Supplementary Materials for

SEGMENTAL DUPLICATIONS AND THEIR VARIATION IN A COMPLETE HUMAN GENOME

Mitchell R. Vollger[1], Xavi Guitart[1], Philip C. Dishuck[1], Ludovica Mercuri[2], William T. Harvey[1], Ariel Gershman[3], Mark Diekhans[4], Arvis Sulovari[1], Katherine M. Munson[1], Alexandra P. Lewis[1], Kendra Hoekzema[1], David Porubsky[1], Ruiyang Li[1], Sergey Nurk[5], Sergey Koren[5], Karen H. Miga[4], Adam M. Phillippy[5], Winston Timp[3], Mario Ventura[2], Evan E. Eichler[1,6]

correspondence to: eee@gs.washington.edu

This PDF file includes:

Materials and Methods

Ancestry of CHM13 and GRCh38.

Based on maximum likelihood admixture analysis, we previously determined (*18*) that CHM13 is primarily of European origin with some slight evidence of Asian or Amerindian admixture. GRCh38, in contrast, is a composite of multiple human samples and haplotypes. However, Green and colleagues *(19)* estimated two-thirds of GRCh38 was derived from a single individual (RP11) with 42% African ancestry.

Determining the final set of SD annotations.

To annotate SDs we identified homologous segments using SEDEF [v1.1-31-g68de243 *(21)*] on a masked version of the T2T-CHM13 v1.0 assembly that included chrY from GRCh38. Masking was performed using Tandem Repeats Finder (TRF) (*89*) and RepeatMasker (*90*) so that only regions with homology outside of common repeat elements would be identified by SEDEF. The resulting homologies identified by SEDEF were then filtered to have: 1) at least 90% gap-compressed identity, 2) at most 50% gapped sequence in the alignment, 3) at least 1 kbp of aligned sequence, and 4) at most 70% satellite sequence as determined by RepeatMasker. The remaining homologies were then used as the final SD annotations for T2T-CHM13 v1.0. SDs were further defined as pericentromeric or telomeric if they were within 5 Mbp of the centromere or 500 kbp of the telomere. The full pipeline for making these annotations is provided at Zenodo (10.5281/zenodo.5498988) under workflows/sedef.smk (*118*). The same workflow was applied to the chromosome-level scaffolds of GRCh38 for all SD comparisons made in the paper.

Repeatmasking.

Common repeats were masked with RepeatMasker v4.1 (*90*) and TRF (*89*). The full pipeline for these masking steps is provided for convenience at Zenodo (10.5281/zenodo.5498988) under workflows/mask.smk (*118*). In brief, RepeatMasker was run with the following settings:

```
RepeatMasker \
  -s \
  -xsmall \
  -e ncbi \
  -species human \
  -dir $(dirname {input.fasta}) \
  -pa {threads} \
  {input.fasta}
```

And TRF was run with:

```
trf {input.fasta} 2 7 7 80 10 50 15 -l 25 -h -ngs > {output.dat}
```

Defining syntenic regions between T2T-CHM13 and GRCh38.

The T2T-CHM13 to GRCh38 synteny track was constructed using the Cactus HAL file available at this link (http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/cactus/t2t-chm13-v1.0.aln1.hal) with 1 Mbp resolution and a maximum anchor distance of 50 kbp. We used the tool halSynteny to construct syntenic blocks from the Cactus alignments—as described in detail in Krasheninnikova et al. *(91)*. After the alignment was constructed, we inspected the alignments to ensure that they were one-to-one best mappings between the two genome assemblies. This track is available at this link (http://t2t.gi.ucsc.edu/chm13/hub/t2t-chm13-v1.0/synteny/synteny.1mb.bigPsl) and on Zenodo (https://doi.org/10.5281/ZENODO.4721956, *116*). To define the previously unresolved and variable regions of T2T-CHM13, we selected the reciprocal segments from the 1 Mbp synteny track and retained all regions without an alignment to GRCh38.

Calculating SD sequence previously-unresolved-by-content versus previously-unresolved-by-structure.

To define SD sequences with no paralogous match in GRCh38 (previously-unresolved-by-content), we gathered all SDs identified in GRCh38 and multi-mapped them to T2T-CHM13 using the following minimap2 command (version 2.22, *101*):

```
minimap2 -x asm20 --eqx -s 500 -N 1000 -p 0.01 \
        {T2T reference} {GRCh38 SDs}
```

We then subtracted these alignments from the 81.34 Mbp of previously unresolved or structurally variable SD sequences to determine all SDs without a paralogous match in GRCh38. In total, we find 16.52 Mbp of SD sequence matching this definition, which we categorize as “previously-unresolved-by-content”, and analogously, 64.81 Mbp of SD sequence that is “previously-unresolved-by-structure.” Of the 6,276 SDs that are previously-unresolved-by-content, 5,071 (80.8%) are contained or map within the acrocentric short arms. We include an ideogram showing all these pairwise alignments in fig. S24.

Counting previously unresolved segmental duplication bases.

We report 81.3 Mbp of SD sequence that overlaps with the previously unresolved or structurally variable sequence in the T2T-CHM13 assembly. This differs from the 68.3 Mbp reported by Nurk et al. *(20)* because we considered all SD pairwise alignments that overlapped with previously unresolved sequences whereas Nurk et al. uses a strict intersection of previously unresolved base pairs overlapping SDs (fig. S25). For example, if a 75 kbp SD overlapped 50 kbp of previously unresolved sequence, we would report 75 kbp of previously unresolved SD sequence whereas Nurk et al. would report 50 kbp. The reason for this difference is because the goal of Nurk et al. was to define which bases were previously unresolved in the T2T-CHM13 genome whereas our goal was to define the previously unresolved or structurally variable SDs. We therefore calculate the number of changed SD bases including the whole SD alignment because this homology is entirely new in T2T-CHM13.

Calculating the number of SD alignments in 5 Mbp windows.

We first offset the SD coordinates in GRCh38 such that the largest gaps [acrocentric short arms, centromeres, and human satellite (HSAT) arrays] matched the length of the assembled sequence in T2T-CHM13. We then normalized the GRCh38 coordinates so that the length of the chromosomes in GRCh38 were equal to those in T2T-CHM13. After this we took 5 Mbp non-overlapping windows from T2T-CHM13 and the normalized GRCh38 and calculated the difference in the number of SDs within each window (table S3).

WSSD detection and genotyping.

As an orthogonal method to estimate copy number of SDs, we applied the whole-genome shotgun sequence detection (WSSD) pipeline, which uses sequence read-depth as a proxy (*14*). Short-read sequence data were processed into 36 bp non-overlapping fragments and mapped to a masked T2T-CHM13 reference using mrsFAST *(92)* with a maximum of two substitution mismatches not allowing for indels. Masking was determined by TRF and RepeatMasker. Read-depth across the genome was corrected for GC bias and copy number was determined using linear regression on read-depth versus known fixed copy number control regions. Finally, integer genotypes were estimated by using the predicted mean and variance of the Gaussian distributions underlying different copy numbers to create a series of models to represent the likely distribution of read-depths underlying a region of specific copy number.

For defining genotyping intervals, we applied the changepoint package in R (*93*) to identify regions where the CHM13 WSSD copy number estimate was consistent. Specifically, we used a log-transformed continuous copy number estimates from WSSD for sliding windows across the assembly and then applied binary segmentation to identify regions where the copy number remained the same. We used the following R command:

```
cpt.mean(Log_cn, method = "BinSeg", Q=Q)
```

Where Log_cn is a vector of log-scaled copy number estimates and Q is the number of independent 50 kbp windows within each chromosome.

Comparing diploid copy numbers to haploid references.

When estimating copy number from short-read data for the SGDP, we report diploid copy number estimates, which are the aggregate copy number from both haplotypes. In order to make these estimates comparable to the haploid references (GRCh38 and T2T-CHM13), we decompose both references to k-mers (k=36) and apply the same read-depth genotyping method (described in WSSD detection and genotyping). This is why, for example, in Figure 5 we show annotations for two copies of DEFB103A/B but it has a diploid copy number estimate of four in table S6. To validate the copy number of assemblies, we used the same the output of WSSD copy number on the k-mer fragmented references. Then every copy number estimate within SD space was compared between the Illumina estimate and assembly estimate and a Pearson's correlation was calculated.

Gene annotations with Liftoff.

Gene annotations on T2T-CHM13 were made using Liftoff (*94*) and then processed with GffRead (*95*) to filter for only transcripts with open reading frames (ORFs). The full pipeline for gene annotation is provided for convenience at Zenodo (10.5281/zenodo.5498988) under workflows/liftoff.smk (*118*). In brief, Liftoff was called with the following command:

```
liftoff -dir {output.temp} \
   -f <(echo "locus") \
   -flank 0.1 \
   -sc 0.85 -copies -p {threads} \
   -g {input.gff} -o {output.gff} -u {output.unmapped} \
   {input.t} {input.r}
```

Using as input the GENCODE Genes track v34 annotation gff3 available at ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_34/gencode.v34.annotation.gff3.gz and GRCh38 FASTA (available at https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13/GRCh38_major_release_seqs_for_alignment_pipelines/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz).

Counting additional genes.

Nurk et al. reports 140 previously uncharacterized protein-coding genes, and we report 182 with multiple exons and ORF *(20)*. This difference comes from different gene annotation sets and different filtering steps. Nurk et al. uses the Comparative Annotation Toolkit (CAT) to annotate genes in the assembly which they then supplement with Liftoff while considering only additional gene copies that are 100% identical to previously annotated paralogs. We used Liftoff to generate our gene annotations and allowed additional paralogs to diverge by up to 15% from previously annotated paralogs if they still had an ORF— and this difference accounts for the additional genes we identify. Of the 140 genes, 58 overlap with the 182 we report. This difference comes from an additional set of filters we used to offset the false positives that were identified by allowing 15% divergence. Our additional filters were that the gene must have an ORF of at least 200 bp and have multiple exons.

Counting the number of high-identity SD genes.

We counted all protein-encoding genes with at least one exon mapping fully within a >95% identical SD and had the additional condition that at least 50% of the full-length gene maps to SD space without the identity limitation.

Cell culture.

CHM13 and CHM1 cells were cultured in complete AmnioMax C-100 Basal Medium (Thermo Fisher Scientific, 17001082) supplemented with 15% AmnioMax C-100 Supplement (Thermo Fisher Scientific, 12556015) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122). GM24385, GM19240, HG00514 and HG00733 cells were cultured in RPMI 1640 with L-glutamine medium (Thermo Fisher Scientific,

11875093) supplemented with 15% FBS (Thermo Fisher Scientific, 16000-044) and 1% penicillin-streptomycin (Thermo Fisher Scientific, 15140122). All cells were cultured in a humidity-controlled environment at 37°C with 5% CO_2.

FISH characterization and validation.

Fosmid probes for FISH experiments were selected by mapping fosmid end sequences from the ABC10 (NA19240 Yoruban) library *(28)* to the T2T-CHM13 reference using BLAST *(96)*. FISH experiments were essentially performed as previously described *(97)*. Human fosmid clones were used as probes in one- or two-color FISH experiments and directly labeled by nick-translation with Cy3-dUTP (PerkinElmer) and fluorescein-dUTP (Enzo), as previously described *(98)*. Briefly, 300 ng of labeled probe was used for the FISH experiments; hybridization was performed at 37°C in 2× SSC, 50% (v/v) formamide, 10% (w/v) dextran sulphate, and 3 mg sonicated salmon sperm DNA in a volume of 10 mL. Posthybridization washing was at 60°C in 0.1 × SSC (three times, high stringency) and post-washing staining was performed by DAPI (5 minutes). The hybridizations were performed on metaphases and nuclei obtained from CHM13, CHM1, GM24385, GM19240, HG00514, and HG00733 lymphoblastoid cell lines.

Slides were imaged on a fluorescence microscope (Leica DM RXA2) equipped with a charge-coupled device camera (CoolSNAP HQ2) and a 100× 1.6–0.6 NA objective lens. DAPI, Cy3, and fluorescein fluorescence signals, detected with specific filters, were recorded separately as grayscale images. Pseudocoloring and merging of images were performed using Adobe Photoshop software. Mapping was performed following comparison to the conventional classical cytogenetics G-banding *(99)*.

Assembly of additional humans and nonhuman primates.

All assemblies except for T2T-CHM13 and GRCh38 were assembled with hifiasm v0.12 using default parameters. The human samples, with the exception of CHM1, were assembled using parental short-read data for phasing. All nonhuman primates and CHM1 samples were assembled without parental phasing information since none exists.

ONT validation.

To validate structural variant configurations predicted by HiFi sequence and assembly, we aligned ultra-long ONT data from two samples (HG002, HG00733) and assessed the uniformity of coverage over the *TBC1D3* assemblies for these four haplotypes. We find no obvious sign of collapsed duplications (read coverage abnormalities) or misjoins in the assemblies (every 25 kbp segment with 1 kbp slide is spanned by four or more reads) in the ultra-long ONT data (figs. S19-S20).

TBC1D3 phylogenetic tree construction.

Orthologous sequences for the two human *TBC1D3* expansion sites were identified in T2T-CHM13 using minimap2 (*99*) and gene models were annotated using Liftoff (*94*). *TBC1D3* transcripts with ORFs were identified using gffread (*95*). Exons were masked and removed using BEDTools maskfasta and getfasta functions (*100*) in order to construct neutrally evolving phylogenetic trees. With exon-free paralogs of both CHM13 and nonhuman primates, a multiple sequence alignment (MSA) was generated using

MAFFT (*102*). To produce the most confident MSA, an iterative refinement algorithm described was used with the option for iterating 1000 times (*104*).

```
mafft --reorder --maxiterate 1000 --thread 16 {input.fasta} > {output.MSA.fasta}
```

The MSA was subsequently used to generate a maximum likelihood phylogeny, using RAxML (*103*). For this phylogeny, the rapid bootstrapping analysis was utilized to identify the best maximum likelihood tree, a gamma model was used to model rate heterogeneity, and macaque *TBC1D3* sequences were used as outgroup sequences.

```
raxmlHPC-PTHREADS -f a -p 12345 -x 12345 \
        -s {input.fasta} -m GTRGAMMA \
        -# 100 -T 8 -n {output.fasta.name} \
        -o {outgroup.sequence.names}
```

Defining structurally variable haplotypes.

To define the set of structurally distinct haplotypes for the evolutionary and biomedically important loci, we performed an all against all pairwise alignment for each of the haplotypes using the following minimap2 command (*101*):

```
minimap2 -r 50000 -ax asm20 --eqx -Y
```

Sequences aligned to the same haplotype for at least 90% of their length at >99% identity without deletions or insertions of 50 kbp or more were grouped into a single structural haplotype and considered not structurally variable. Structurally variable haplotypes were then defined as the mutually exclusive groups where every haplotype in a given group did not align to the haplotype of any other group for >90% of its length at >99% identity. Similarly, haplotypes that were grouped together with GRCh38 or CHM13 were deemed to have recapitulated the structural organization of that particular reference.

Once these structural alleles were defined, we calculated the % structural heterozygosity as follows:

$$1 - \sum_{i=1}^{k} p_i^2$$

where p_i is the allele frequency of the i^{th} of k structurally distinct alleles.

Variation graphs for SD loci.

We applied minigraph v0.14 (*61*) to construct variation graphs using all structurally distinct haplotypes with the parameters:

```
minigraph -xggs -L 5000 -r 100000 -t {threads} *.fasta
```

All haplotypes were aligned back to the graph to call variants:

```
minigraph -x asm -t {threads} {input.gfa} {input.fasta}
```

Methylation analysis.

Methylation analysis was performed using the same data and methods described by Gershman et al., bioRxiv (*106*). In brief, CHM13 ultra-long ONT reads were aligned to the CHM13 reference with Winnowmap2 (*107*) with a k-mer size of 15 and filtered for primary alignments for read lengths greater than 50 kbp. To measure CpG methylation in nanopore data, we used Nanopolish (v0.13.2) (*65*) filtered methylation calls using the nanopore_methylation_utilities tool (Zenodo: 10.5281/zenodo.5498988, *118*), which uses a log-likelihood ratio of 1.5 as a threshold for calling methylation. Methylation data was then loaded into R for all downstream analysis with GenomicRanges (*109*) and dplyr (*110*).

Testing for enrichment of untranscribed SD genes in hypomethylated genomic regions.

To test if hypomethylated regions were enriched for transcriptionally untranscribed SD genes, we performed a permutation test. Specifically, we took the total number of untranscribed SD genes and randomly labeled them as being in hypomethylated or hypermethylated genomic regions with probability proportional to the number of bases in each category. This experiment was repeated 10,000 times and compared to our actual observation to determine a p-value.

Custom ideogram and homology visualizations.

Linear ideograms were constructed using the karyoploteR package (*111*) and circular ideograms were made using circlize (*110*). R code used to make these figures is shared for convenience at Zenodo (10.5281/zenodo.5498993) (*117*); however, this is not a software package and is provided without extensive documentation or installation instructions.

Sequence homology plots were made with a modified version of Miropeats (*108*) that uses minimap2 to identify alignments. Code for the homology plots can be found at Zenodo (10.5281/zenodo.5498988) under workflows/minimiro.smk (*113–115, 118*). In brief, sequences are aligned using the following minimap2 parameters:

```
minimap2 -x asm20 -r 200000 -s 100000 \
   -N 1000 --secondary=no \
    --cs {input.ref} {input.query} > {output.paf}
```

and then processed into a postscript file using scripts/minimiro.py and converted into a PDF.

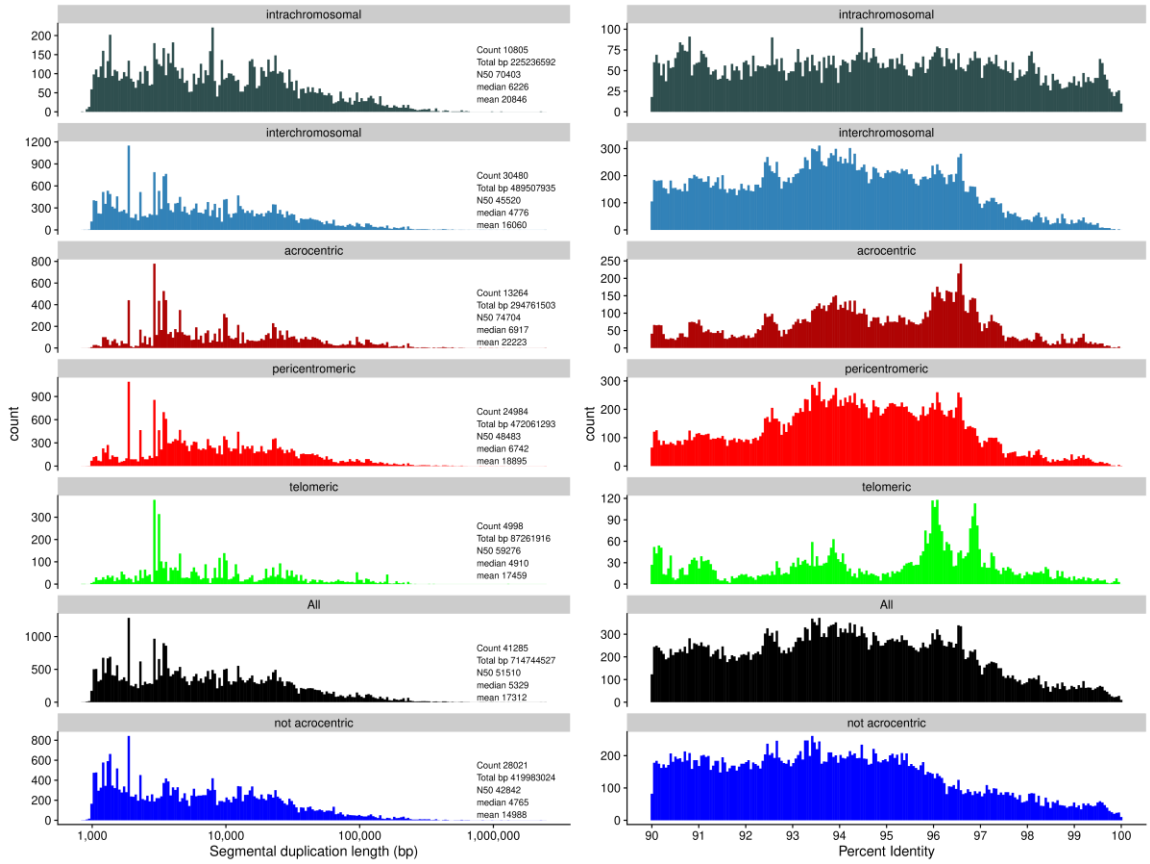

Fig. S1

Comparison of SD length and identity in different regions of the genome.
The length (left) and identity (right) of SDs across commonly delineated regions of the genome (colors). Acrocentric SDs are significantly longer than all other SD categories (p-value < 0.01, one-sided Wilcoxon rank-sum test).

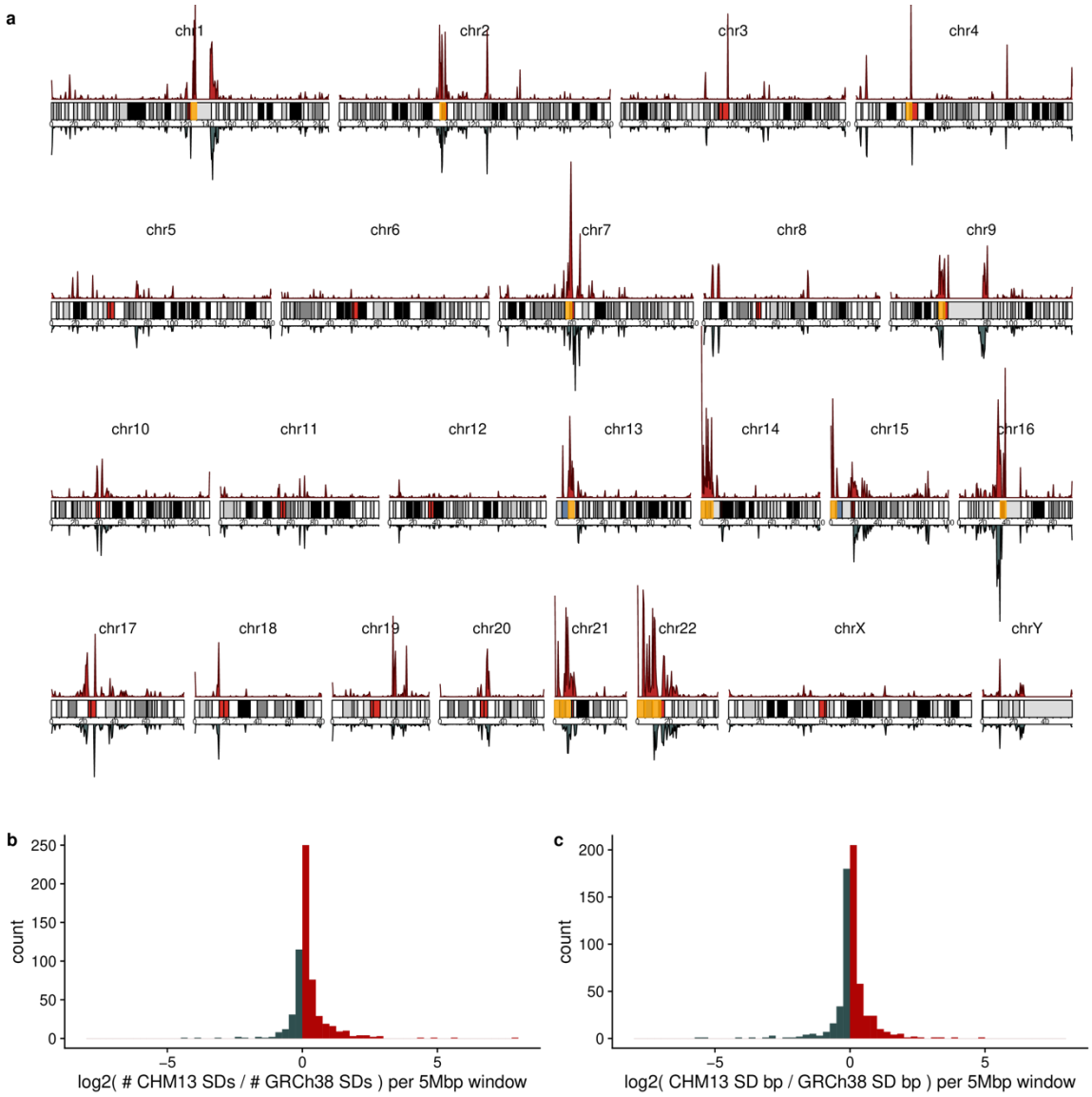

Fig. S2

SD density comparison between T2T-CHM13 and GRCh38.

a) Density of SDs in T2T-CHM13 (red) and GRCh38 (blue). In the ideogram highlighted in orange are the 15 regions with the largest increase in the number of SDs. **b)** Histogram showing the $\log_2$ fold change between the number of SDs in T2T-CHM13 and GRCh38 per non-overlapping 5 Mbp window. **c)** Histogram showing the $\log_2$ fold change between the number of base pairs in SDs for T2T-CHM13 and GRCh38 per non-overlapping 5 Mbp window.

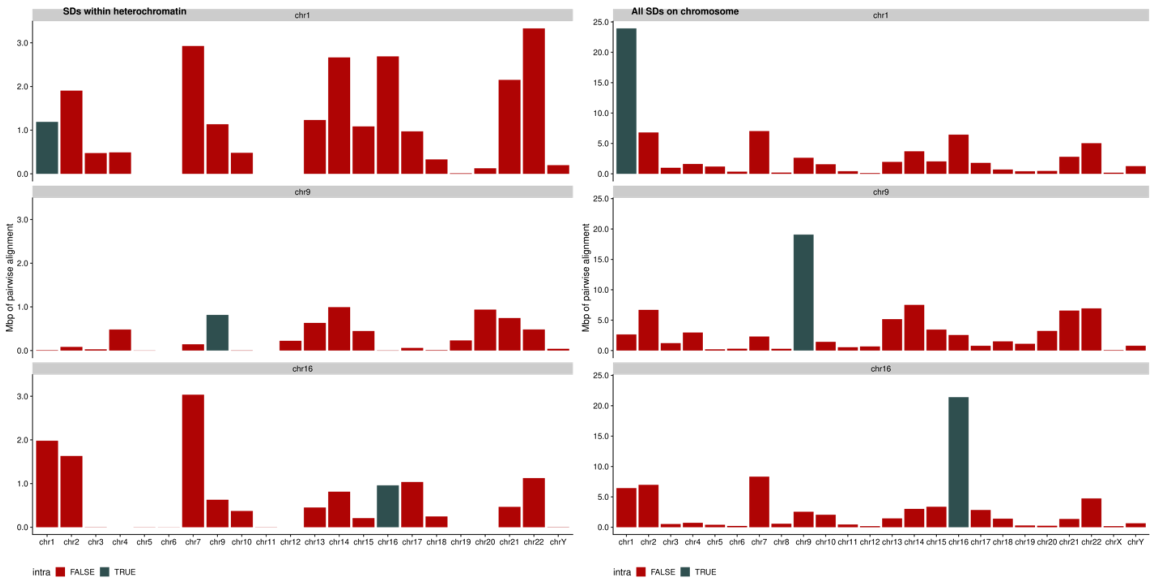

Fig. S3

SDs within heterochromatin on chromosomes 1, 9, and 16.
This figure shows where the SDs that separate the HSAT and centromere arrays on chromosomes 1, 9, and 16 align to (left) compared to the overall distribution of that chromosome (right). Blue are intrachromosomal SDs and red are interchromosomal.

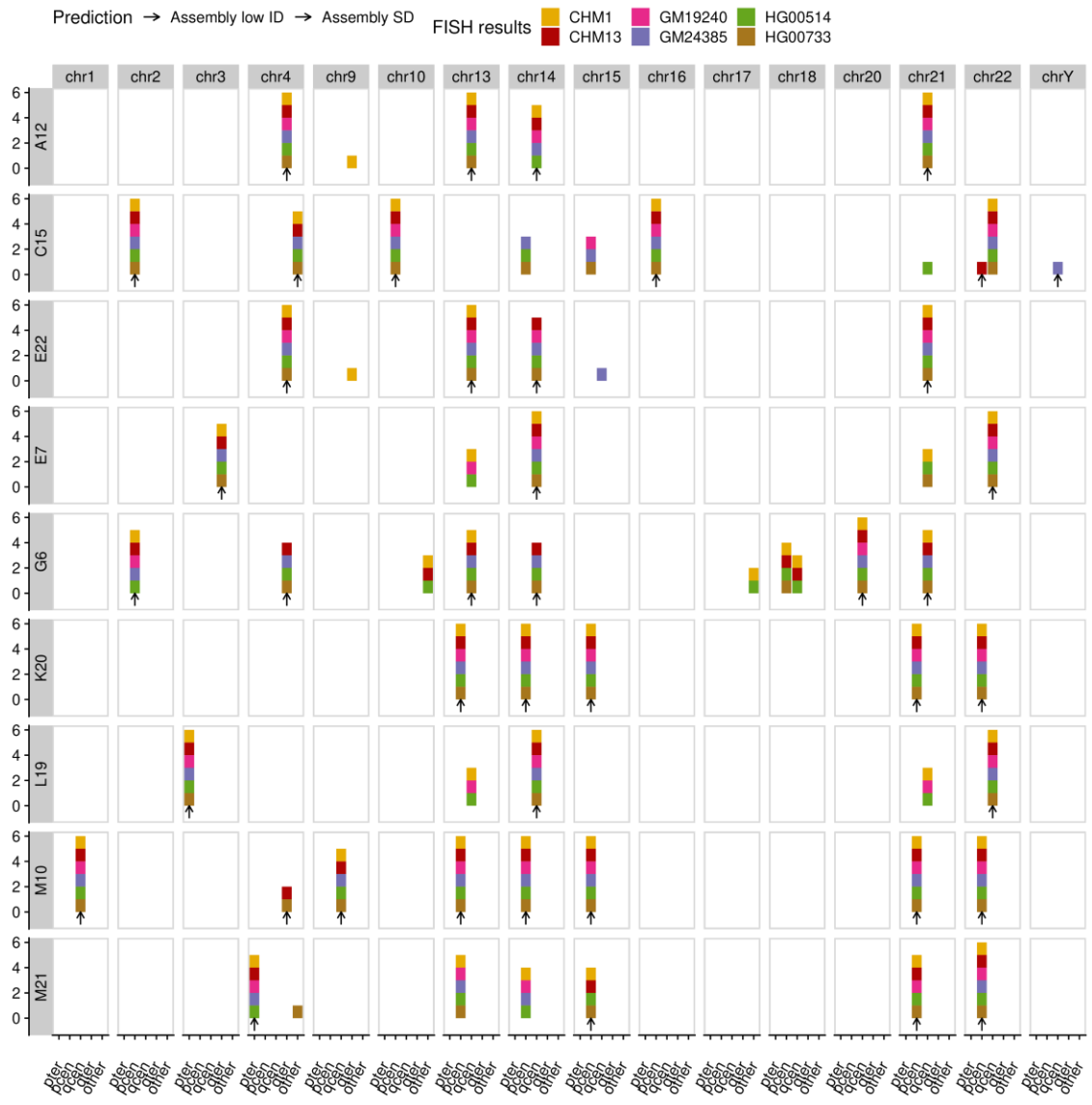

Fig. S4

FISH support for previously unresolved duplications in T2T-CHM13.
This table shows the location and chromosomes of FISH signals (x) for each probe (y) across the different cell lines (color). Black shows the predicted locations of FISH signals from the T2T-CHM13 assembly.

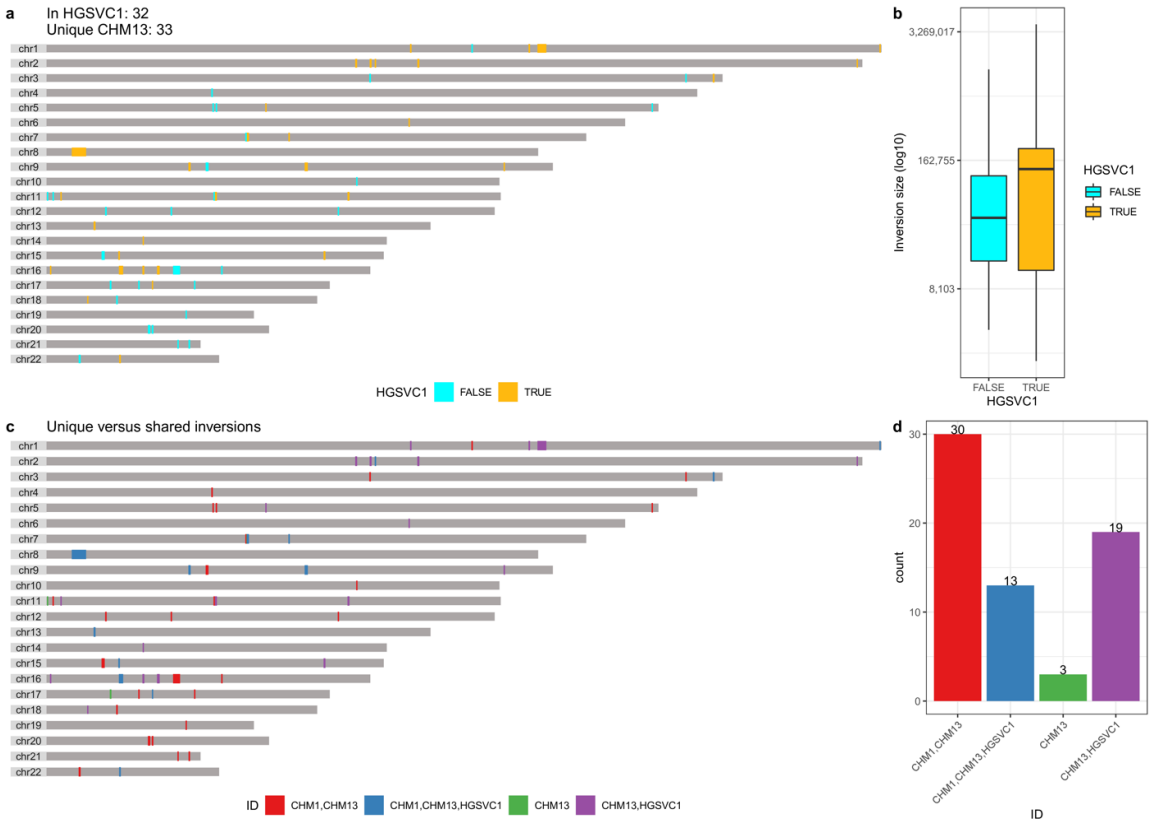

Fig. S5

CHM13 inversions supported by Strand-seq.
a) Inversion locations in CHM13 relative to GRCh38 as identified with Strand-seq. The color indicates whether the inversion is shared (orange) with at least one sample from HGSVC1 (*4*) or unique to CHM13 (cyan). **b)** Size distribution of inversions in CHM13. **c)** Comparison of inversions shared between CHM1 and CHM13. **d)** Bar chart showing the counts of shared and unique inversions in CHM13.

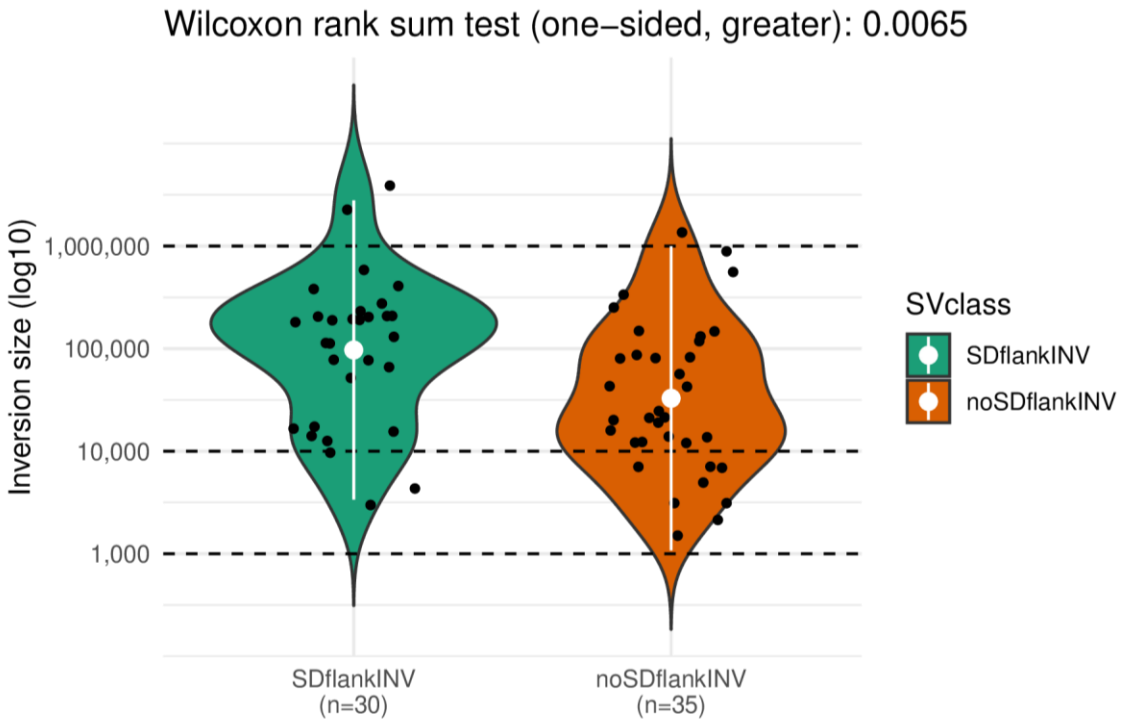

Fig. S6

Length of CHM13 inversions.
The length of inversions in CHM13 as predicted by Strand-seq stratified by the presence of flanking SDs (green) or lack thereof (orange). Inversions flanked by SDs are significantly longer than other inversions (p = 0.0065, one-sided Wilcoxon rank-sum test).

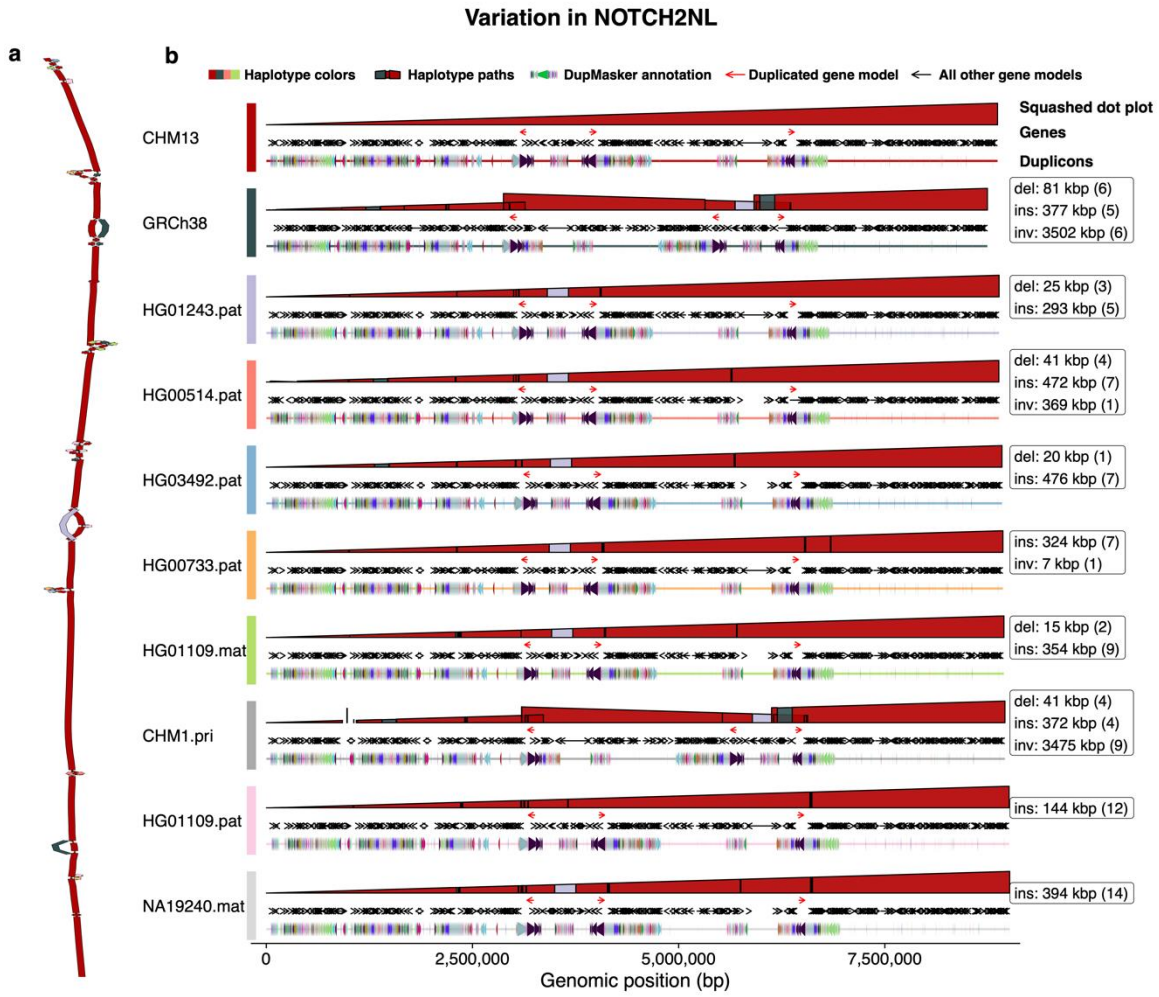

Fig. S7

Pangenome graph of *NOTCH2NL* and *SRGAP2*.

a) Variation in human haplotypes across the *NOTCH2NL* expansion site: a graph representation [rGFA generated using minigraph (*5*)] of the locus where colors indicate the source genome for the sequence. The graph visualization was created using the software tool Bandage (*6*). **b)** The path for each haplotype-resolved assembly through the graph. The "squashed dot plot" represents a vertically compressed dot plot comparing the haplotype-resolved sequence (horizontal) against the graph (vertical). Color represents the source haplotype for the vertical sequence. Structural variants can be identified from discontinuities in height (deletion), changes between colors (insertion), or changes in the direction of the polygon (inversion). *NOTCH2NL*, the gene of interest, is shown with red arrows and other genic content in the region are shown with black arrows. The final line is a duplicon track showing the ancestral duplications (color) that make up the larger duplication block. The coordinates in T2T-CHM13 v1.0 chr1:142242498-151009743 and GRCh38 chr1:143189295-151936076.

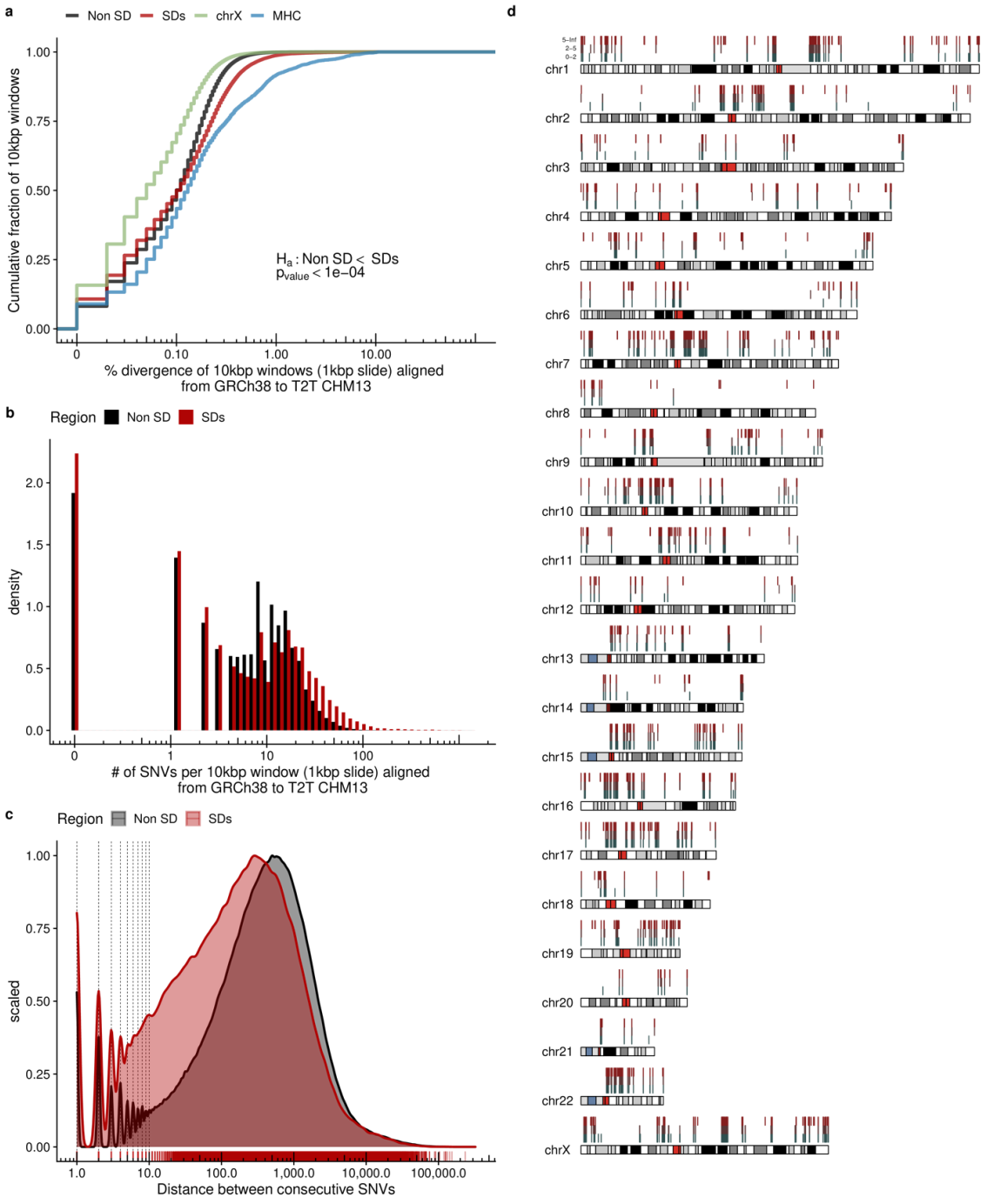

Fig. S8

Single-nucleotide variants (SNVs) in SDs between T2T-CHM13 and GRCh38.
a) Divergence of 10 kbp windows with synteny between GRCh38 and T2T-CHM13. **b)** Distribution of the number of SNVs per 10 kbp windows aligned from GRCh38 to T2T-CHM13 in unique and SD regions. **c)** Distribution of the distance between SNVs in the syntenic regions of GRCh38 and T2T-CHM13. **d)** SD regions with synteny between T2T-CHM13 and GRCh38 and their average levels of single-nucleotide variation in 1 kbp windows. The bottom row has SD windows with 0-2 SNVs per kbp, middle row 2-5 SNVs per kbp, and top row is greater than 5 SNVs per kbp.

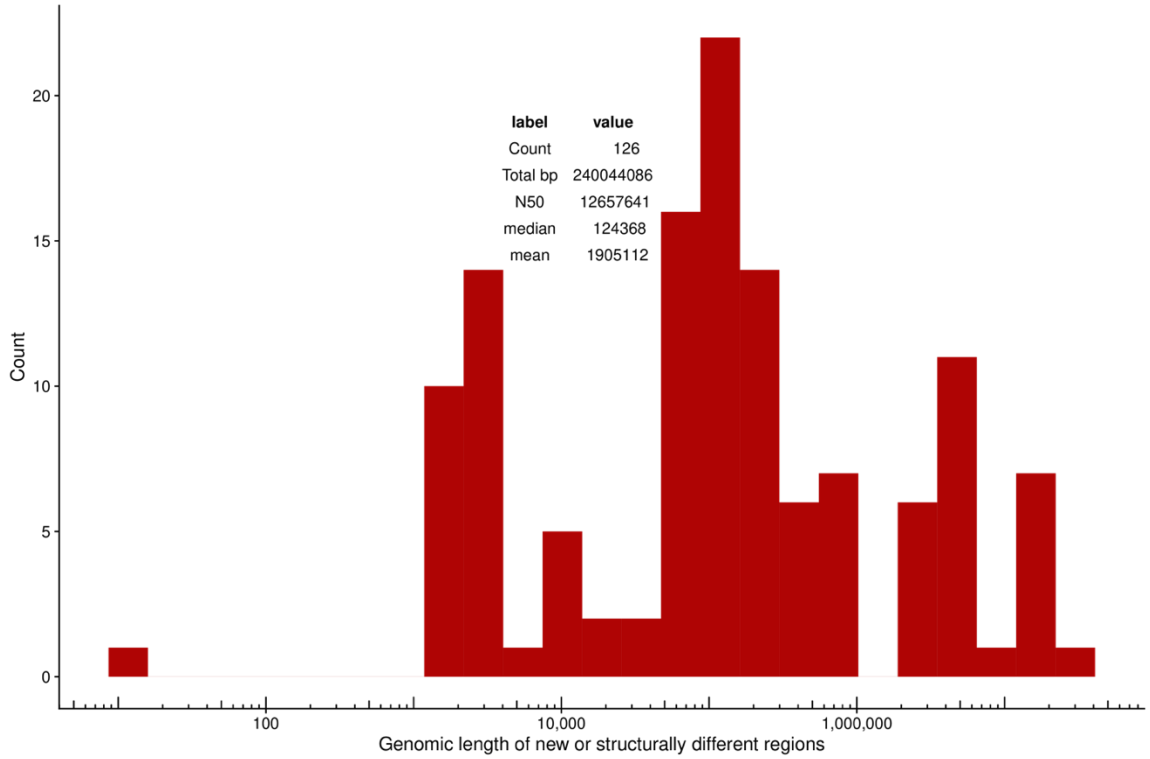

Fig. S9

Size distribution of non-syntenic regions between GRCh38 and T2T-CHM13. Histogram showing the size of non-syntenic regions (Methods) between GRCh38 and T2T-CHM13, and a table of statistics on the lengths of the region.

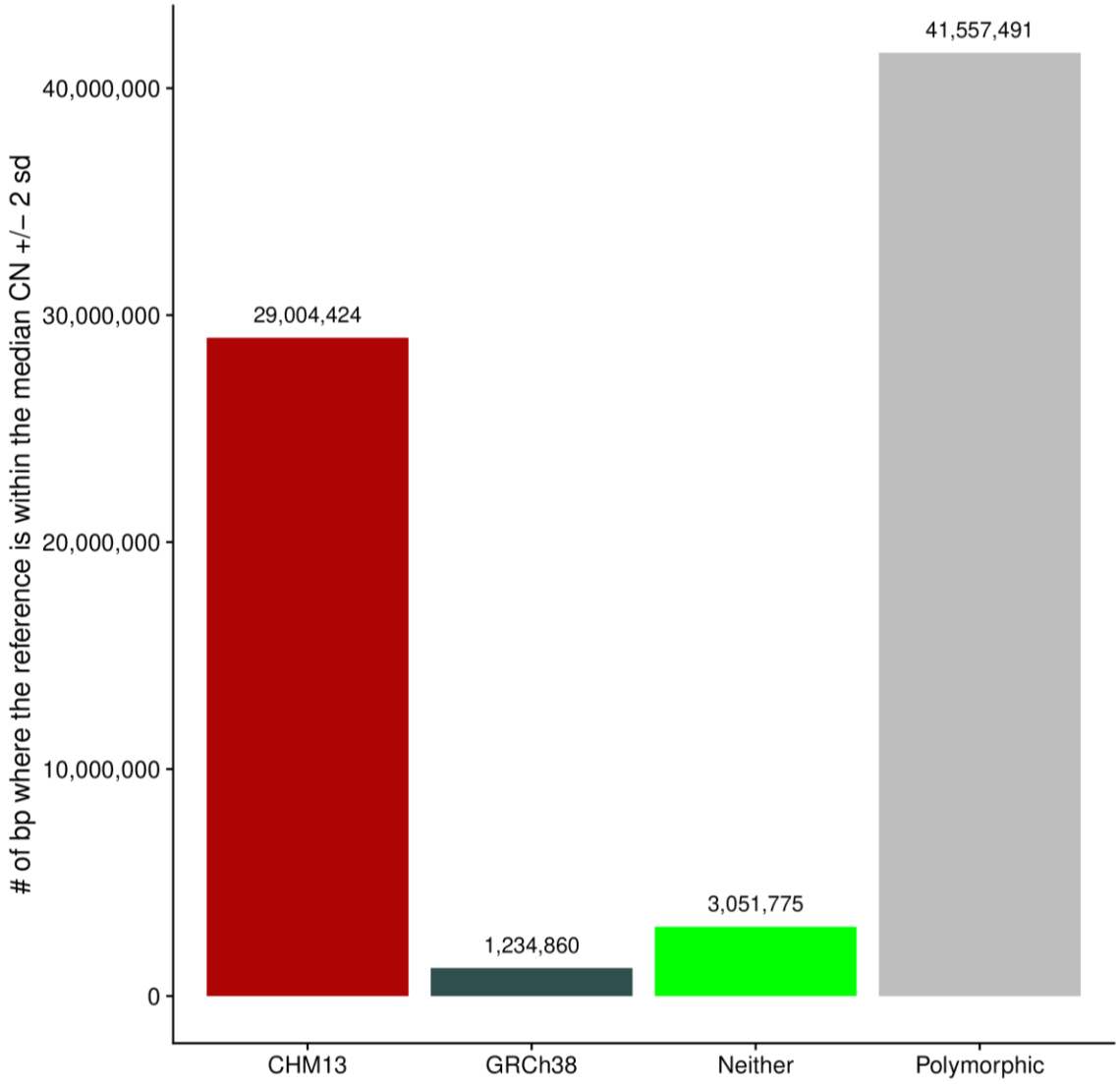

Fig. S10

Non-syntenic regions where the reference copy number reflects SGDP.
Copy number of SD regions that are previously unresolved or structurally different in T2T-CHM13 compared to GRCh38 and 268 individuals from the SGDP (*7*). The histogram shows the number in Mbp where the median sample copy number from SGDP was within two standard deviations (s.d.) of the given assembly [T2T-CHM13 (red), GRCh38 (blue), neither (green), or both (equal copy number)].

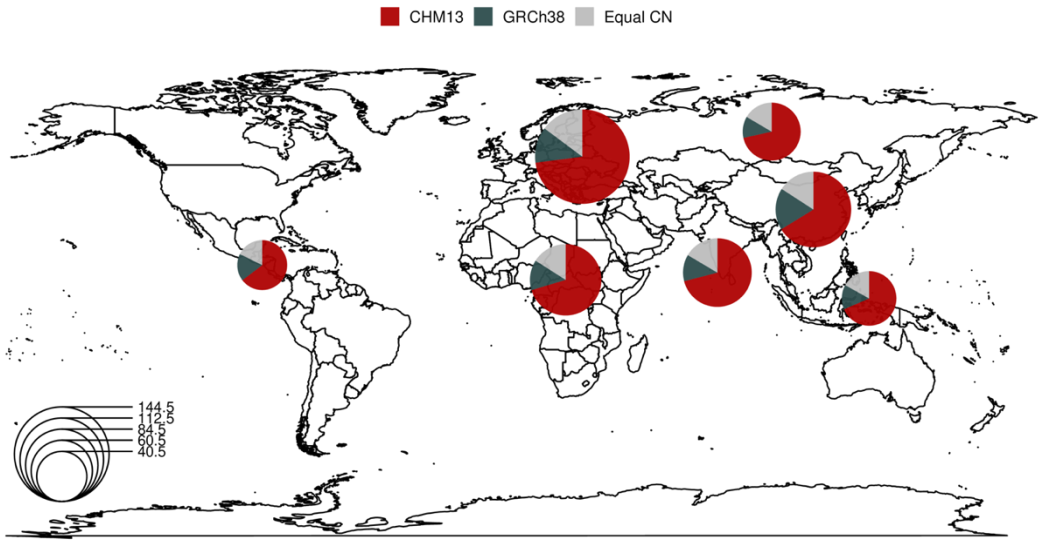

Fig. S11

Copy number representation across human superpopulations.
This figure indicates the reference genome which better represents human copy number based on Illumina read-depth genotyping across different superpopulations in the SGDP (n = 268). Individual pie charts show the relative fraction of SGDP samples summed across non-syntenic SD regions where the individual's genotype more closely matches the copy number of a reference. The colors show whether the copy number genotype can be better represented by T2T-CHM13 (red), GRCh38 (blue), or equally well by both (gray).

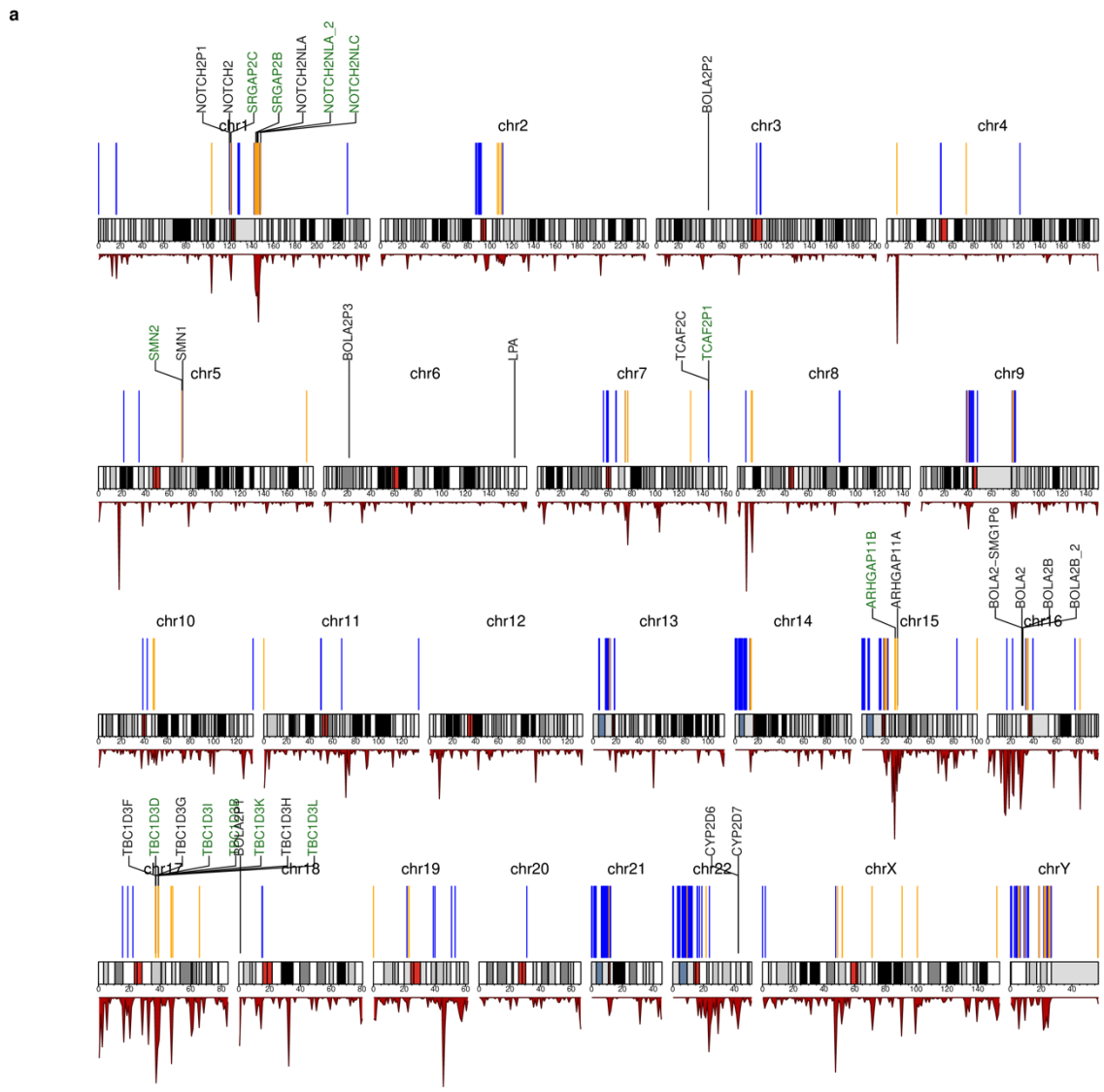

Fig. S12

Genic SD expansions in T2T-CHM13 relative to chimpanzee.
The blue (no genes) and orange (containing genes) peaks in the ideogram show regions of expansion in T2T-CHM13 v1.0 relative to the Clint_PTR assembly within SD space. The bottom panel shows the density of genic SDs in T2T-CHM13. The genes highlighted as biomedically or evolutionarily important loci are labeled and colored green if they are part of a human expansion.

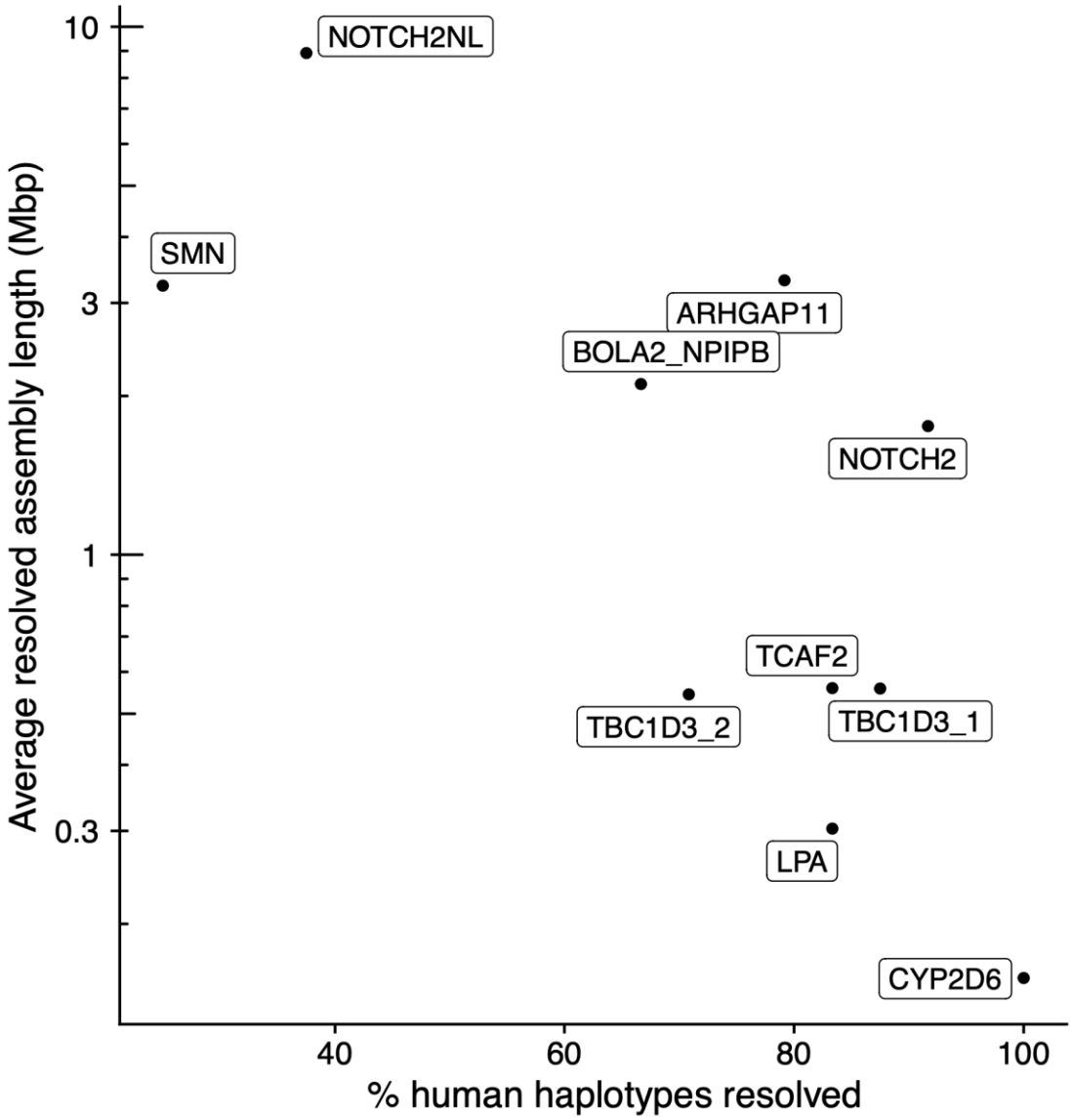

Fig. S13

Average locus length versus percent resolution in human haplotypes.
This figure shows the relationship between the percent of human haplotypes that are resolved using hifiasm and the average length of the SD locus for the 10 loci investigated in this study.

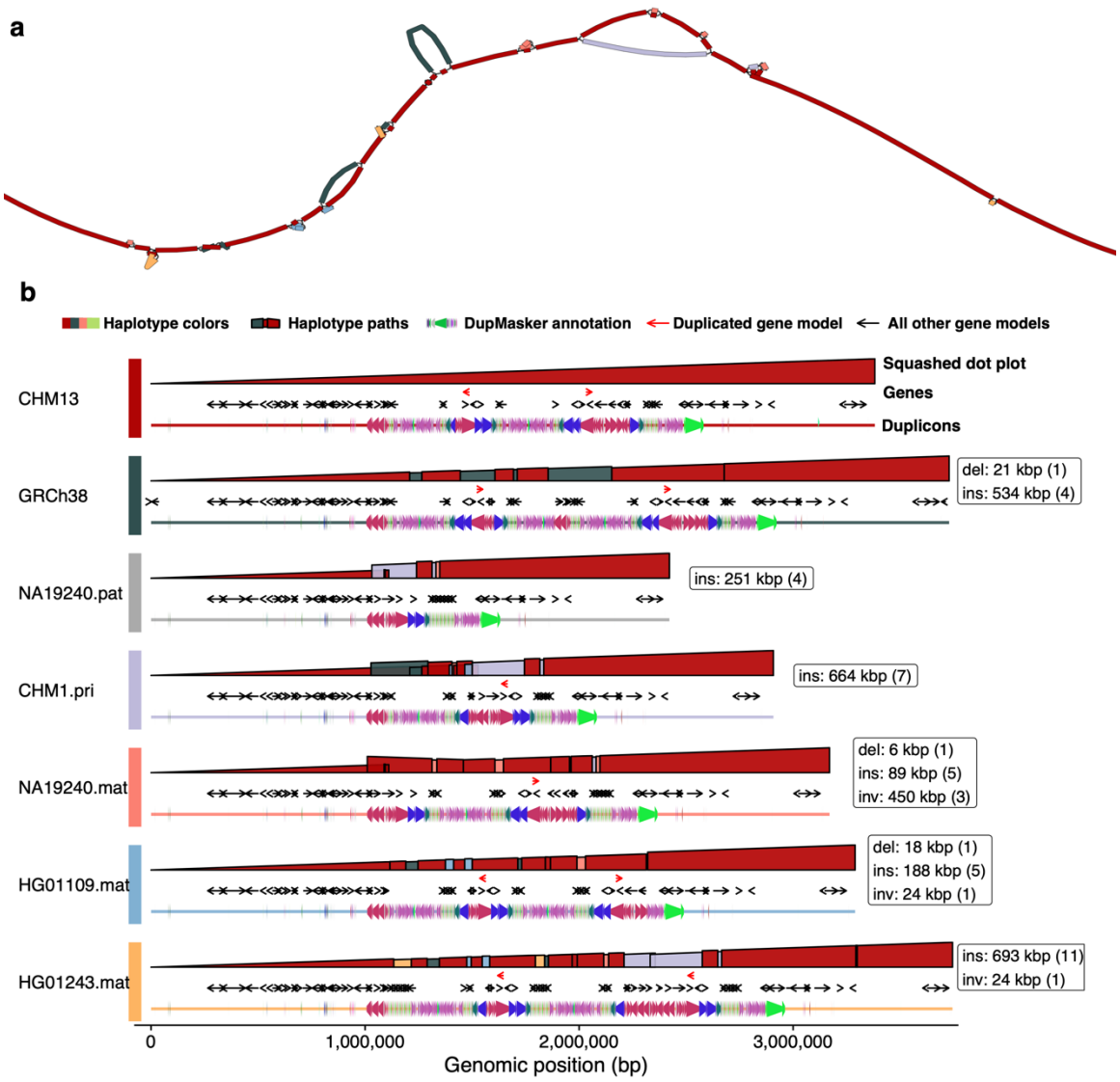

Fig. S14

Pangenome graph of *SMN*.
For a description of the elements within this figure, see fig. S7. The coordinates in T2T-CHM13 v1.0 chr5:69399944-72682017 and GRCh38 chr5:68527618-72250798.

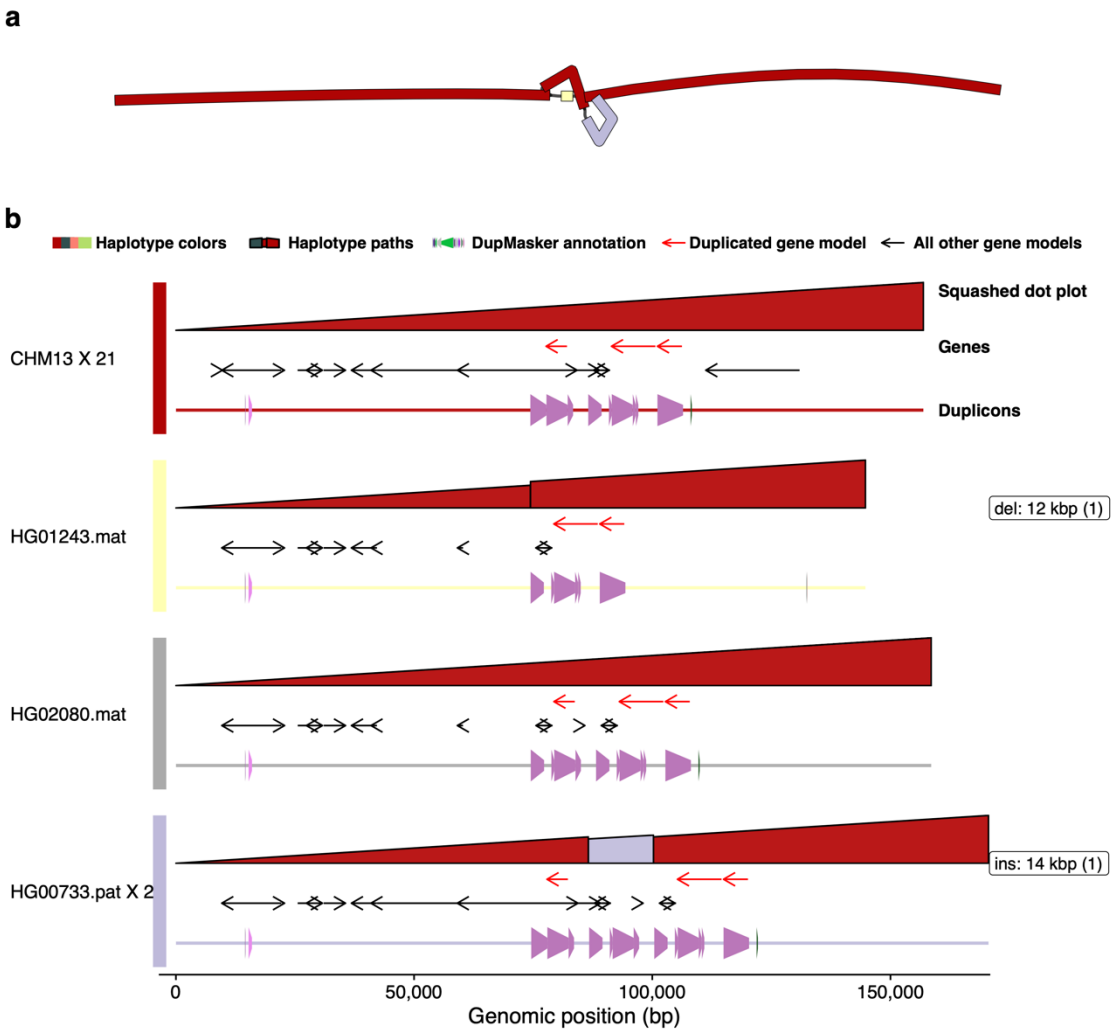

Fig. S15

Pangenome graph of *CYP2D6*.
For a description of the elements within this figure, see fig. S7. The coordinates in T2T-CHM13 v1.0 chr22:42,606,937-42,663,967 and GRCh38 chr22:42,048,703-42,205,690.

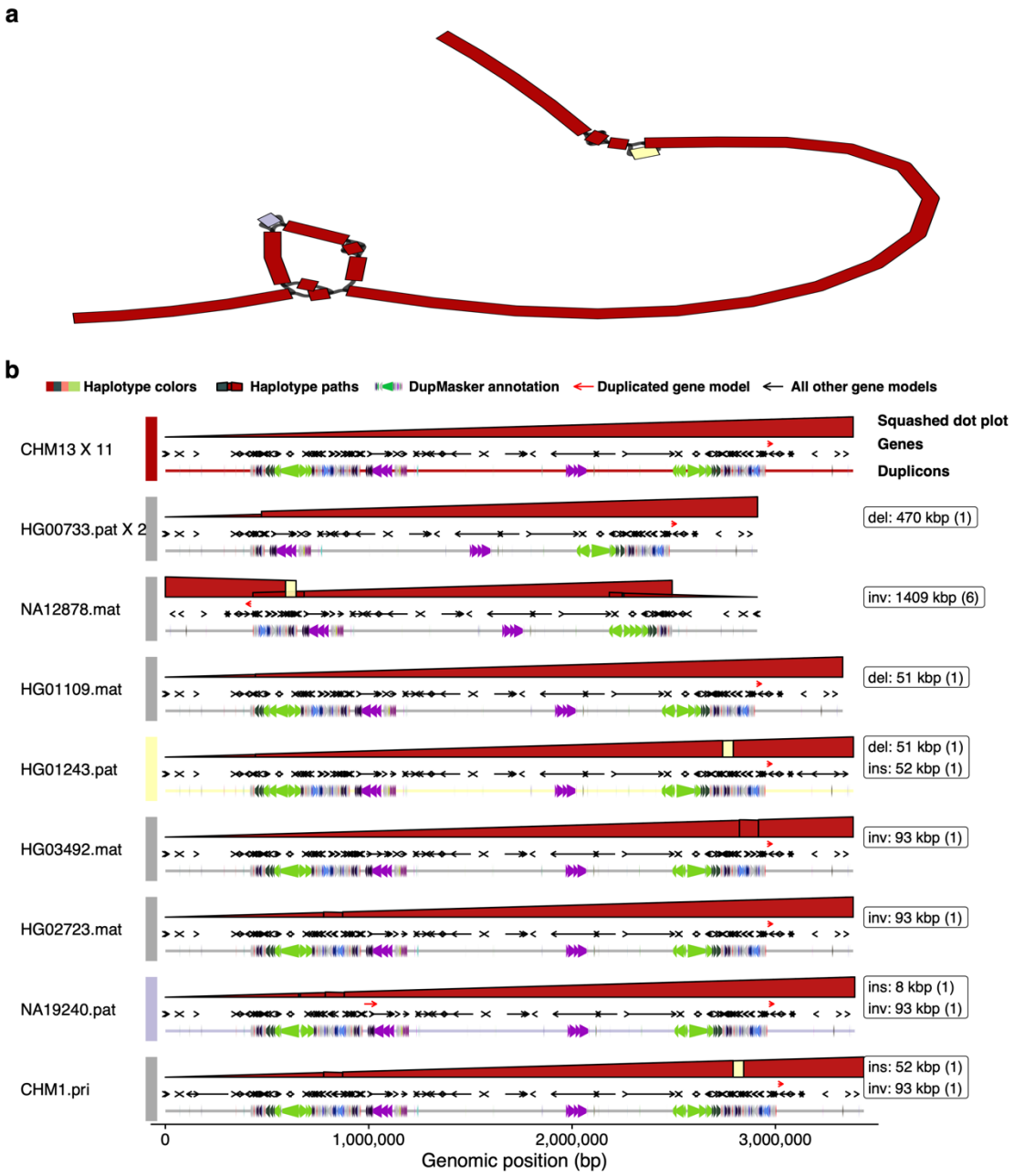

Fig. S16

Pangenome graph of *ARHGAP11*.
For a description of the elements within this figure, see fig. S7. The coordinates in T2T-CHM13 v1.0 chr15:28086550-31369052 and GRCh38 chr15:29660681-33037568.

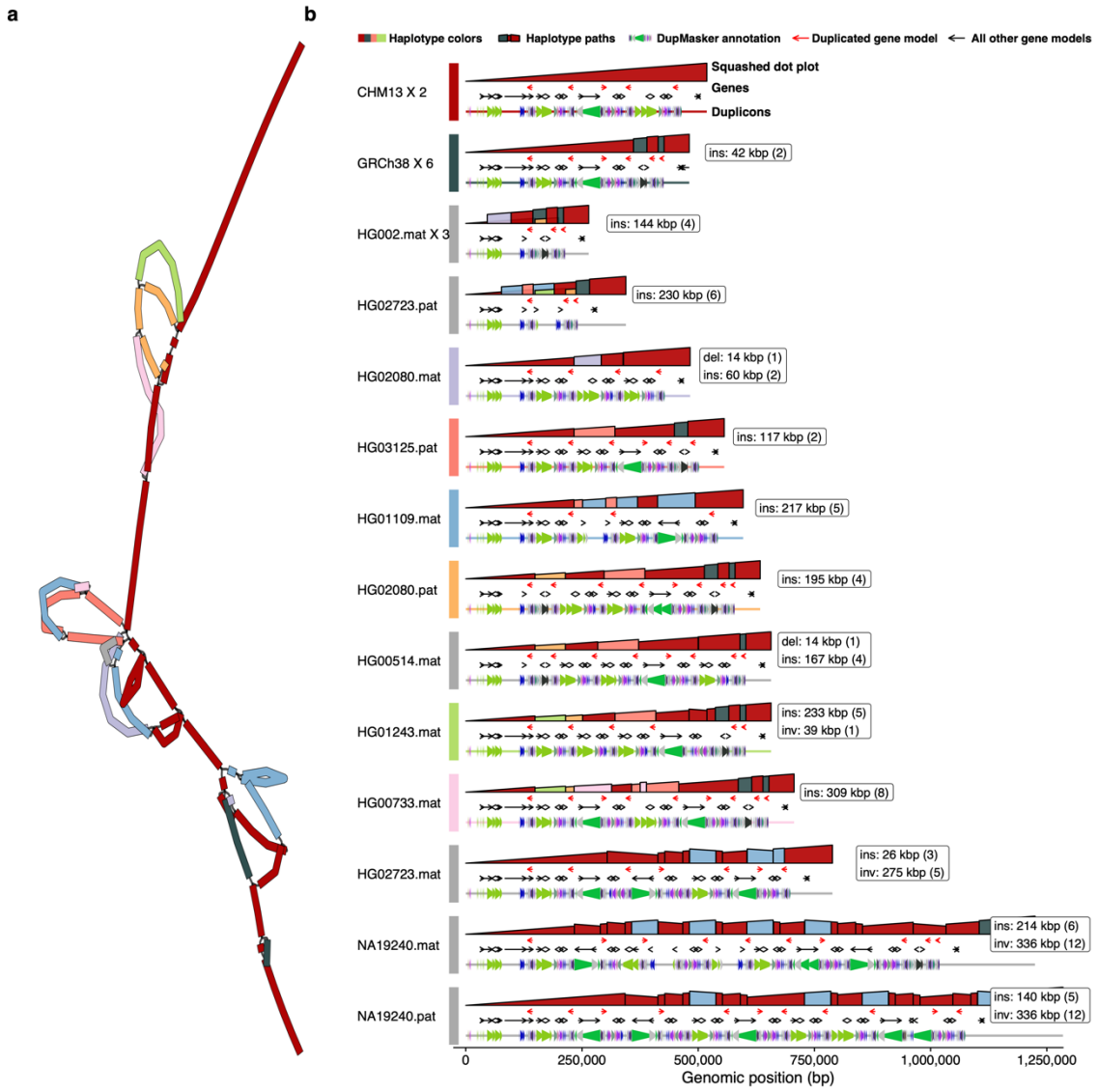

Fig. S17

Pangenome graph of *TBC1D3* expansion site one.
For a description of the elements within this figure, see fig. S7. The coordinates in T2T-CHM13 v1.0 chr17:37030899-37449510 and GRCh38 chr17:36032706-36513461.

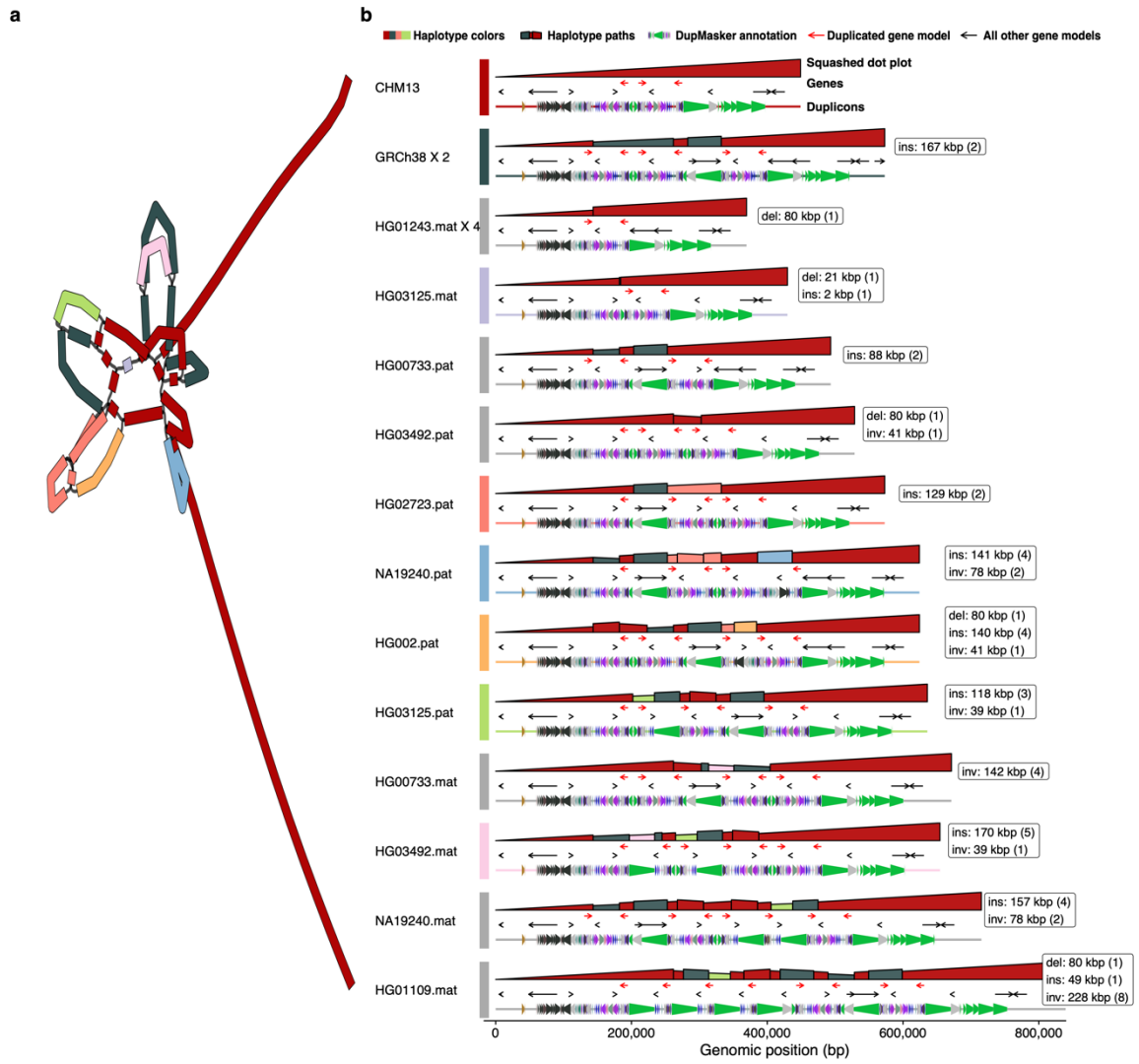

Fig. S18

Pangenome graph of *TBC1D3* expansion site two.
For a description of the elements within this figure, see fig. S7. The coordinates in T2T-CHM13 v1.0 chr17:38831091-39180264 and GRCh38 chr17:37793841-38367055.

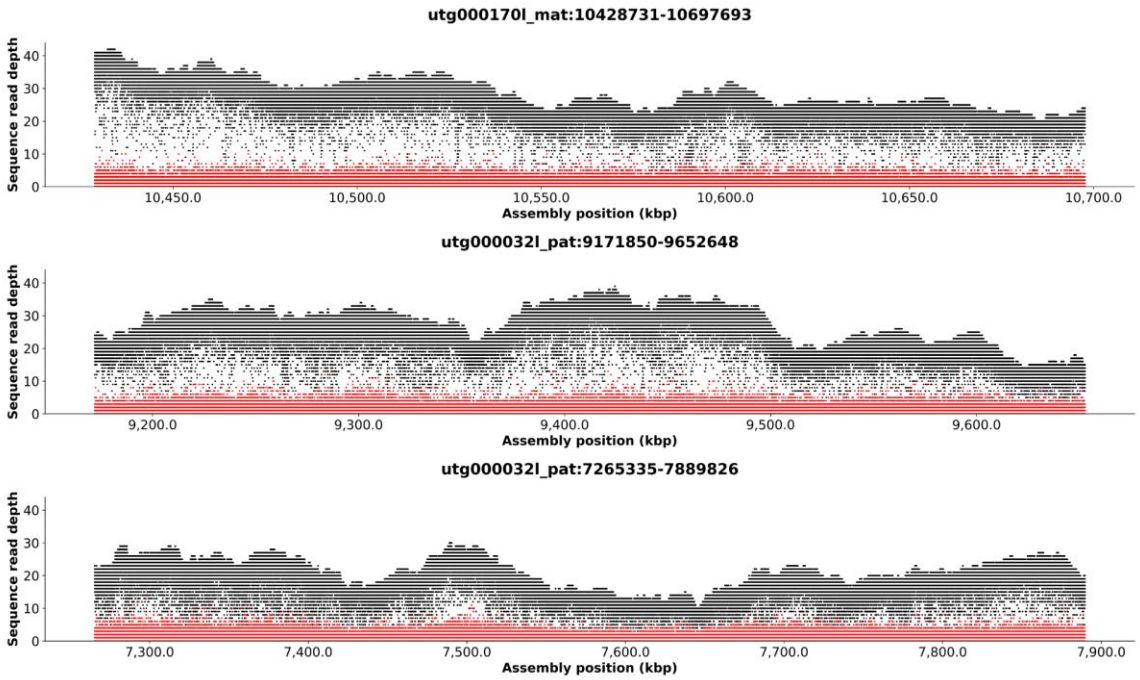

Fig. S19

Validation of assembly using ONT coverage over *TBC1D3* for HG002.
Ultra-long ONT coverage of HG002 across the maternal haplotype of *TBC1D3* expansion site one, and the coverage across the maternal and paternal haplotypes of *TBC1D3* expansion site two. Black dots show the coverage of the most frequent base at each genomic position and red dots show the coverage of the second most frequent base.

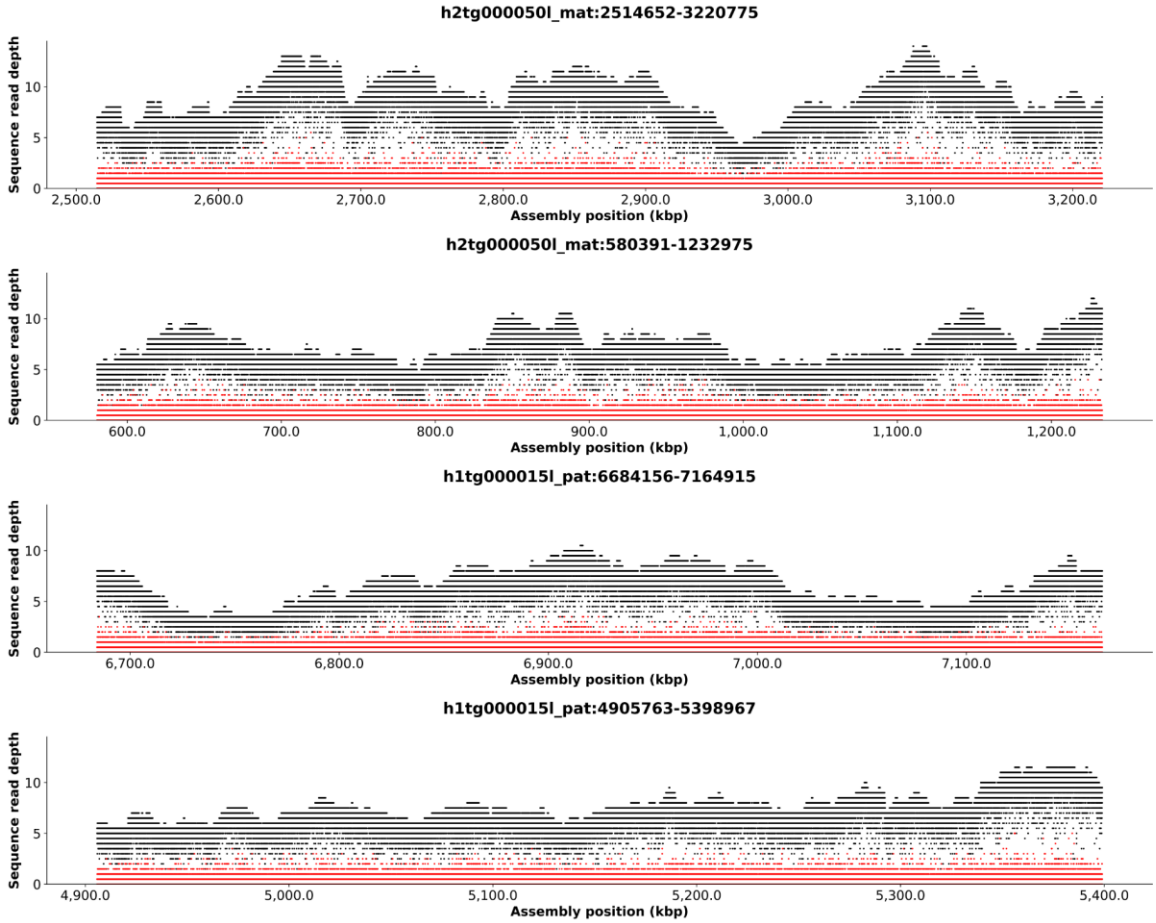

Fig. S20

Validation of assembly using ONT coverage over *TBC1D3* for HG00733.
Ultra-long ONT coverage of HG00733 across the maternal and paternal haplotypes of *TBC1D3* expansion site one, and the coverage across the maternal and paternal haplotypes of *TBC1D3* expansion site two. Black dots show the coverage of the most frequent base at each genomic position and red dots show the coverage of the second most frequent base.

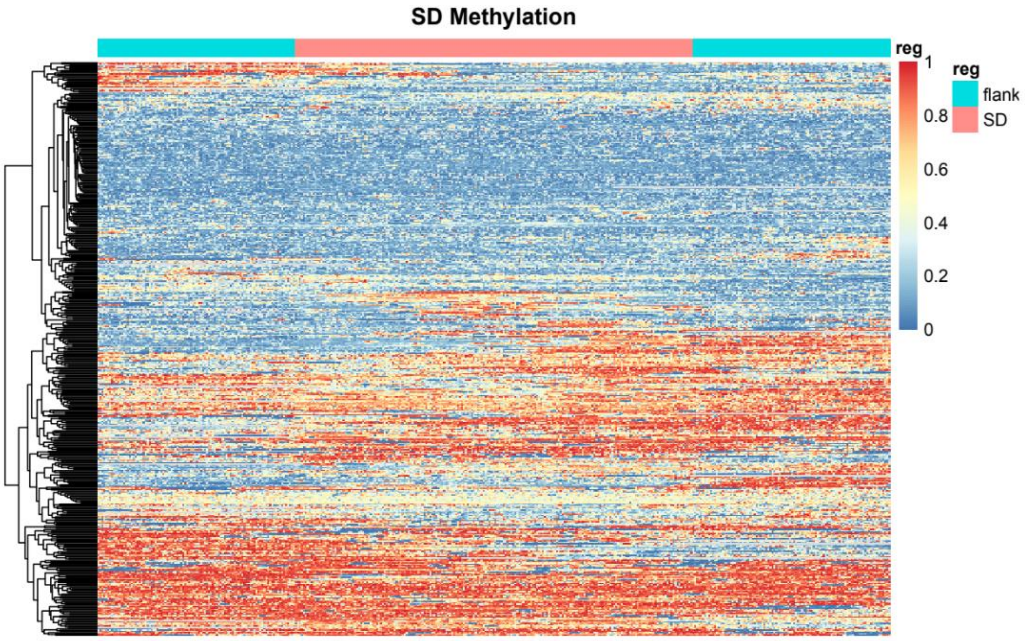

Fig. S21

Clustering of methylation status in SD blocks.
Heatmap of CpG methylation of all SD blocks with at least 50 kbp of flanking sequence clustered using the "pheatmap" package in R. The horizontal annotation shows in cyan the 50 kbp of unique flanking sequence and red the SD block.

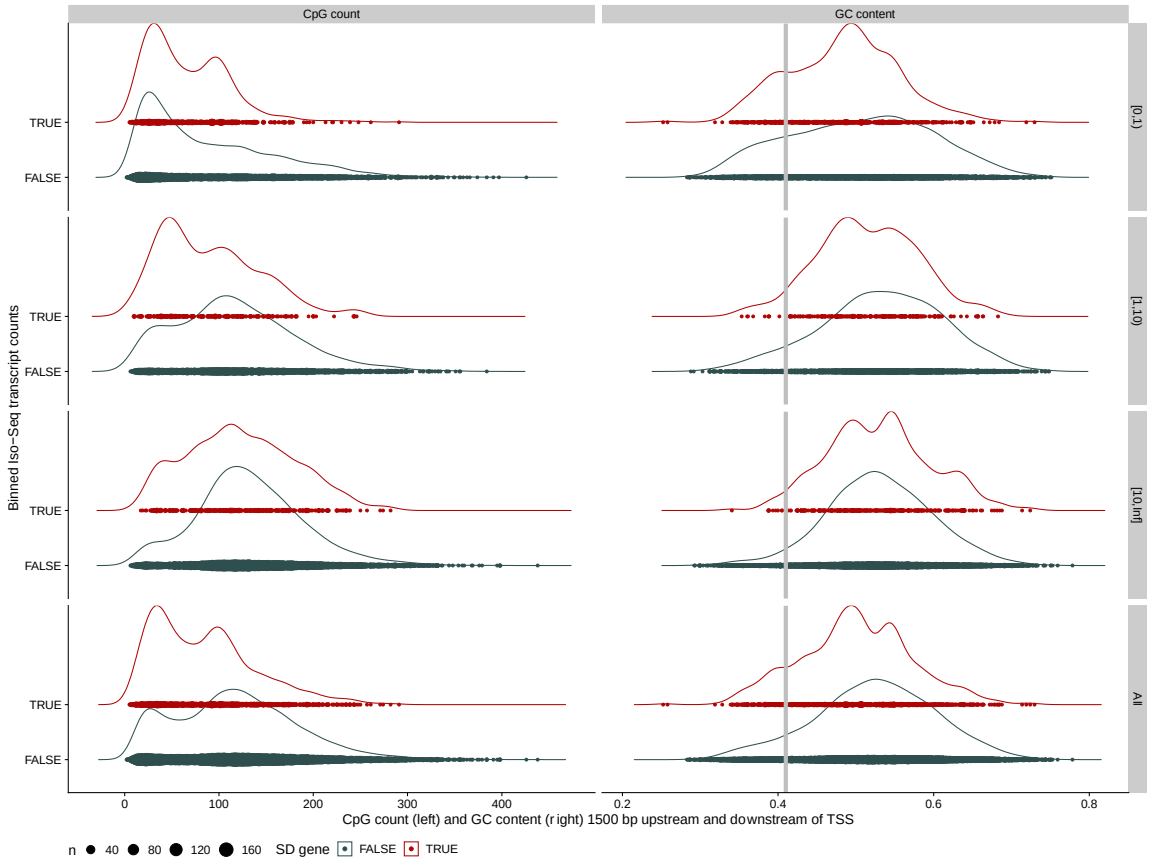

Fig. S22

CpG and GC content within 1,500 bp of the transcription start site (TSS). Shown are the density of the number of CpGs within +/-1,500 bp of the TSS (left), and the density of bases that are G or C within +/-1,500 bp of the TSS (right), both stratified by the level of Iso-Seq transcription (vertical positioning) and SD content (color).

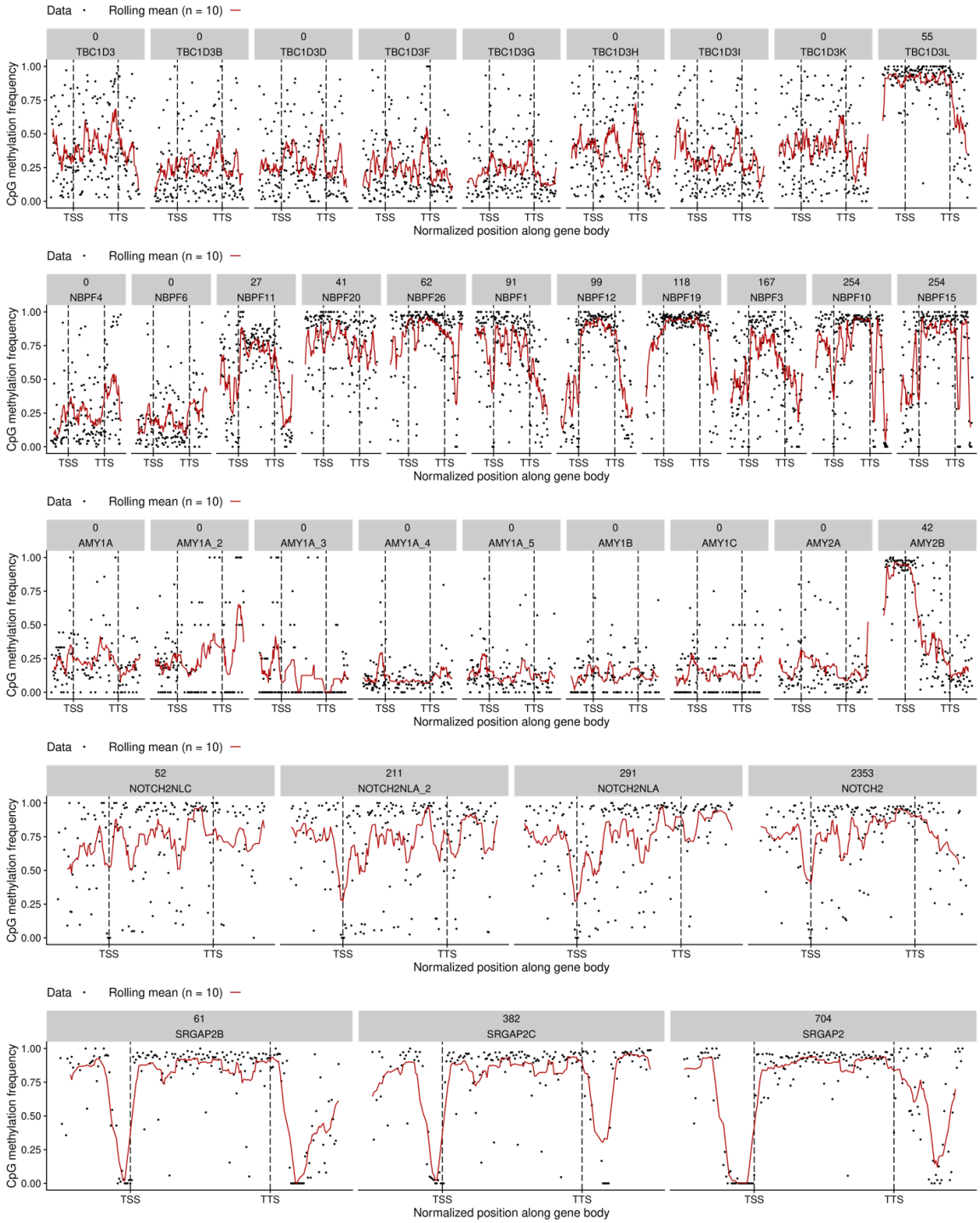

Fig. S23

Methylation and transcription levels across multi-copy gene families.
Shown are the methylation signals across recently duplicated gene families in T2T-CHM13 (*TBC1D3*, *NBPF*, *AMY*, *NOTCH2*, and *SRGAP2)*. Black points represent individual methylation calls, and the red line is a rolling mean across 10 methylation sites. The number of CHM13 Iso-Seq transcripts and the gene name are indicated in gray.

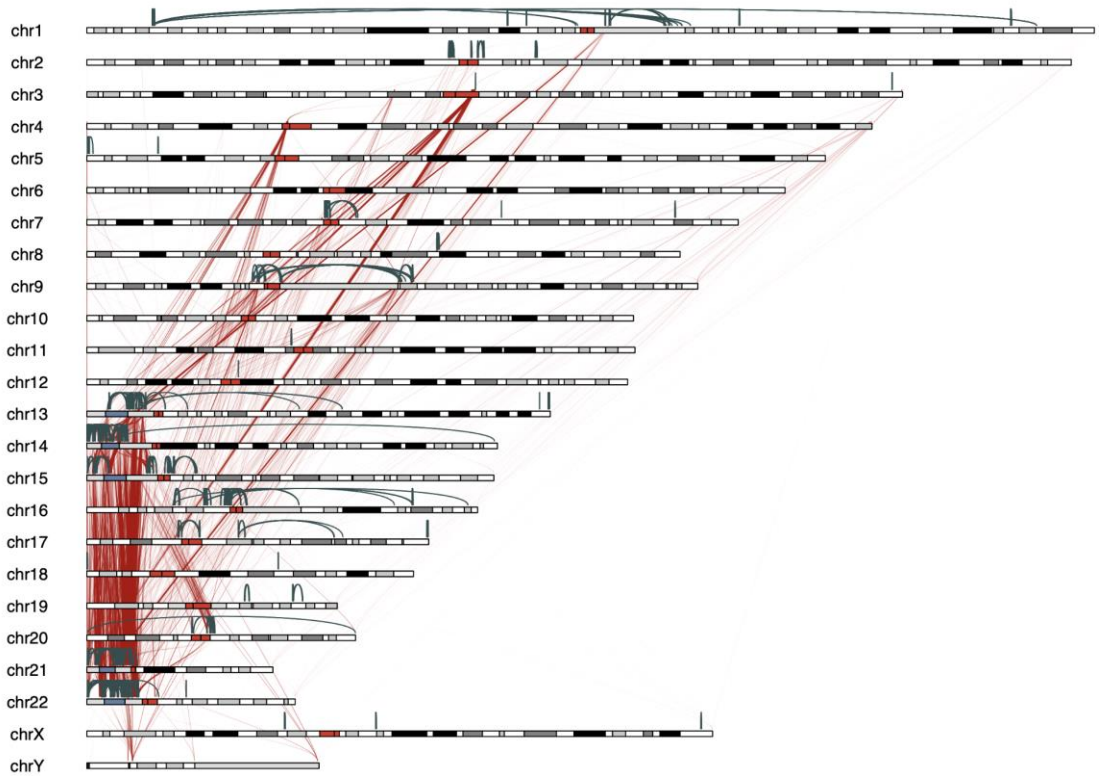

Fig. S24

Pairwise alignments of SDs that are previously-unresolved-by-content (unique to T2T).

Shown are the inter (red) and intrachromosomal (blue) SDs that have no paralogous sequence match in GRCh38.

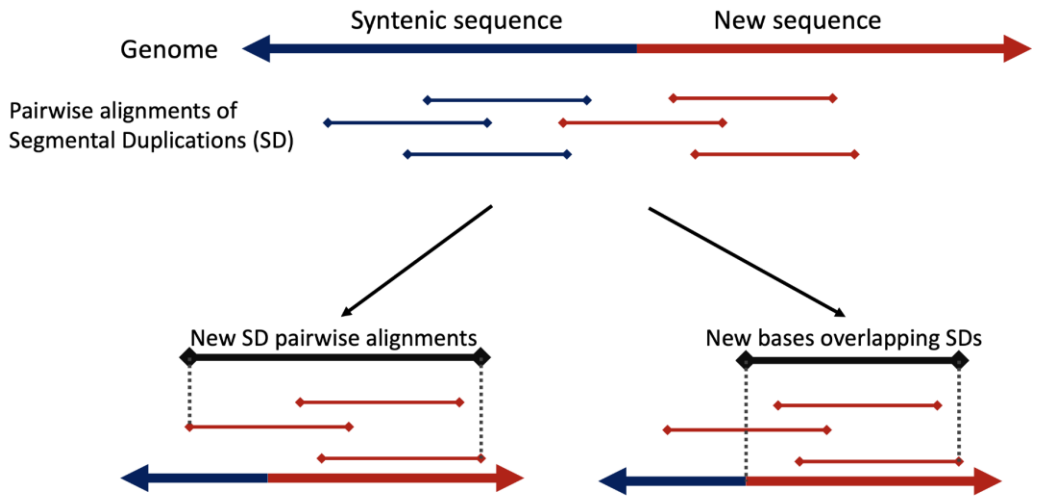

Fig. S25

Previously unresolved SDs versus previously unresolved sequences that overlap SDs.

This figure illustrates the different methods used by this work (left) and Nurk *et al.* (right) for counting previously unresolved SD bases *(20)*.

List of tables available as Supplementary Material online

Table S1: Segmental duplications shared between acrocentric and non-acrocentric chromosomes.
Listed are the non-acrocentric SD regions with the most alignments back to the acrocentric short arms, including the number of average percent identity of these alignments.

Table S2: Duplicon content of T2T-CHM13 and GRCh38 as determined by DupMasker.
A comparison of the duplicon content of T2T-CHM13 and GRCh38 as annotated by DupMasker. Includes the total length, number, GC content, and genes for each duplicon in the two assemblies.

Table S3: Genomic ranges in T2T-CHM13 with the largest increases in SDs compared to GRCh38.
A comparison of the SD content in 5 Mbp windows between T2T-CHM13 and GRCh38. The list is organized from largest to smallest difference in the number of SDs within the window.

Table S4: FISH results from acrocentric fosmid probes.
Summary of FISH validation experiments done on acrocentric duplications. The table shows the assembly predictions (colors) of where there would be FISH signals and the where actual FISH signals were observed (text) for each of the nine probes and six samples.

Table S5: Summary of variation in unique versus duplicated sequence.
Summary statistics on the number of variants (SNVs, insertions, and deletions) seen in unique regions of the genome compared to the MHC region, chrX, and SD sequence. Results were only tabulated in regions with at least 1 Mbp of synteny between the two references.

Table S6: Copy number of genes in SGDP that are in non-syntenic regions of T2T-CHM13.
This table shows the number of SGDP samples that agree with either the T2T-CHM13 copy number or the GRCh38 copy number for genic SDs in non-syntenic regions between the two references.

Table S7: Core duplicon copy number in AFR vs. non-AFR.
This table contains the average copy number for several core-duplicon gene families in Africans and non-Africans.

Table S8: Targeted loci of biomedical or evolutionary importance.
Regions targeted for assembly in additional humans and nonhuman primates. Includes the number of successfully resolved haplotypes as well as average assembly statistics for each locus.

Table S9: Genic SD expansions in T2T-CHM13 relative to chimpanzee.
Genic SD regions in T2T-CHM13 where there are at least 50 kbp of sequence with no correspondence to a HiFi assembly of a chimpanzee genome (Clint PTR) indicating a likely human-specific expansion.

Table S10: Variation in targeted loci of biomedical or evolutionary importance.
A summary of a structural variation identified by minigraph in the 10 loci targeted for their biomedical or evolutionary importance and assembled with HiFi in multiple humans and nonhuman primates.

Table S11: Intersection of previously unresolved or non-syntenic genes with Iso-Seq data.
A complete list of all previously unresolved or non-syntenic SD genes identified by Liftoff and the number of Iso-Seq transcripts that support these gene models. Human Iso-Seq data from 59 different experiments were included (table S13).

Table S12: Samples used in the variation analysis of biomedical and evolutionary loci.
List of human and nonhuman primate samples used in the assembly of the 10 loci of biomedical or evolutionary importance including the population, superpopulation, and species where appropriate.

Table S13: Accessions for Iso-Seq data.
Accessions for all Iso-Seq data used in generating support for the additional candidate gene models in T2T-CHM13.

Table S14: *TBC1D3* short-read copy number estimates.
Short-read copy number estimates for *TBC1D3* in all SGDP samples and nonhuman primates. This data is displayed in Figure 4.