study that has identified genetic variants causing serious adverse drug reactions (Hallberg et al. 2020). In another study, transposable elements (TEs) in protein-coding genes have been characterized in SweGen and used to diagnose two unsolved cases of rare diseases (Bilgrav Saether et al. 2023). SV frequency data from SweGen has been used to map complex chromosomal rearrangements (Eisfeldt et al. 2019). Several studies have used the frequency data from SweGen for identifying potentially pathogenic rare mutations usually not found in healthy individuals e.g., to identify germline mutations associated with breast

cancer risk (Helgadottir et al. 2021) and somatic variation in relapsed pediatric acute lymphoblastic leukemia (Sayyab et al. 2021).

In this study, we aim to evaluate the improvement of mapping short read WGS data to the new T2T-CHM13 reference compared to the reference genomes that are currently used in clinical decision tools. We remapped the WGS data from 1000 individuals from SweGen and compared the mapping, variant calling, and annotation performance as well as extrapolated from what impact a transition from the current reference to the T2T-CHM13 will have for clinical diagnostics. We also provide an open resource with updated summary data for variants in the SweGen cohort based on the T2T-CHM13 reference to be used for research and clinical applications.

Results

We obtained the original WGS alignments from SweGen, which were BAM files mapped to GRCh37 and contained all generated reads, including those that were not mapped. We reanalyzed the SweGen dataset with the freely available nf-core/sarek pipeline using BWA-MEM to T2T-CHM13 v2.0 (Garcia et al. 2020a; Md et al. 2019). We performed variant calling using GATK HaplotypeCaller in joint germline calling mode (Van der Auwera and O'Connor 2020). The same data had also been analyzed using GRCh38.p13 so we were able to leverage the results from that analysis for comparison with our results.

T2T-CHM13 improves quality of alignments

In comparison to GRCh37, the number of mapped reads increased by, on average, 558,955 reads per individual, which corresponded to 0.0692% (Figure 1A,B, Supplemental Table S1). On the other hand, fewer reads were mapped when comparing to GRCh38, with a mean difference of 128,094 per individual, corresponding to 0.0158% of all reads (Figure 1A,B, Supplemental Table S1). However, we observed that a larger portion of read pairs were

properly paired according to BWA-MEM when using T2T-CHM13 compared to the other two references (Figure 1C, Supplemental Table S1). BWA-MEM considers a read pair properly paired if it consists of two reads where one mate is mapped in forward and the other in reverse direction, both mates' read directions point towards each other and the distance between mates does not deviate too much from the mean insert size. In this case, this threshold is approximately six to seven standard deviations from the mean insert size. On average, 99.4% of reads per individual were paired properly (IQR: 99.4% – 99.5%) when using T2T-CHM13 (Figure 1C). In comparison, 97.9% of read pairs had proper orientation when using GRCh37 and 98.2% when using GRCh38. Furthermore, the alignments to T2T-CHM13 showed a lower per-sample mismatch rate (Figure 1D, Supplemental Table S1). The mean mismatch rate was 0.54% (IQR: 0.48% – 0.58%) for T2T-CHM13, compared to 0.78% (IQR: 0.73% – 0.82%) for GRCh37 and 0.65% (IQR: 0.60% – 0.69%) for GRCh38. This means that reads were mapped to T2T-CHM13 with higher accuracy.

Alignments previously considered unique are ambiguous

We noticed an increase in the number of alignments that received mapping quality 0 (MQ0) i.e., they mapped equally well to more than one position (Figure 1E) as well as a higher percentage of mapped reads with MQ0 (Figure 1F) with the T2T-CHM13 reference (Supplemental Table S1). MQ0 reads are also significantly overrepresented in SDs (χ^2 p-value $< 2.2 \times 10^{-16}$). Overall, 7.98% of all reads were assigned MQ0 but among those that map to SDs, 31.9% were assigned MQ0. This is likely a result of T2T-CHM13 resolving many SDs that appeared as unique regions in GRCh37 and GRCh38 as well as low-complexity regions not included in GRCh37 and GRCh38.

Coverage of genes is more uniform

To investigate the coverage of genes in our sequencing data and assess potential improvements to detection of functionally relevant variants, we calculated the read depth for all non-overlapping 500-bp bins in gene regions annotated by the T2T Consortium. While we did not observe a change in the mean coverage in genes, there was a reduction in the per-individual standard deviation of the read counts compared to GRCh37 and GRCh38, indicating a more uniform coverage of genes when aligning to T2T-CHM13 (Figure 2).

Reads from unplaced contigs map mostly to acrocentric short arms

We investigated the mapping positions of reads which were mapped to unplaced contigs in GRCh38. Virtually all reads (99.99974%) that uniquely mapped to unplaced contigs in GRCh38 could be mapped to canonical chromosomes in T2T-CHM13. The short arms of Chromosomes 13, 14, 15, 21, 22 and Y were overrepresented among the regions where these reads mapped, accounting for 84% of reads (Figure 3, Supplemental Table S2). However, we could only assign a unique canonical chromosome to five unplaced contigs out of 126. Reads from the remaining contigs mapped to positions on multiple chromosomes.

More rare variants with T2T-CHM13

To assess the general characteristics of the T2T-CHM13 variant calls, we collected statistics about number, frequencies, and types on the cohort level. Overall, we detected 47,744,487 unique high-quality variants in the cohort, of which 19,077,986 (40%) were singletons; an increase from GRCh37 and GRCh38 (Table 1, Figure 4A). This difference is not explained by the higher number of resolved bases, as the T2T-CHM13 call set shows a higher density of variants overall (15.3 variants/kbp) and singletons in particular (6.12 singletons/kbp) than the older call sets (Table 1). Moreover, we observed higher numbers of rare (allele frequency (AF) < 1%)) and low-frequency variants (AF < 5%) (Table 1, Figure 4C, Supplemental Table

S3). On the other hand, the number of common variants (AF ≥ 5%) remained roughly the same, and the number of non-reference alleles with an AF > 50% was generally lower with T2T-CHM13 than the other references (Figure 4C). SweGen was established as a resource to capture genetic variability within the Swedish cohort, and it has been shown that these samples cluster closely with other European populations, although they are distinct from central European populations (Ameur et al. 2017). Since T2T-CHM13 matches European ancestries more closely than GRCh37 and GRCh38 (Aganezov et al. 2022), we would expect to observe fewer variants where the non-reference allele is the major allele in our cohort when T2T-CHM13 is used as the reference. Noticeable is also the peak of variants with AF ≈ 0.5. with GRCh37 and GRCh38 (Figure 4C) which was not seen with T2T-CHM13. When removing variants that deviated strongly from Hardy-Weinberg equilibrium ($P < 10^{-20}$), this peak disappeared (Supplemental Figure S1). This points to it being caused by previously unresolved segmental duplications, that give rise to stretches of heterozygous variant calls in GRCh37 and GRCh38.

Proportions of SNVs and indels remain unchanged

To assess whether T2T-CHM13 affected our power to call SNVs or indels, we compared the numbers and proportions of these variant types between references and in previously unresolved regions. Around 87.8% of all called variants were SNVs, which is similar to what was found for GRCh37 (86.7%) and GRCh38 (87.3%) (Figure 4A). The overall fraction of indels remained roughly the same as well, constituting 13.9% of variants in T2T-CHM13, 14.3% in GRCh37 and 13.7% in GRCh38. In general, low-frequency variants skewed more towards indels and mixed variation (i.e. indels and SNVs at the same site) than common variants in T2T-CHM13 (Figure 4D). Indels were generally skewed towards deletions in all assemblies. Overall, more bases were deleted than inserted, resulting in an excess of deleted bases of 11,028,492 bases in T2T-CHM13, which was an increase from GRCh37 (1,065,087)

and GRCh38 (816,639) (Supplemental Figure S2). This excess was also visible in the number of 1-bp deletions, which accounted for 1,425,614 variants. 729,804 of these deletions were situated in genes, independent of predicted impact. 1-bp deletions were more likely to be singletons than variants in general; 43.4% (619,334) were private to one individual, opposed to 40% overall; but less likely to be fixed in the cohort (0.004% (170) vs 0.036% (16967) overall).

Higher numbers of loss-of-function variants and variants with predicted functional impact with T2T-CHM13

To confirm our previous assessment of improved detection of variants with functional impact thanks to better gene coverage, we predicted variant effects using SnpEff, including loss of function (LoF). We observed an increased number of variants with an impact rating, i.e., a prediction of having impact on the protein function by, for example, causing truncation, loss of function or triggering nonsense mediated decay (Figure 4B). Overall, using T2T-CHM13 we identified 474,050 variants with an impact rating of “low”, “moderate” or “high”, corresponding to approximately 0.993% of all called variants. For comparison, 354,087 variants called with GRCh37 (0.996%) and 359,977 called with GRCh38 (0.949%) received an impact rating. Among the variants with an impact rating with T2T-CHM13, 48.4% and 45.2% of variants with were predicted to have moderate and low impact respectively and the remaining 6.4% of were rated high impact (Figure 4B). Our results thus suggest that improved gene coverage facilitates detection of functionally relevant variation.

We identified 17,446 LoF variants using T2T-CHM13, an increase from GRCh37 and GRCh38 (Table 1). Of these, 17,297 variants were annotated to 8,287 genes with existing ClinVar annotations. Most affected genes were shared among all three references (Figure 5). Across all three references, *MUC4*, which encodes mucin 4, was the gene with the highest

number of LoF variants (Table 2, Supplemental Table S4). However, the majority of LoF variants in *MUC4* fall into a 48-bp variable tandem repeat in exon 2, which is problematic for short reads and might have led to an inflated number of variants. Tandem repeats are a general characteristic of the mucin gene family, which made up five of the ten top affected genes across all assemblies (Ferez-Vilar and Hill 1999). Additionally, the *HLA-DRB1*, belonging to the human leukocyte antigen (HLA) gene family, was present across all references. Another *HLA* gene, *HLA-DRB5*, was among the top ten for GRCh37 and GRCh38 but not T2T-CHM13. However, as with the mucin gene family, *HLA* genes tend to be difficult to analyze with short reads since they are highly polymorphic.

With these considerations in mind, we curated a random list of ten novel LoF variants which were called in genes which had no such variants in GRCh37 and GRCh38. We could confirm the existence of six but not the other four (Supplemental Figure S3). Previous studies have suggested that an increased number of high-impact predictions are likely to represent artifacts arising from lower quality or incomplete assemblies (Cooper and Shendure 2011). It is therefore plausible that some of the novel LoF variants we did not curate are artifacts rather than true LoF variants, particularly given the improved quality and completeness of the T2T-CHM13 assembly. Moreover, we validated these LoF variants in two individuals (SweGen0945 and SweGen0970) with available *de-novo* assemblies from 75× PacBio continuous long-read (CLR) sequencing, where we could confirm 15 of 130 (11.5%) and 16 of 139 (11.5%) variants, respectively (Supplemental Table S5a,b). For a third individual (SweGen0969), we could make use of available 35× WGS PacBio CLR data and confirm 21 of 143 variants (14.7%) (Supplemental Table S5c).

Among the top ten affected genes in T2T-CHM13, only four did not belong to the aforementioned gene families. Melanoma antigen gene C1 (*MAGEC1*) has close links to

melanoma and myeloma development (Caballero et al. 2010; Jungbluth et al. 2005). AHNAK Nucleoprotein 2 (AHNAK2) may play a role in calcium signalling and is up regulated in lung cancer (Liu et al. 2020; Komuro et al. 2004). Nascent polypeptide associated complex (NACA) binds to newly created polypeptides to prevent wrong translocation to the endoplasmatic reticulum and plays a role in bone-formation, red blood cell differentiation and has been linked to dermatitis (Lauring et al. 1995; Natter et al. 1998; Lopez et al. 2005). Immunoglobulin-Like And Fibronectin Type III Domain-Containing Protein 1 (IGFN1) may play a role in skeletal muscle development and has been associated with polypoidal choroidal vasculopathy (Wen et al. 2018; Cracknell et al. 2020).

Previously unresolved regions are not enriched for singletons

To assess whether the observed increase in singleton calls was driven by regions that were unresolved in GRCh37 and GRCh38, we collected general variant call statistics restricted to these regions. There were 8,321,170 variants in regions unresolved in GRCh37 and in regions which were unresolved in GRCh38, we identified 7,826,115 variants (Table 3). Within these regions, we observed a higher density of variants (unresolved in GRCh37: 30.9 variants / kbp, unresolved in GRCh38: 31.1 variants / kbp) than in the remaining genome (13.9 variants / kbp). By that, we can conclude that these variants contribute to a majority, but not all, of the excess in the number of variants called with T2T-CHM13. This was also apparent in unique regions syntenic between GRCh38 and T2T-CHM13, where T2T-CHM13 showed a higher variant density (13.9 variants / kbp) than GRCh38 (12.5 variants / kbp).

Variants in regions that were previously unresolved in GRCh37 and GRCh38 were more biased towards SNVs (which made up 92.6% and 92.8% of variants there, respectively) than other regions (Figure 6A). However, the density of deletions was higher in these regions (unresolved in GRCh37: 2.45 deletions / kbp, unresolved in GRCh38: 2.46 deletions / kbp)

than in the rest of the genome (1.65 deletions / kbp). We detected an excess of 4,705,752 deleted bases in regions which were unresolved in GRCh38, (4,947,747 bp in regions not in GRCh37) despite only contributing 13% of all indels. The excess in the regions syntenic with GRCh38 is still higher, however, with 6,322,740 bp more bases deleted than inserted.

The fraction of singletons was lower than in the genome in general (χ^2 p value < 2.2×10^{-16}) (Table 3), suggesting that the newly resolved regions do not contribute disproportionately to the number of singletons despite a higher density of singletons (Table 3). However, they were significantly enriched for rare variants (χ^2 p value < 2.2×10^{-16}) and contributed the majority of the excess of rare alleles that were not observed in GRCh37 (61.1% of 12,594,438) or in GRCh38 (68.6% of 10,661,072).

Previously unresolved regions contain a low rate of variants with predicted impact on a protein

Focusing on previously unresolved regions, we found that they had a much lower fraction of variants with predicted impact on a protein according to SnpEff. The fractions of variants with at least low impact were significantly lower than those of all variants called with T2T-CHM13. In the regions unresolved in GRCh37, there were 3,823 variants with a rating of at least low impact, corresponding to 0.045% (χ^2 P-value < 2.2×10^{-16}). In the parts of the genome, which were unresolved in GRCh38, there were 1,922 variants with assigned impact rating, which corresponded to 0.025% (χ^2 P-value < 2.2×10^{-16}). The distribution of impact ratings among these variants also differed slightly between regions that were unresolved in GRCh37 compared to GRCh38, with a larger fraction of high-impact variants in regions that were unresolved in GRCh37 (Figure 6B). Together, this suggests that variations in the previously unresolved regions generally have lower impact and might therefore be under purifying selective pressure. However, this difference could also be due to a lower density of

genes or a lack of annotations compared to regions that have been available to functional annotation before.

Proportion of SNVs of per-individual variants decreases with use of T2T-CHM13

Considering the differences we observed on the cohort level between GRCh37, GRCh38 and T2T-CHM13, we also investigated how variant calls differed at the individual level. Despite the higher total number of called variants in the cohort with T2T-CHM13, we did not observe this increase in the number of variants called per individual (Figure 7A, Table 4). After limiting the comparison to syntenic regions with high uniqueness in GRCh38 and T2T-CHM13, we observed a slight decrease in the number of variants per individual (mean difference = -212,696.8, p value < 2.2×10^{-16}) and their density (mean difference = -0.039 variants / kbp, p value < 2.2×10^{-16}) (Supplemental Table S6).

We also assessed whether the proportions of SNVs and indels changed when calling variants with T2T-CHM13. The number of SNVs per individual decreased and the number of indels increased slightly. (Figure 7B, Table 4). These changes also became apparent in the fraction of SNVs and indels in all called variants per individual (Figure 7C). The mean fraction of SNVs decreased whereas we observed an increased proportion of indels. This discrepancy can be explained by the previously described observation that the fraction of indels tends to be higher among rare variants and that we detected fewer high frequency but a larger number of rare variants. The called indels also tended to skew more towards deletions than the rest of the genome.

Increased power to detect rare and loss-of-function variants per individual

The number of singletons called per individual was higher with T2T-CHM13 (Table 4, Figure 7D). There were four individuals with singleton counts of more than three standard

deviations above the mean, which we considered outliers. These four individuals were different from the four individuals with the largest number of singletons with GRCh38. However, they made up four of the six individuals with the lowest fraction of properly paired reads and showed anomalies in certain other QC metrics, such as the number of read pairs on different chromosomes. Considering that this was consistent across references, we believe these high numbers of singletons to be caused by technical issues with these four samples, which had not been identified before (Supplemental Table S7). While these individuals accounted for approximately 645,000 singletons, they did not drive the observed increase in per-individual singletons, as every individual in the cohort had an increased number of singleton variants (Supplemental Figure S4). Furthermore, we observed a slightly increased number of variants with predicted LoF effects per individual over GRCh38 (mean difference = 4.778, p value = 0.000722), which agreed with the higher number in the cohort as a whole and showed the benefit of improved sequencing coverage in gene regions (Figure 7E, Table 4). However, in comparison to GRCh37, the per-individual number of LoF variants actually decreased (mean difference = -174.429, p value < 2.2×10^{-16}). This excess can be explained by variants whose reference alleles were the LoF alleles, which were corrected in GRCh38 and appeared as common or fixed variants in SweGen. For instance, there were 194 LoF variants in the GRCh37 call set with AF > 90%, as compared to 83 in GRCh38 and 24 in T2T-CHM13.

Discussion

In this study, we have shown that using the new T2T-CHM13 reference can improve alignments, variant calls, and sensitivity for rare, singleton and LoF variants over GRCh37 and GRCh38, which are currently in widespread use, including in genes with clinical relevance. This is an important prospect for clinical genetics as rare, ultra-rare and *de-novo*

variants are often the cause of rare disease. Continuing efforts are aiming to bring sequencing, and especially WGS, into the diagnostic routine because of its power to detect these variants (Tesi et al. 2023). For instance, the UK 100,000 Genomes Project found that genome sequencing significantly increases the diagnostic yield of rare diseases (Smedley et al. 2021). The Genomics England project is finding an increasing number of rare variants with clinical significance and enabling diagnoses of previously undiagnosed patients (Joyce et al. 2022; Han et al. 2022; Kojic et al. 2023; Vadgama et al. 2022). By employing T2T-CHM13 as a reference for first-line WGS in the clinic, the number of people receiving a diagnosis for a rare disease is likely to improve further. Moreover, it is becoming increasingly clear that reanalysis of sequencing data can lead to better diagnoses and improved patient outcomes (Machini et al. 2019; Kingsmore et al. 2022). Through reanalysis, it is possible to detect new causal variations using improved tools and annotation and provide treatment to affected individuals. Considering this, *FixItFelix*, which realigns reads falling into medically relevant genes using a modified GRCh38 reference with certain genes replaced by their T2T-CHM13 assemblies, was developed (Behera et al. 2023). This tool only requires reanalysis of smaller regions and, as such, provides a faster turnaround. However, it requires limiting the scope of the analysis to manually curated regions and, therefore, is not suitable for a whole-genome study. As such, a re-evaluation of inconclusive clinical samples using T2T-CHM13, including reads that failed to map to canonical chromosomes, can provide valuable diagnoses to patients.

Rare variants are also becoming more prominent in genetic epidemiology. GWAS have only been able to explain part of the observed heritability. Recent studies showed that part of this missing heritability can be explained by rare variants and successfully quantified their effect on complex traits and disease risk (Höglund et al. 2019; Kierczak et al. 2022; Wang et al. 2021; The UK10K Consortium 2015). Consequently, the possibility of detecting more rare

variants, especially in regions which lacked an assembly but are functionally relevant, such as centromeres, promises to close the gap in quantifying the missing heritability.

The resources provided by the T2T Consortium, such as lifted over versions of dbSNP, allow for the adoption of existing workflows to the new reference as well as alignments and variant calls generated using it. However, there is currently a lack of original annotation and databases based on T2T-CHM13. Reliance on annotations generated using older references and lifted over to T2T-CHM13 might negatively impact results, as not all variants can be lifted over. We expect this issue to diminish as more resources, such as this one, become available. Another limitation was that we were unable to unambiguously elucidate why four individuals were outliers with regards to number of singletons. As these four individuals showed some anomalies regarding some QC measures, a possible explanation might have been a higher sensitivity to sample quality or technical issues when aligning to T2T-CHM13. However, we could not conclusively confirm this hypothesis at this point. Furthermore, as the version of T2T-CHM13 we used included unmasked pseudoautosomal regions of Chromosome Y, variants called in these regions suffer from lower accuracy. However, there were only 415 variants with a predicted functional impact, so these issues did not significantly impact our results (Supplemental Table S8). In addition, we did not restrict our analyses to short-read accessible regions. Even though variants outside these regions have a higher risk of being false positives, general trends regarding variant impacts and types within them, agree with our findings from the whole genome (Supplemental Table S9, Supplemental Figure S6).

We observed, in general, improvements in mapping and variant calling performance similar to those presented by the T2T Consortium in their investigation (Aganezov et al. 2022). These included an increase in quality of alignments as both the fraction of properly paired

and oriented read pairs increased as well as the mismatch rate decreased. The newly assembled regions also allowed us to identify reads that appeared to map uniquely in older references, but that actually map to duplicated regions which were previously unknown. This allowed for greater confidence in the variant calls since spurious variants, which were called as common alleles but originated from separate copies of the same SD, could be removed. Furthermore, common variants might have disappeared because of altered alleles in T2T-CHM13.

On the other hand, we identified generally more variants overall, in terms of absolute number and density, including more deletions than with previous references, which might be a sign of reference bias associated with the increased number of duplicated sequences in T2T-CHM13. This difference was also present in syntenic regions between GRCh38 and T2T-CHM13 with high uniqueness, where we expected calls to be largely consistent between references. However, at the individual level, we observed a slight decrease in variant number and density in these regions. Therefore, we argue that the majority of variants are called consistently and that the observed opposite trends of these metrics between cohort and individual level are due to minor differences in the syntenic regions and the changes to variant composition mentioned above.

The markedly elevated number of LoF variants in specific genes, many of which are implicated in disease-related pathways, suggests that variation in these genes have been systematically underrepresented in previous reference assemblies. This finding underscores the importance of T2T-based mapping, particularly for diseases where accurate variant detection in these pathways is critical for understanding pathogenesis and improving diagnostic resolution. However, some additional LoF variants we identified, including in syntenic regions, were technical artifacts. However, through our curation of LoF variants in

T2T-CHM13 in genes without any detected LoF variation in GRCh37 and GRCh38, and additional validation using PacBio CLR data, which, despite its higher error rate, allowed us to confirm up to 14.7% of novel LoF variants, we could show that there was, in fact, a higher sensitivity for functionally relevant variation. As such, the balance between detection of novel LoF variants and technical artifact is an important consideration for future projects using T2T-CHM13, especially considering that, as before, LoF variants should be confirmed experimentally before they can be used to inform clinical decisions.

In conclusion, the T2T-CHM13 reference provides a valuable resource that gives improved alignments and variant calling. Based on our results we recommend, going forward, the adoption of T2T-CHM13 for novel sequencing studies. Re-evaluation of sequenced samples, especially inconclusive ones, using T2T-CHM13 can provide further insights into genetic variation, including high-impact events, that might otherwise be missed. Considering the computational complexity of remapping and recalling large cohorts, however, it can be sensible to take an iterative approach. But as more resources based on T2T-CHM13 become available, we can expect increased power and precision in statistical and clinical genetics applications.

Methods

The SweGen Cohort

The SweGen cohort is a cross-sectional cohort of the Swedish population consisting of 1000 individuals that all passed quality control as described in the original SweGen paper (Ameur et al. 2017). The majority of SweGen comprises 942 unrelated individuals from the Swedish Twin Register (STR) which were selected to mirror the overall distribution of genetic principal components (PCs) of all 10,000 STR individuals that underwent SNV genotyping (Lichtenstein et al. 2002). The remaining 58 individuals were selected from the Northern

Swedish Population Health Study (NSPHS) (Igl et al. 2010). NSPHS individuals were selected based on their genetic PCs, as well. Library preparation and WGS were responsibility of Sweden's National Genomics Infrastructure (NGI) in Stockholm (NGI-S) and Uppsala (NGI-U). After fragmenting to 350 bp inserts and library preparation, DNA samples underwent WGS on the Illumina HiSeq X platform using v2.5 chemistry (Ameur et al. 2017). The 1000 individuals included in the dataset all passed QC. The original mappings were obtained by aligning the raw reads to GRCh37.p13 using BWA-MEM 0.7.12. The mappings underwent quality control, which was described previously, including base quality recalibration and duplication marking (Ameur et al. 2017). In addition, alignments and variant calls based on GRCh38.p14 including alternative haplotypes and decoys were available. The final GRCh37 BAM files served as the input for our analysis pipeline as they contained all reads, including unmapped ones, and allowed us to take advantage of the already performed quality control.

Remapping and Variant Calling

We obtained the T2T-CHM13 reference genome version 2.0, a lifted over version of dbSNP and a BED file of segmental duplications from the T2T Consortium's GitHub page (https://github.com/marbl/CHM13). This version of the reference genome contains an assembly of Chromosome Y and softmasked repeats. We performed mapping and variant calling using nf-core/sarek version 3.2.0, which uses GATK 4.3.0.0, with default parameters corresponding to GATK best practices (Garcia et al. 2020b; DI Tommaso et al. 2017; Ewels et al. 2020; McKenna et al. 2010). Shortly, the reads contained in the delivered BAM files, i.e., the post-QC alignments to GRCh37p13, which contained all generated reads, even those that failed original QC or were not aligned, underwent read QC using FastQC 0.11.9 and fastp 0.23.2 (Chen et al. 2018; Anders 2010). Reads failed QC if more than 40% of bases had a quality score below 15, or more than five base calls were indeterminate, i.e., they were

called as “N”. On average, 4.1% of reads (inter-quartile range (IQR): 3.6% – 4.5%) failed this filter. The remaining reads were aligned using BWA-MEM version 0.7.17 (Md et al. 2019). The resulting CRAM files were merged, sorted, and indexed using SAMtools 1.16.1 (Li et al. 2009).

Variants were called using GATK HaplotypeCaller with the joint germline calling feature enabled. For Variant QC, GATK VariantRecalibrator was used with lifted over versions of dbSNPv155, 1000 Genomes Omni 2.5 variants, as well as 1000 Genomes phase I and gold standard indels. We included all variants passing the more lenient default threshold (truth sensitivity 99.9% – 100%) for increased sensitivity and because those were available in the call sets of the older assemblies. General statistics on the performance of the T2T-CHM13 reference were based on the reports automatically generated by SAMtools and BCFtools 1.17 as part of the sarek pipeline (Danecek et al. 2021; Li 2011). We added annotations using BCFtools and as well as functional effect predictions, such as loss of function (LoF) using SnpEff based on the lifted over version of dbSNP and curated transcripts from RefSeq 5.1 provided by the T2T Consortium (Cingolani et al. 2012a, 2012b). We created a subset of putatively clinically relevant LoF variants by filtering for annotation to genes in ClinVar (version 20240215) (Landrum et al. 2014). We identified variants in previously unresolved regions, both in relation to GRCh37 and GRCh38, by subsetting the joint VCF with BCFtools using the “CHM13 unique” track in UCSC Genome Browser (Kent et al. 2002). We obtained already existing alignments to GRCh37 and GRCh38 from the SweGen dataset as well as variant calls, both per-individual and joint calls for the whole cohort, generated from these alignments. We generated mapping and variant reports using SAMtools and BCFtools. We calculated read depth in gene regions by creating non-overlapping 500-bp bins covering these regions. Additionally, we created a subset of variants expected to be consistent between GRch38 and T2T-CHM13 by subsetting for syntenic regions and the “UMAP S100” track

from UCSC Genome Browser, as these regions are consistent between assemblies and allow for high-confidence short-read mapping (Karimzadeh et al. 2018).

Validation of putative LoF variants

We visually inspected ten randomly selected predicted LoF variants from genes that had no reported LoF variants in the GRCh37 and GRCh38 call sets in IGV 2.8.13. We validated putative LoF variants in the same set of genes in individuals SweGen0945 and SweGen0970 using available *de-novo* assemblies generated from 75× coverage PacBio continuous long read (CLR) sequencing (Ameur et al. 2018). We mapped the assembled contigs to T2T-CHM13 using minimap2 2.26 and called variants using the included version of paftools (Li 2018). We compared the variants found in the *de-novo* assemblies with those in the short reads using RTG Tools vcfeval 3.12.1, restricting the analysis to regions deemed reliable for variant calling by paftools i.e., covered by exactly one contig, excluding alignments shorter than 10 kbp (Cleary et al. 2015).

We validated putative LoF variants in individual SweGen0969 using previously generated PacBio CLR sequencing data (Schmitz et al. 2023). We aligned the reads to T2T-CHM13 using minimap2 version 2.26 and called variants using Longshot 1.0 (Edge and Bansal 2019). We compared the variants found in the PacBio data with those in the short reads using RTG Tools vcfeval 3.12.1, restricting the analysis to regions with coverage higher than 15× and lower than 60× to exclude copy-number variable regions.

Ethics Declaration

All participants in the Swedish Twin Register gave informed consent and the study was approved by the regional ethics committee in Stockholm (Regionala Etikprövningsnämnden, Stockholm, dnr 2007-644-31 and 2014/521-32). Participants of the Northern Swedish Population Health Study gave their informed consent and the study was approved by the

regional ethics committee in Uppsala (Regionala Etikprövningsnämnden, Uppsala, dnr 2005:325 and 2016-03-09).

Data Access

Cohort-level allele frequencies and variant annotation are publicly available on the website of SweGen (https://swefreq.nbis.se/dataset/SweGen) and Zenodo (https://doi.org/10.5281/zenodo.13739348). All data generated in this study have been deposited in FEGA Sweden and made findable through the European Genome-Phenome Archive web portal (EGA, https://ega-archive.org) under study number EGAS50000000906 (https://ega-archive.org/studies/EGAS50000000906).

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

AA and DS conceived the study. DS performed the analyses, interpreted the results, wrote the initial manuscript and generated figures and tables under supervision of ÅJ and AA. All authors discussed and interpreted the results and contributed to the final manuscript.

This project was funded by the Swedish Research Council (Dnr. 2019-01497). The computations were enabled by resources in project sens2016003 and sens2022031 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al. 2022. A complete reference genome improves analysis of human genetic variation. *Science (1979)* **376**. doi:10.1126/SCIENCE.ABL3533.

Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, Höijer I, Häggqvist S, Vezzi F, Nordlund J, Olason P, et al. 2018. De Novo Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes 2018, Vol 9, Page 486* **9**: 486. doi:10.3390/GENES9100486.

Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, Viklund J, Kähäri AK, Lundin P, Che H, et al. 2017. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics 2017 25:11* **25**: 1253–1260. doi:10.1038/ejhg.2017.130.

Anders S. 2010. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. *Soil* **5**: http://www.bioinformatics.babraham.ac.uk/projects/.

Behera S, LeFaive J, Orchard P, Mahmoud M, Paulin LF, Farek J, Soto DC, Parker SCJ, Smith A V, Dennis MY, et al. 2023. FixItFelix: improving genomic analysis by fixing reference errors. *Genome Biol* **24**: 31. doi:10.1186/s13059-023-02863-7.

Bilgrav Saether K, Nilsson D, Thonberg H, Tham E, Ameur A, Eisfeldt J, Lindstrand A. 2023. Transposable element insertions in 1000 Swedish individuals. *PLoS One* **18**: e0289346. doi:10.1371/JOURNAL.PONE.0289346.

Caballero OL, Zhao Q, Rimoldi D, Stevenson BJ, Svobodová S, Devalle S, Róhrig UF, Pagotto A, Michielin O, Speiser D, et al. 2010. Frequent MAGE Mutations in Human Melanoma. *PLoS One* **5**: e12773. doi:10.1371/JOURNAL.PONE.0012773.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890. doi:10.1093/BIOINFORMATICS/BTY560.

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. 2012a. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* **3**: 35. doi:10.3389/FGENE.2012.00035/BIBTEX.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012b. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**: 80–92. doi:10.4161/fly.19695.

Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, Jackson A, Littin R, Rathod M, Ware D, et al. 2015. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv* 023754. doi:10.1101/023754.

Cooper GM, Shendure J. 2011. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* **12**: 628–640. doi:10.1038/nrg3046.

Cracknell T, Mannsverk S, Nichols A, Dowle A, Blanco G. 2020. Proteomic resolution of IGFN1 complexes reveals a functional interaction with the actin nucleating protein COBL. *Exp Cell Res* **395**: 112179. doi:10.1016/J.YEXCR.2020.112179.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* **10**: 1–4. doi:10.1093/GIGASCIENCE/GIAB008.

DI Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**: 316–319. doi:10.1038/NBT.3820.

Edge P, Bansal V. 2019. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nature Communications 2019 10:1* **10**: 1–10. doi:10.1038/s41467-019-12493-y.

Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Käller M, Gruselius J, Nilsson D, Syk Lundberg E, Carvalho CMB, Lindstrand A. 2019. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. *PLoS Genet* **15**: e1007858. doi:10.1371/JOURNAL.PGEN.1007858.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology 2020 38:3* **38**: 276–278. doi:10.1038/s41587-020-0439-x.

Ferez-Vilar J, Hill RL. 1999. The structure and assembly of secreted mucins. *Journal of Biological Chemistry* **274**: 31751–31754. doi:10.1074/jbc.274.45.31751.

Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, Myers EW, Jarvis ED, Phillippy AM. 2022. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nature Methods 2022 19:6* **19**: 696–704. doi:10.1038/s41592-022-01445-y.

Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, DiLorenzo S, Sandgren J, Díaz De Ståhl T, Ewels P, et al. 2020a. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research 2020 9:63* **9**: 63. doi:10.12688/f1000research.16665.2.

Garcia M, Juhos S, Larsson M, Olason PI, Martin M, Eisfeldt J, DiLorenzo S, Sandgren J, Díaz De Ståhl T, Ewels P, et al. 2020b. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research 2020 9:63* **9**: 63. doi:10.12688/f1000research.16665.2.

Gershman A, Sauria MEG, Guitart X, Vollger MR, Hook PW, Hoyt SJ, Jain M, Shumate A, Razaghi R, Koren S, et al. 2022. Epigenetic patterns in a complete human genome. *Science (1979)* **376**. doi:10.1126/SCIENCE.ABJ5089/SUPPL_FILE/SCIENCE.ABJ5089_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.

Hallberg P, Yue QY, Eliasson E, Melhus H, Ås J, Wadelius M. 2020. SWEDEGENE—a Swedish nation-wide DNA sample collection for pharmacogenomic studies of serious adverse drug reactions. *The Pharmacogenomics Journal 2020 20:4* **20**: 579–585. doi:10.1038/s41397-020-0148-3.

Han JH, Ryan G, Guy A, Liu L, Quinodoz M, Helbling I, Lai-Cheong JE, Barwell J, Folcher M, McGrath JA, et al. 2022. Mutations in the ribosome biogenesis factor gene LTV1 are linked to LIPHAK syndrome, a novel poikiloderma-like disorder. *Hum Mol Genet* **31**: 1970–1978. doi:10.1093/HMG/DDAB368.

Helgadottir HT, Thutkawkorapin J, Lagerstedt-Robinson K, Lindblom A. 2021. Sequencing for germline mutations in Swedish breast cancer families reveals novel breast cancer risk genes. *Scientific Reports 2021 11:1* **11**: 1–7. doi:10.1038/s41598-021-94316-z.

Höglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, Johansson Å. 2019. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific Reports 2019 9:1* **9**: 1–14. doi:10.1038/s41598-019-53111-7.

Igl W, Johansson A, Gyllensten U. 2010. The Northern Swedish Population Health Study (NSPHS)--a paradigmatic study in a rural population combining community health and basic research. *Rural Remote Health* **10**: 1363. doi:10.22605/RRH1363.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921. doi:10.1038/35057062.

Joyce KE, Onabanjo E, Brownlow S, Nur F, Olupona K, Fakayode K, Sroya M, Thomas GA, Ferguson T, Redhead J, et al. 2022. Whole genome sequences discriminate hereditary hemorrhagic telangiectasia phenotypes by non-HHT deleterious DNA variation. *Blood Adv* **6**: 3956–3969. doi:10.1182/BLOODADVANCES.2022007136.

Jungbluth AA, Ely S, DiLiberto M, Niesvizky R, Williamson B, Frosina D, Chen YT, Bhardwaj N, Chen-Kiang S, Old LJ, et al. 2005. The cancer-testis antigens CT7 (MAGE-C1) and MAGE-A3/6 are commonly expressed in multiple myeloma and correlate with plasma-cell proliferation. *Blood* **106**: 167–174. doi:10.1182/BLOOD-2004-12-4931.

Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. 2018. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* **46**: e120–e120. doi:10.1093/NAR/GKY677.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006. doi:10.1101/gr.229102.

Kierczak M, Rafati N, Höglund J, Gourlé H, Lo Faro V, Schmitz D, Ek WE, Gyllensten U, Enroth S, Ekman D, et al. 2022. Contribution of rare whole-genome sequencing variants to plasma protein levels and the missing heritability. *Nat Commun* **13**: 2532. doi:10.1038/s41467-022-30208-8.

Kingsmore SF, Smith LD, Kunard CM, Bainbridge M, Batalov S, Benson W, Blincow E, Caylor S, Chambers C, Del Angel G, et al. 2022. A genome sequencing system for universal newborn screening, diagnosis, and precision medicine for severe genetic diseases. *Am J Hum Genet* **109**: 1605–1619. doi:10.1016/j.ajhg.2022.08.003.

Kojic M, Abbassi NEH, Lin TY, Jones A, Wakeling EL, Clement E, Nakou V, Singleton M, Dobosz D, Kaliakatsos M, et al. 2023. A novel ELP1 mutation impairs the function of the Elongator complex and causes a severe neurodevelopmental phenotype. *Journal of Human Genetics 2023 68:7* **68**: 445–453. doi:10.1038/s10038-023-01135-3.

Komuro A, Masuda Y, Kobayashi K, Babbitt R, Gunel M, Flavell RA, Marchesi VT. 2004. The AHNAKs are a class of giant propeller-like proteins that associate with calcium channel proteins of cardiomyocytes and other cells. *Proceedings of the National Academy of Sciences* **101**: 4053–4058. doi:10.1073/PNAS.0308619101.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980–D985. doi:10.1093/nar/gkt1113.

Lauring B, Sakai H, Kreibich G, Wiedmann M. 1995. Nascent polypeptide-associated complex protein prevents mistargeting of nascent chains to the endoplasmic reticulum. *Proceedings of the National Academy of Sciences* **92**: 5411–5415. doi:10.1073/PNAS.92.12.5411.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993. doi:10.1093/BIOINFORMATICS/BTR509.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. doi:10.1093/BIOINFORMATICS/BTY191.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. doi:10.1093/BIOINFORMATICS/BTP352.

Lichtenstein P, De Faire U, Floderus B, Svartengren M, Svedberg P, Pedersen NL. 2002. The Swedish Twin Registry: a unique resource for clinical, epidemiological and genetic studies. *J Intern Med* **252**: 184–205. doi:10.1046/J.1365-2796.2002.01032.X.

Liu G, Guo Z, Zhang Q, Liu Z, Zhu D. 2020. Ahnak2 promotes migration, invasion, and epithelial-mesenchymal transition in lung adenocarcinoma cells via the tgf-β/smad3 pathway. *Onco Targets Ther* **13**: 12893–12903. doi:10.2147/OTT.S281517.

Lopez S, Stuhl L, Fichelson S, Dubart-Kupperschmitt A, St Arnaud R, Galindo JR, Murati A, Berda N, Dubreuil P, Gomez S. 2005. NACA is a positive regulator of human erythroid-cell differentiation. *J Cell Sci* **118**: 1595–1605. doi:10.1242/JCS.02295.

Machini K, Ceyhan-Birsoy O, Azzariti DR, Sharma H, Rossetti P, Mahanta L, Hutchinson L, McLaughlin H, Green RC, Lebo M, et al. 2019. Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project. *Am J Hum Genet* **105**: 177–188. doi:10.1016/j.ajhg.2019.05.017.

Mc Cartney AM, Shafin K, Alonge M, Bzikadze A V., Formenti G, Fungtammasan A, Howe K, Jain C, Koren S, Logsdon GA, et al. 2022. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nature Methods 2022 19:6* **19**: 687–695. doi:10.1038/s41592-022-01440-3.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297–1303. doi:10.1101/GR.107524.110.

Md V, Misra S, Li H, Aluru S. 2019. Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019* 314–324. doi:10.1109/IPDPS.2019.00041.

Molster C, Urwin D, Di Pietro L, Fookes M, Petrie D, Van Der Laan S, Dawkins H. 2016. Survey of healthcare experiences of Australian adults living with rare diseases. *Orphanet J Rare Dis* **11**: 1–12. doi:10.1186/S13023-016-0409-Z/TABLES/9.

Natter S, Seiberler S, Hufnagl P, Binder BR, Hirschl AM, Ring J, Abeck D, Schmidt T, Valent P, Valenta R. 1998. Isolation of cDNA clones coding for IgE autoantigens with serum IgE from atopic dermatitis patients. *FASEB J* **12**: 1559–1569. doi:10.1096/FASEBJ.12.14.1559.

Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Le Cam Y, Rath A. 2019. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *European Journal of Human Genetics 2019 28:2* **28**: 165–173. doi:10.1038/s41431-019-0508-0.

Nurchis MC, Altamura G, Riccardi MT, Radio FC, Chillemi G, Bertini ES, Garlasco J, Tartaglia M, Dallapiccola B, Damiani G. 2023. Whole genome sequencing diagnostic yield for paediatric patients with suspected genetic disorders: systematic review, meta-analysis, and GRADE assessment. *Archives of Public Health* **81**: 93. doi:10.1186/s13690-023-01112-4.

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze A V., Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. 2022. The complete sequence of a human genome. *Science (1979)* **376**: 44–53. doi:10.1126/science.abj6987.

Prodanov T, Bansal V. 2022. Robust and accurate estimation of paralog-specific copy number for duplicated genes using whole-genome sequencing. *Nature Communications 2022 13:1* **13**: 1–12. doi:10.1038/s41467-022-30930-3.

Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, Hook PW, Koren S, Rautiainen M, Alexandrov IA, et al. 2023. The complete sequence of a human Y chromosome. *Nature* **621**: 344–354. doi:10.1038/s41586-023-06457-y.

Sayyab S, Lundmark A, Larsson M, Ringnér M, Nystedt S, Marincevic-Zuniga Y, Tamm KP, Abrahamsson J, Fogelstrand L, Heyman M, et al. 2021. Mutational patterns and clonal evolution from diagnosis to relapse in pediatric acute lymphoblastic leukemia. *Scientific Reports 2021 11:1* **11**: 1–17. doi:10.1038/s41598-021-95109-0.

Schmitz D, Li Z, Lo Faro V, Rask-Andersen M, Ameur A, Rafati N, Johansson Å. 2023. Copy number variations and their effect on the plasma proteome. *Genetics* **225**. doi:10.1093/GENETICS/IYAD179.

Smedley D, Smith KR, Martin A, Thomas EA, McDonagh EM, Cipriani V, Ellingford JM, Arno G, Tucci A, Vandrovcova J, et al. 2021. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *New England Journal of Medicine* **385**: 1868–1880. doi:10.1056/NEJMoa2035790.

Tesi B, Boileau C, Boycott KM, Canaud G, Caulfield M, Choukair D, Hill S, Spielmann M, Wedell A, Wirta V, et al. 2023. Precision medicine in rare diseases: What is next? *J Intern Med* **294**: 397–412. doi:10.1111/joim.13655.

The UK10K Consortium. 2015. The UK10K project identifies rare variants in health and disease. *Nature* **526**: 82–90. doi:10.1038/nature14962.

Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, Allen HL, Sanchis-Juan A, Frontini M, Thys C, et al. 2020. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature 2020 583:7814* **583**: 96–102. doi:10.1038/s41586-020-2434-2.

Vadgama N, Ameen M, Sundaram L, Gaddam S, Gifford C, Nasir J, Karakikes I. 2022. De novo and inherited variants in coding and regulatory regions in genetic

cardiomyopathies. *Hum Genomics* **16**: 1–20. doi:10.1186/S40246-022-00420-0/FIGURES/4.

Van der Auwera GA, O'Connor BD. 2020. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media.

Vollger MR, Guitart X, Dishuck PC, Mercuri L, Harvey WT, Gershman A, Diekhans M, Sulovari A, Munson KM, Lewis AP, et al. 2022. Segmental duplications and their variation in a complete human genome. *Science (1979)* **376**. doi:10.1126/science.abj6965.

Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature 2021 597:7877* **597**: 527–532. doi:10.1038/s41586-021-03855-y.

Wen X, Liu Y, Yan Q, Liang M, Tang M, Liu R, Pan J, Liu Q, Chen T, Guo S, et al. 2018. Association of IGFN1 variant with polypoidal choroidal vasculopathy. *J Gene Med* **20**: e3007. doi:10.1002/JGM.3007.

Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, Roberts AM, Quaife NM, Schafer S, Rackham O, et al. 2020. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. *Nature Communications 2020 11:1* **11**: 1–12. doi:10.1038/s41467-019-10717-9.

Yan X, He S, Dong D. 2020. Determining How Far an Adult Rare Disease Patient Needs to Travel for a Definitive Diagnosis: A Cross-Sectional Examination of the 2018 National Rare Disease Survey in China. *International Journal of Environmental Research and Public Health 2020, Vol 17, Page 1757* **17**: 1757. doi:10.3390/IJERPH17051757.

1

Fig. 1

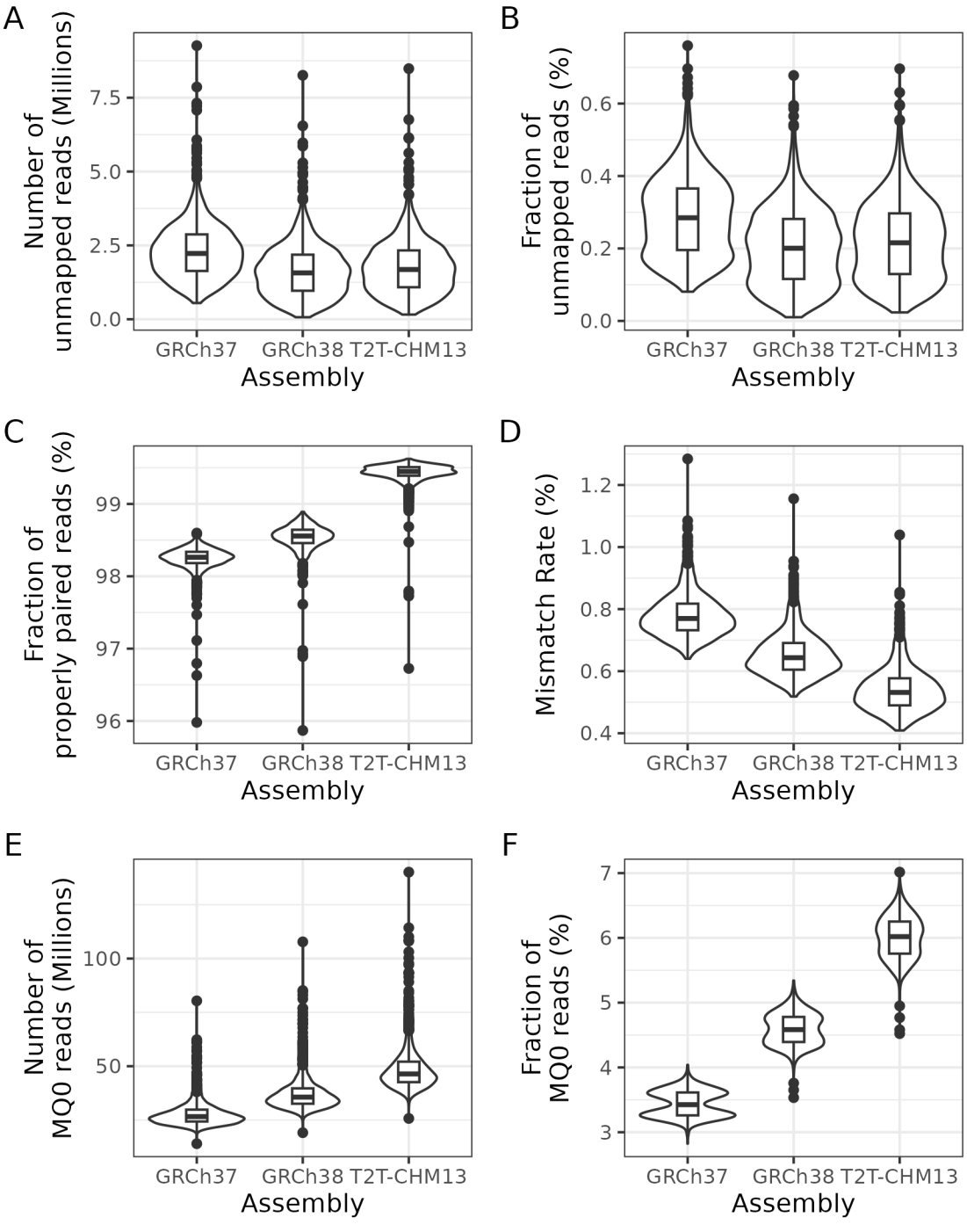
A
Number of unmapped reads (Millions)
7.5
5.0
2.5
0.0
GRCh37
GRCh38
T2T-CHM13
Assembly
B
Fraction of unmapped reads (%)
0.6
0.4
0.2
0.0
GRCh37
GRCh38
T2T-CHM13
Assembly
C
Fraction of properly paired reads (%)
99
98
97
96
GRCh37
GRCh38
T2T-CHM13
Assembly
D
Mismatch Rate (%)
1.2
1.0
0.8
0.6
0.4
GRCh37
GRCh38
T2T-CHM13
Assembly
E
Number of MQ0 reads (Millions)
100
50
GRCh37
GRCh38
T2T-CHM13
Assembly
F
Fraction of MQ0 reads (%)
7
6
5
4
3
GRCh37
GRCh38
T2T-CHM13
Assembly

Fig. 2

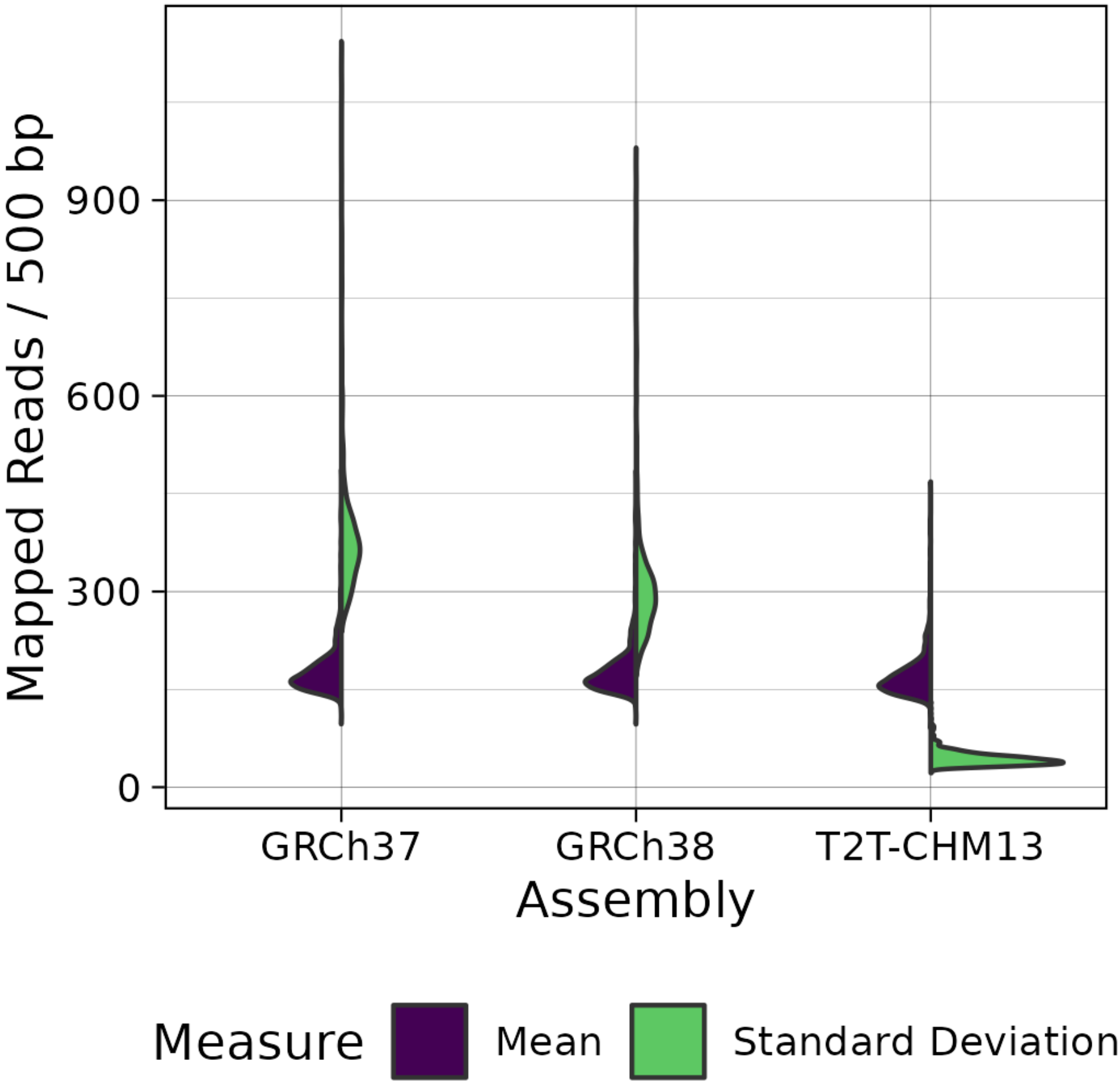
Mapped Reads / 500 bp
900
600
300
0
GRCh37
GRCh38
T2T-CHM13
Assembly
Measure
Mean
Standard Deviation

Fig. 3

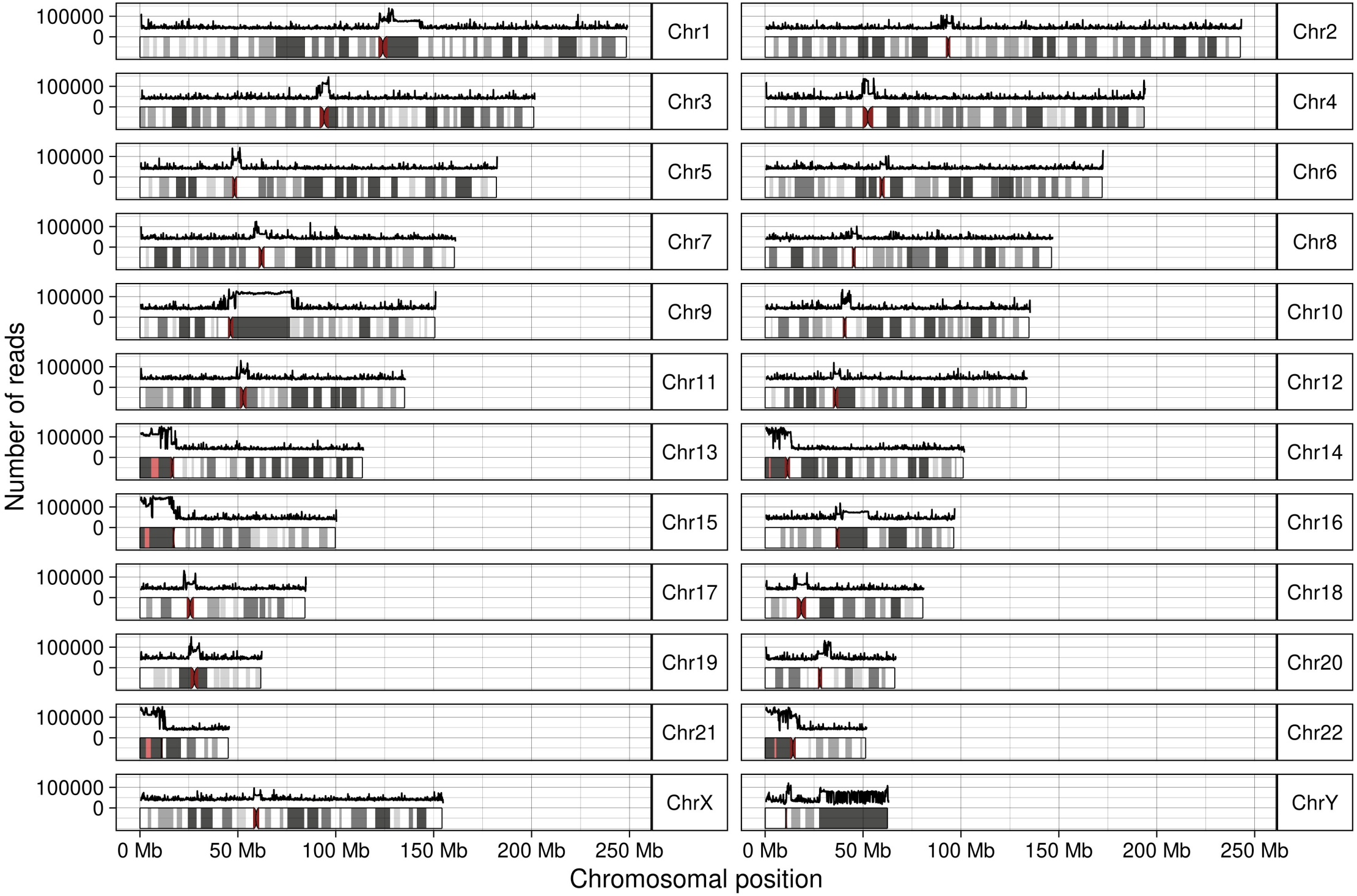
100000
0
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
ChrY
Number of reads
0 Mb
50 Mb
100 Mb
150 Mb
200 Mb
250 Mb
Chromosomal position

Fig.4

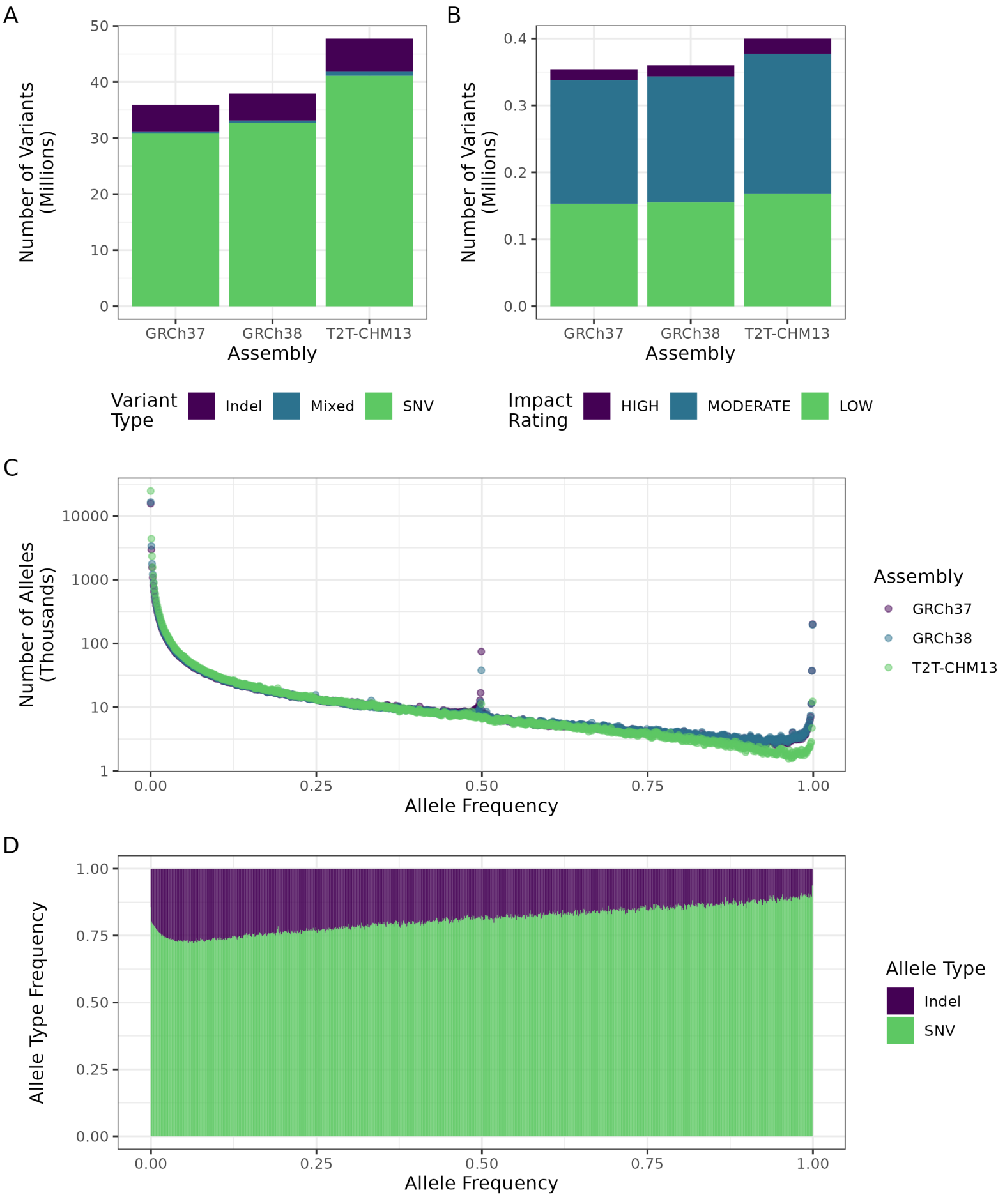
A
Number of Variants (Millions)
50
40
30
20
10
0
GRCh37
GRCh38
T2T-CHM13
Assembly
Variant Type
Indel
Mixed
SNV
B
Number of Variants (Millions)
0.4
0.3
0.2
0.1
0.0
GRCh37
GRCh38
T2T-CHM13
Assembly
Impact Rating
HIGH
MODERATE
LOW
C
Number of Alleles (Thousands)
10000
1000
100
10
1
0.00
0.25
0.50
0.75
1.00
Allele Frequency
Assembly
GRCh37
GRCh38
T2T-CHM13
D
Allele Type Frequency
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Allele Frequency
Allele Type
Indel
SNV

Fig. 5

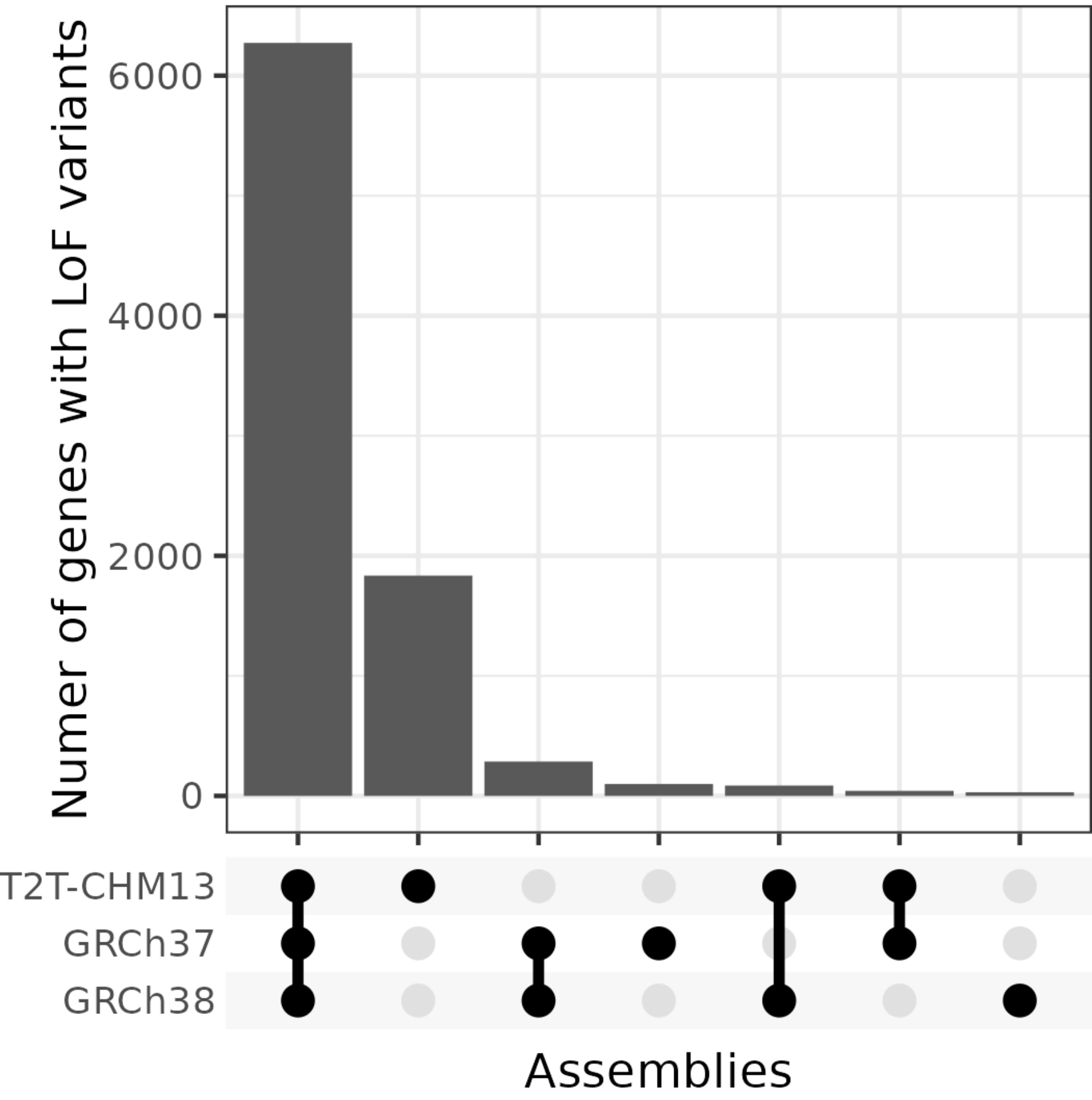
Numer of genes with LoF variants
6000
4000
2000
0
T2T-CHM13
GRCh37
GRCh38
Assemblies

Fig. 6

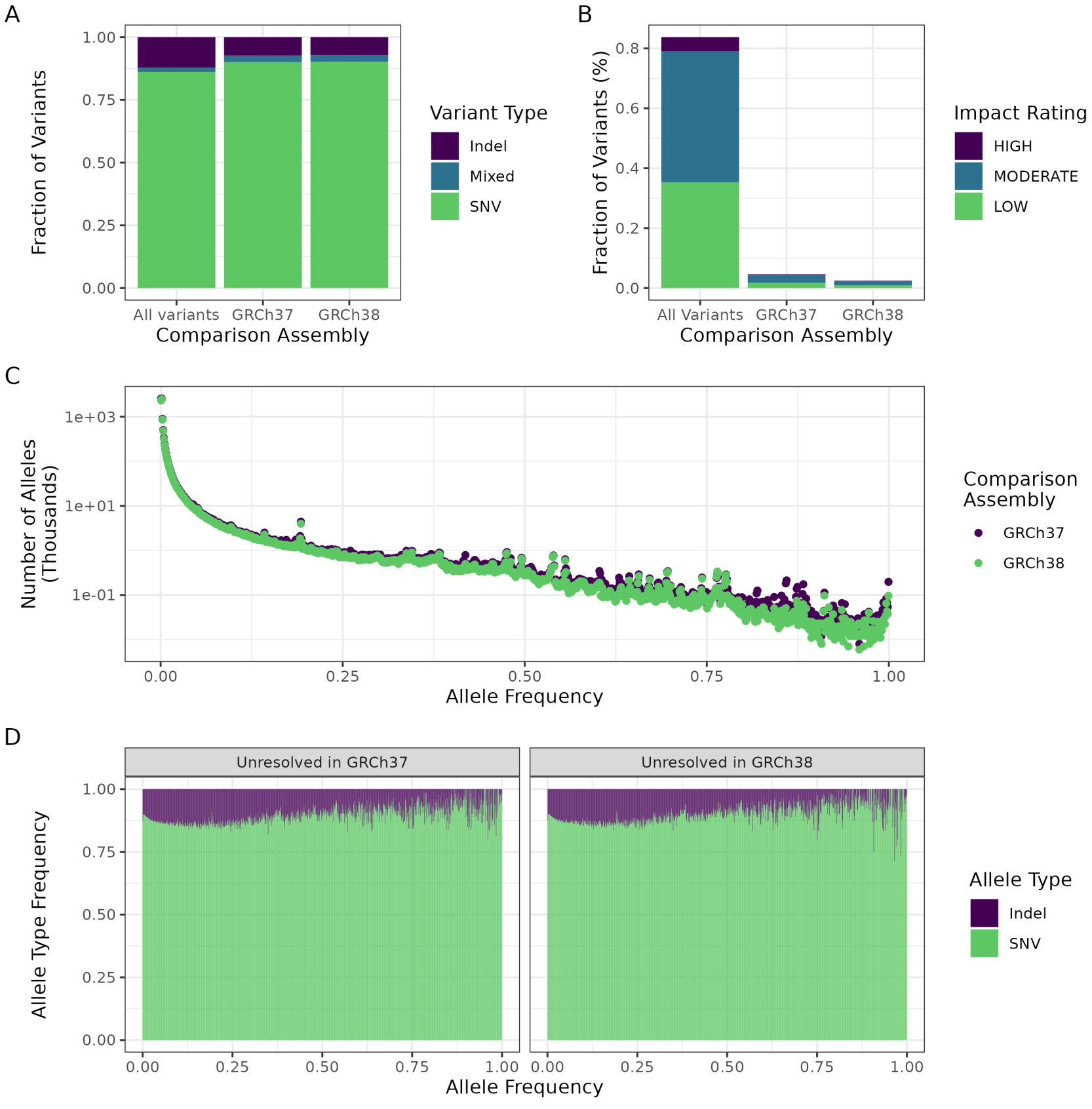
A
Fraction of Variants
1.00
0.75
0.50
0.25
0.00
All variants
GRCh37
GRCh38
Comparison Assembly
Variant Type
Indel
Mixed
SNV
B
Fraction of Variants (%)
0.8
0.6
0.4
0.2
0.0
All Variants
GRCh37
GRCh38
Comparison Assembly
Impact Rating
HIGH
MODERATE
LOW
C
Number of Alleles (Thousands)
1e+03
1e+01
1e-01
0.00
0.25
0.50
0.75
1.00
Allele Frequency
Comparison Assembly
GRCh37
GRCh38
D
Unresolved in GRCh37
Unresolved in GRCh38
Allele Type Frequency
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Allele Frequency
Allele Type
Indel
SNV

Fig. 7

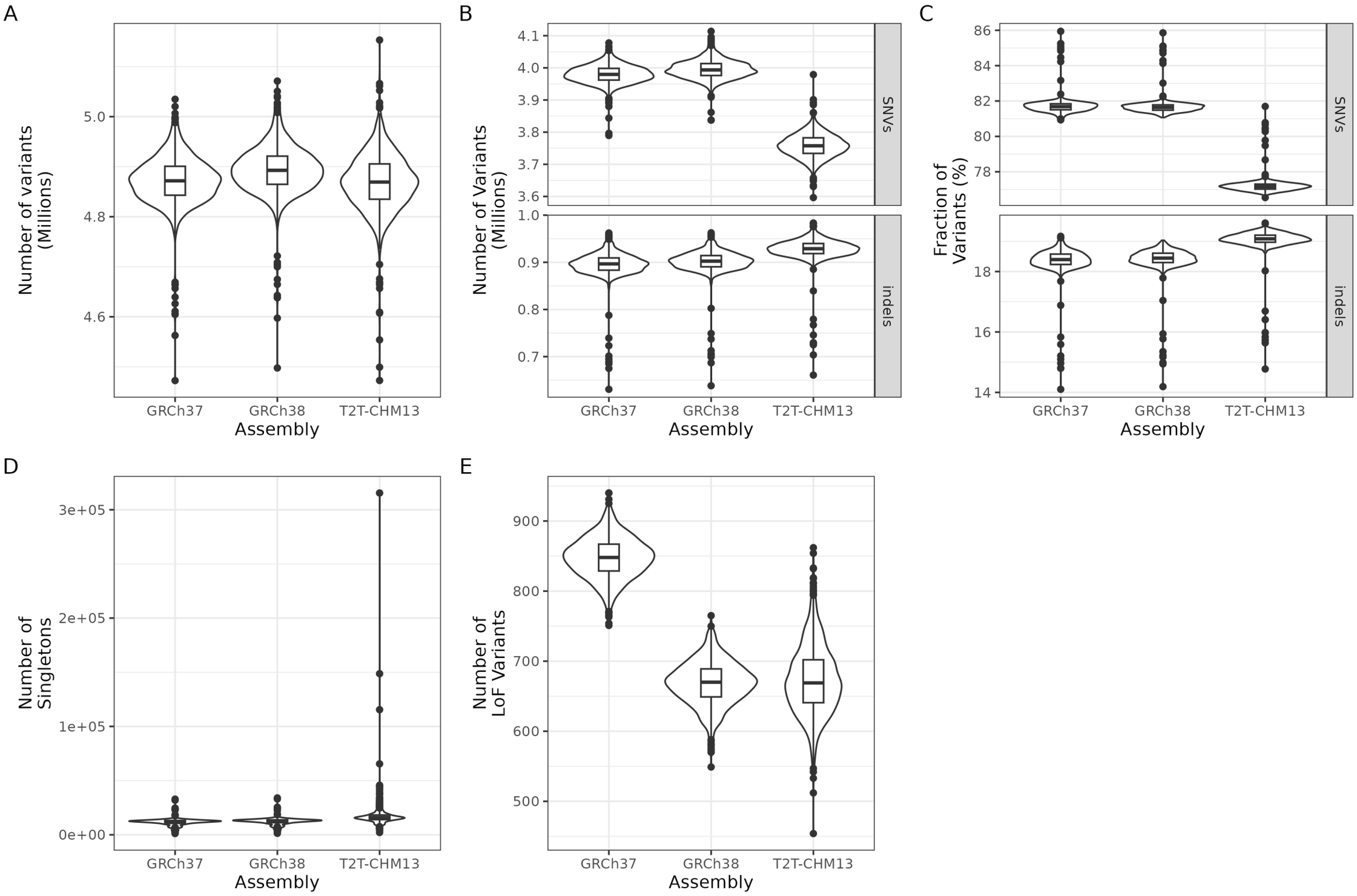
A
Number of variants (Millions)
5.0
4.8
4.6
GRCh37
GRCh38
T2T-CHM13
Assembly
B
Number of Variants (Millions)
4.1
4.0
3.9
3.8
3.7
3.6
1.0
0.9
0.8
0.7
SNVs
indels
GRCh37
GRCh38
T2T-CHM13
Assembly
C
Fraction of Variants (%)
86
84
82
80
78
18
16
14
SNVs
indels
GRCh37
GRCh38
T2T-CHM13
Assembly
D
Number of Singletons
3e+05
2e+05
1e+05
0e+00
GRCh37
GRCh38
T2T-CHM13
Assembly
E
Number of LoF Variants
900
800
700
600
500
GRCh37
GRCh38
T2T-CHM13
Assembly

Figure Legends

Figure 1. General mapping statistics for all 1000 samples. A) Distribution of the total number of unmapped reads. B) Percentage of unmapped reads per individual. C) Percentage of mapped read pairs that were properly paired i.e., both mates were mapped with the expected distance and orientation. D) Mismatch rate, i.e., the fraction of mapped bases that did not match the reference. E) Number of mappings with quality 0 i.e., not mapping uniquely. F) Percentage of mappings that had quality 0. MQ0 = mapping quality 0.

Figure 2. Summary of read depth in gene regions. Genes were divided into non-overlapping 500-bp bins and the read depth of each bin was calculated. From these, per-individual means (blue, left facing) and standard deviations (green, right facing) were obtained which are shown here.

Figure 3. Number of reads from unplaced contigs mapping across T2T-CHM13 in bins of 100 kbp. For each chromosome, an ideogram including the bands is shown; centromeres are dark red and thinner. The x axis corresponds to the position on the chromosome. The y axis is logarithmic and corresponds to the number of reads The line above each chromosome shows the number of reads mapping to a 100 kbp window.

Figure 4. Summaries of variant characteristics from the whole cohort. A) Comparison of overall variant counts, separated by variant type (SNV, indel or mixed, i.e., both) between variant calls from the different assemblies. B) Overall numbers of variants by predicted impact in the whole call set as determined by SnpEff for each assembly. Variants with impact rating MODIFIER were excluded. C) Distribution of allele frequencies in the three assemblies. D) Fraction of each variant type by allele frequency using T2T-CHM13. Frequencies were rounded to the closest multiple of 0.001 (1/1000).

Figure 5. Number of genes with predicted LoF variants shared among assemblies.

Figure 6. Summaries of variant characteristics from the whole cohort, restricted to newly resolved regions. A) Fraction of variants by type in the novel regions in the T2T-CHM13 assembly that were unresolved in the reference as shown on the x axis and the whole genome. B) Fractions of variants by predicted impact in the whole call set as determined by SnpEff in previously unresolved regions and the whole genome. Variants with impact rating MODIFIER were excluded. C) Distribution of allele frequencies for T2T-CHM13 specific variants mapped to regions not resolved in the two previous assemblies. D) Fraction of each variant type by allele frequency. Frequencies were rounded to the closest multiple of 0.001 (1/1000).

Figure 7. Summaries of several per-individual variant call statistics. Statistics included only variants passing filters. A) Number of overall variants. B) Number of SNVs and indels. C) Fraction of SNVs and indels of called variants. D) Number of singletons. E) Number of LoF variants.

Table 1. General cohort-level characteristics of alignments and variant calls with T2T-CHM13 in comparison with previous assemblies

	T2T-CHM13	**GRCh37**	**GRCh38**
Ungapped length (bp)	3,117,275,501	2,861,327,195	2,937,639,396
Avg coverage (min /max)	39.0 (20.3/105.9)	39.8 (21.3/107.7)	40.2 (21.5/108.7)
Number of variant sites*	47,744,487	35,571,130	37,938,450
Variant rate (variants / kbp)*,**	15.3	12.4	12.9
Number of SNV sites (not in dbSNPv155)	41,931,920 (10,806,250)	30,866,176 (14,806,371)	33,133,695 (12,070,889)
Number of indel sites (not in dbSNPv155)	6,650,809 (3,709,821)	5,089,447 (3,170,927)	5,221,969 (2,457,457)
Ts/Tv ratio	1.50	1.95	1.80
Number of common variants (MAF ≥ 5%)***	9,468,024	9,800,121	9,832,496
Number of low-frequency variants (MAF < 5%)***	43,151,191	29,714,010	31,895,019
Number of rare variants (MAF < 1%)***	37,402,076	24,807,638	26,741,004
Number of singletons (per kbp) [% of total variants]	19,077,986 (6.12) [40.0]	11,502,698 (4.02) [32.3]	13,547,273 (4.61) [35.7]
Number of non-reference alleles with AF > 50%	1,940,331	2,424,122	2,465,936
Number of LoF variants (% of total variants) [in ClinVar genes]	17,446 (0.037) [17,297]	12,628 (0.036) [12,326]	12,829 (0.034) [12,517]
Number of variants with low/medium/high impact (% of total variants)	168,420/208,625/ 22,848 (0.353/0.437/0.048)	153,001/184,694/ 16,392 (0.426/0.514/0.045)	155,088/188,347/ 16,542 (0.409/0.496/0.044)

Abbreviations: Ts: transition, Tv: transversion, MAF: minor allele frequency, AF: allele frequency, LoF: loss of function

** Multiallelic variants were only counted once.*

*** based on ungapped length*

**** Each alternative allele was counted separately. If an alternative allele had a frequency > 50%, the reference allele was considered the minor allele.*

Table 2. Top 10 ClinVar genes by number of loss-of-function variants by assembly

GRCh37		**GRCh38**		**T2T-CHM13**	
Gene	Number of LoF variants	Gene	Number of LoF variants	Gene	Number of LoF variants
MUC4	143	*MUC4*	172	*MUC4*	245
MUC19	99	*MUC3A*	127	*MUC19*	61
MUC3A	71	*MUC19*	100	*MUC2*	54
MUC16	48	*MUC6*	47	*MAGEC1*	46
MUC6	43	*HLA-DRB1*	39	*MUC5AC*	46
HLA-DRB1	37	*MUC16*	36	*HLA-DRB1*	35
HLA-DRB5	29	*HLA-DRB5*	33	*AHNAK2*	30
ZNF717	23	*ZNF717*	25	*MUC17*	29
PDE4DIP	22	*ZNF880*	19	*NACA*	28
ANKRD36	21	*GOLGA6L6*	18	*IGFN1*	26

Abbreviations: LoF – loss of function

Table 3. Overall characteristics of variant calls in previously unresolved regions

	Regions unresolved in GRCh37	GRCh38	All of T2T-CHM13
Total length (bp)	268,965,814	251,330,203	3,117,275,501
Number of variant sites*	8,321,170	7,826,115	47,744,487
Variant rate (variants / kbp)*	30.9	31.1	15.3
Number of SNV sites	7,710,249	7,265,225	41,931,920
Number of indel sites	829,457	771,429	6,650,809
Number of common variants (MAF ≥ 5%)**	695,892	621,512	9,293,456
Number of low-frequency variants (MAF < 5%)**	8,670,575	8,211,562	43,103,217
Number of rare variants (MAF < 1%)**	7,721,885	7,316,843	37,800,790
Number of singletons [% of total variants]	2,540,127 (30.5)	2,335,181 (29.8)	19,077,986 (40.0)
Number of LoF variants	185	99	17,446
Number of variants with low/medium/high impact (% of total variants)	1,435/2,102/286 (0.017/0.025/0.003)	719/1,048/155 (0.009/0.013/0.002)	168,420/208,625/22,848 (0.353/0.437/0.048)

Abbreviations: MAF: minor allele frequency, LoF: loss of function
** Multiallelic variants were only counted once.*
*** Each alternative allele was counted separately. If an alternative allele had a frequency > 50%, the reference allele was considered the minor allele.*

Table 4. Per-individual statistics of variant calls.

	T2T-CHM13	**GRCh37**	**GRCh38**
Mean number of variants per individual	4,868,620	4,939,144	4,892,944
Mean number of SNVs (Fraction of total variants)	3,757,819 (80.2%)	4,041,524 (81.8%)	3,995,225 (81.6%)
Mean number of indels (Fraction of total variants)	928,186 (19.8%)	901,550 (18.2%)	901,818 (18.4%)
Mean number of homozygous variants (IQR)	1,218,295 (1,197,992–1,235,602)	1,491,412 (1,476,977–1,503,380)	1,491,407 (1,476,992–1,503,377)
Mean number of heterozygous variants (IQR)	2,479,761 (2,444,590–2,512,655)	2,490,186 (2,462,633–2,520,624)	2,490,402 (2,462,660–2,520,740)
Mean number of singletons (IQR)	17,034 (14,007-18,263)	11,501 (10,550–13,122)	12,145 (11,181–13,756)
Mean number of LoF variants (IQR)	669 (641-702)	848 (829-867)	668 (649-689)

Abbreviations: IQR: Inter-quartile range, LoF: loss of function

T2T-CHM13 improves read mapping and detection of clinically relevant genetic variation in the Swedish population

Daniel Schmitz, Adam Ameur and Åsa Johansson

Genome Res. published online September 16, 2025
Access the most recent version at doi:10.1101/gr.279320.124