The complete sequence of a human Y chromosome

Arang Rhie[1†], Sergey Nurk[1†‡], Monika Cechova[2,3†], Savannah J. Hoyt[4†], Dylan J. Taylor[5†], Nicolas Altemose[6], Paul W. Hook[7], Sergey Koren[1], Mikko Rautiainen[1], Ivan A. Alexandrov[8,9§], Jamie Allen[10], Mobin Asri[11], Andrey V. Bzikadze[12], Nae-Chyun Chen[13], Chen-Shan Chin[14,15], Mark Diekhans[11], Paul Flicek[10,16], Giulio Formenti[17], Arkarachai Fungtammasan[18], Carlos Garcia Giron[10], Erik Garrison[19], Ariel Gershman[7], Jennifer L. Gerton[20,21], Patrick G.S. Grady[4], Andrea Guarracino[22,19], Leanne Haggerty[10], Reza Halabian[23], Nancy F. Hansen[24,1], Robert Harris[25], Gabrielle A. Hartley[4], William T. Harvey[26], Marina Haukness[11], Jakob Heinz[7], Thibaut Hourlier[10], Robert M. Hubley[27], Sarah E. Hunt[10], Stephen Hwang[28], Miten Jain[29], Rupesh K. Kesharwani[30], Alexandra P. Lewis[26], Heng Li[31,32], Glennis A. Logsdon[26], Julian K. Lucas[3,11], Wojciech Makalowski[23], Christopher Markovic[33], Fergal J. Martin[10], Ann M. Mc Cartney[1], Rajiv C. McCoy[5], Jennifer McDaniel[34], Brandy M. McNulty[3,11], Paul Medvedev[35,36,37], Alla Mikheenko[9,38], Katherine M. Munson[26], Terence D. Murphy[39], Hugh E. Olsen[3,11], Nathan D. Olson[34], Luis F. Paulin[30], David Porubsky[26], Tamara Potapova[20], Fedor Ryabov[40], Steven L. Salzberg[41], Michael E.G. Sauria[5], Fritz J. Sedlazeck[30,42], Kishwar Shafin[43], Valery A. Shepelev[44#], Alaina Shumate[7], Jessica M. Storer[27], Likhitha Surapaneni[10], Angela M. Taravella Oill[45], Françoise Thibaud-Nissen[39], Winston Timp[7], Marta Tomaszkiewicz[25,46], Mitchell R. Vollger[26], Brian P. Walenz[1], Allison C. Watwood[25], Matthias H. Weissensteiner[25], Aaron M. Wenger[47], Melissa A. Wilson[45], Samantha Zarate[13], Yiming Zhu[30], Justin M. Zook[34], Evan E. Eichler[26,48], Rachel J. O'Neill[4,49,50], Michael C. Schatz[13,5], Karen H. Miga[3,11], Kateryna D. Makova[25], Adam M. Phillippy[1*]

1. Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
2. Faculty of Informatics, Masaryk University, Brno, Czech Republic
3. Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, USA
4. Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA
5. Department of Biology, Johns Hopkins University, Baltimore, MD, USA
6. Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA
7. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA
8. Federal Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia
9. Center for Algorithmic Biotechnology, Saint Petersburg State University, Russia
10. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
11. UC Santa Cruz Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, USA
12. Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, CA, USA
13. Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
14. GeneDX Holdings Corp, Stamford, CT, USA
15. Foundation of Biological Data Science, Belmont, CA, USA
16. Department of Genetics, University of Cambridge, Cambridge, United Kingdom
17. The Rockefeller University, New York, NY, USA
18. DNAnexus, Inc., Mountain View, CA, USA
19. Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, Tennessee, USA

20. Stowers Institute for Medical Research, Kansas City, MO, USA
21. University of Kansas Medical Center, Kansas City, MO, USA
22. Genomics Research Centre, Human Technopole, Viale Rita Levi-Montalcini 1, Milan, Italy
23. Institute of Bioinformatics, Faculty of Medicine, University of Münster, Münster, Germany
24. Cancer Genetics and Comparative Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
25. Department of Biology, Pennsylvania State University, University Park, PA, USA
26. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
27. Institute for Systems Biology, Seattle, WA, USA
28. XDBio Program, Johns Hopkins University, Baltimore, MD, USA
29. Department of Bioengineering, Department of Physics, Northeastern University, Boston, MA, USA
30. Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX, USA
31. Department of data sciences, Dana-Farber Cancer Institute, Boston, MA, USA
32. Department of biomedical informatics, Harvard Medical School, Boston, MA, USA
33. Genome Technology Access Center at the McDonnell Genome Institute (GTAC@MGI) Washington University, St. Louis, MO, USA
34. Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
35. Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA, USA
36. Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA, USA
37. Center for Computational Biology and Bioinformatics, Pennsylvania State University, University Park, PA, USA
38. UCL Queen Square Institute of Neurology, UCL, London, UK
39. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
40. Masters Program in National Research University Higher School of Economics, Moscow, Russia
41. Departments of Biomedical Engineering, Computer Science, and Biostatistics, Johns Hopkins University, Baltimore, MD
42. Department of Computer Science, Rice University, Houston, TX, USA
43. Google Inc, 1600 Amphitheatre Pkwy, Mountain View, CA, USA
44. Institute of Molecular Genetics, Moscow, Russia
45. Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, AZ, USA
46. Department of Biomedical Engineering, Pennsylvania State University, State College, PA, USA
47. Pacific Biosciences, Menlo Park, CA, USA
48. Investigator, Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA
49. Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA
50. Department of Genetics and Genome Sciences, UConn Health, Farmington, CT, USA

[†] These authors contributed equally
[‡] Present address: Oxford Nanopore Technologies Inc., Oxford, United Kingdom
[§] Present address: Department of Anatomy and Anthropology and Department of Human Molecular Genetics and Biochemistry, Sackler Faculty of Medicine, Tel Aviv University, Israel
[#] Retired
[*] Corresponding author. Email: adam.phillippy@nih.gov

The human Y chromosome has been notoriously difficult to sequence and assemble because of its complex repeat structure including long palindromes, tandem repeats, and segmental duplications[1–3]. As a result, more than half of the Y chromosome is missing from the GRCh38 reference sequence and it remains the last human chromosome to be finished[4,5]. Here, the Telomere-to-Telomere (T2T) consortium presents the complete 62,460,029 base pair sequence of a human Y chromosome from the HG002 genome (T2T-Y) that corrects multiple errors in GRCh38-Y and adds over 30 million base pairs of sequence to the reference, revealing the complete ampliconic structures of *TSPY*, *DAZ*, and *RBMY* gene families; 41 additional protein-coding genes, mostly from the *TSPY* family; and an alternating pattern of human satellite 1 and 3 blocks in the heterochromatic Yq12 region. We have combined T2T-Y with a prior assembly of the CHM13 genome[4] and mapped available population variation, clinical variants, and functional genomics data to produce a complete and comprehensive reference sequence for all 24 human chromosomes.

The human Y chromosome plays critical roles in fertility and hosts genes important for spermatogenesis, as well as *SRY*, the mammalian sex-determining locus[6]. However, in the human reference genome, GRCh38, the Y chromosome remains the most incomplete chromosome, with >50% of bases represented by gaps. These multi-megabase gaps have persisted for decades and represent sequences flanking the endogenous model centromere, parts of the ampliconic regions, and large heterochromatic regions. The architecture of the Y chromosome, specifically the presence of large tandemly arrayed and inverted repeats (i.e. palindromes)[1], makes assembly difficult and hinders the study of rearrangements, inversions, duplications, and deletions in several critical regions such as AZFa, AZFb, and AZFc (azoospermia factor), which are linked to clinical phenotypes, including infertility[7].

Following the first complete assemblies of chromosomes X[8] and 8[9], the Telomere-to-Telomere (T2T) consortium successfully assembled all chromosomes of the CHM13 cell line[4]. This first complete human genome assembly was enabled by innovative technological improvements in generating Pacific Biosciences (PacBio) high-fidelity reads (HiFi)[10] and Oxford Nanopore ultra-long reads (ONT)[11], the development of better assembly algorithms for utilizing HiFi reads and generating assembly graphs[12], the use of ONT reads for better resolving the graph[13], new methods for validating and polishing[14–18], and a coordinated curation effort to finish the assembly. Having been derived from a complete hydatidiform mole, CHM13 has a 46,XX karyotype but is almost entirely homozygous. This simplified assembly of its genome, but prevented assembly of a Y chromosome.

In parallel, with the goal of including broader genomic diversity across populations[19], the Human Pangenome Reference Consortium (HPRC) has evaluated various methods for generating high-quality diploid genome assemblies[20] using a well characterized human genome, HG002, which has been previously assembled[21] and is commonly used for benchmarking by the Genome in a Bottle consortium[22]. Using this rich set of data, and integrating the lessons learned from assembling CHM13, we successfully reconstructed the complete sequence of the HG002 Y chromosome, hereafter referred to as T2T-Y.

Here we analyze the composition of the newly assembled pseudoautosomal regions (PARs), ampliconic and palindromic sequences, centromeric satellites, and q-arm heterochromatin of a complete Y chromosome. We have annotated T2T-Y and combined it with the prior T2T-CHM13 assembly to form a new, complete reference for all human chromosomes, T2T-CHM13+Y. To enable the use of this new reference sequence, we have lifted over available variation datasets from ClinVar[23], GWAS[24], dbSNP[25] and gnomAD[26]. In addition, we have recalled variants from 1000 Genomes Project (1KGP)[27] and Simons Genome Diversity Panel (SGDP)[28] data, as well as epigenetic profiles from ENCODE data[29]. These experiments demonstrate improved mappability and variant calling for XY individuals when using T2T-Y as a reference.

Assembly and validation of T2T-Y

Assembly of the HG002 Y chromosome followed the strategy used for the T2T-CHM13 genome[4] (**Supplementary Table 1** and **Supplementary Fig. 1**). We used PacBio HiFi reads (60× haploid genome coverage) and ONT ultra-long reads (90× in reads longer than 100 kb) generated from HG002. An assembly string graph was first constructed for the whole HG002 genome using PacBio HiFi reads. The ChrX and ChrY string graph components shared connections to one another at the PARs, but to no other chromosomes in the genome and could be independently analyzed (**Extended Data Fig. 1a**). The remaining tangles in these XY subgraphs were resolved using ONT reads (**Extended Data Fig. 1b**). ChrX and ChrY chromosomal walks were identified using haplotype-specific k-mers from parental Illumina reads (**Extended Data Fig. 1c**), and a consensus sequence was computed for each. PAR1 was enriched for GA-microsatellites, which reduced HiFi coverage in this region and led to a more fragmented graph (due to a known HiFi sequencing bias[12]). These gaps were manually patched using a *de novo* assembly of trio-binned parental ONT reads[14].

The ChrY draft assembly was further polished and validated using sequencing reads from Illumina (66× haploid genome coverage), HiFi (84×), and ONT (250×). During four rounds of polishing, 1,520 small and 10 large (>50 base) errors were detected and corrected (**Extended Data Fig. 2a**). Conservatively filtered long-read alignments identified two potential assembly issues remaining in the satellite (HSat) arrays around positions 40 Mb and 59.1 Mb, and Strand-seq[30,31] identified one inversion error within palindromic sequence P5 around position 18.8 Mb (**Extended Data Fig. 2b-c, Supplementary Table 2, Supplementary Figs. 2-4**). The validation signal at the two HSat positions was ambiguous, and the P5 inversion appears as a true recurrent inversion[32], so these regions were noted but left uncorrected in this release. The remaining sequences showed no signs of collapse or false duplication, with even HiFi coverage (mean 39.3x ± SD 12.5 on ChrXY) except for regions associated with known sequencing biases[17], all of which had supporting ONT coverage (reads over 25 kb, mean 78.1x ± SD 13.6 on ChrXY). The base error is estimated as less than 1 error per 10 Mb (Phred Q73.8, **Supplementary Table 3**). Mapped HiFi and ONT reads from the paternal HG003 genome are also consistent with the HG002 T2T-Y assembly, suggesting no large, structural variants were introduced during cell line immortalization and culture (**Supplementary Fig 5**).

The resulting T2T-Y assembly is 62,460,029 bases in length with no gaps or model sequences, revealing the previously uncharacterized ~30 Mb of sequence within the heterochromatic region

of the q-arm (**Table 1**). In comparison, ChrY in the human reference genome (GRCh38-Y) consists of two sequences, with the longer sequence totaling 57.2 Mb (NC_000024.10), for which 53.8% (30.8 Mb) of the bases are unresolved gaps. The shorter GRCh38-Y sequence (NT_187395.1) is 37.2 kb in length, not placed in the primary Y assembly, and has been omitted from most prior genomic studies. The PAR1 (2.77 Mb) and PAR2 (329.5 kb) sequences in GRCh38-Y are duplicated from ChrX rather than assembled *de novo*, and the centromere is represented by a 227 kb model sequence. Direct sequence comparison between T2T-Y and GRCh38-Y yields an average sequence identity of ~99.8% in the alignable regions, but with multiple structural differences including an incorrectly oriented centromere model for GRCh38-Y (**Fig. 1** and **Extended Data Fig. 3**). We identified the Y-chromosome haplogroup of HG002 as J-L816 (J1) and that of GRCh38 as R-L20 (R1b). These haplogroups are most commonly found among Ashkenazi Jews[33] and Europeans[34], respectively, consistent with the established ancestry of these genomes. T2T-Y was combined with the T2T-CHM13v1.1 assembly to create a new Y-bearing reference, T2T-CHM13v2.0, referred to here as T2T-CHM13+Y.

Comprehensive annotation of the Y

Gene annotation

We annotated T2T-CHM13+Y by mapping RefSeq (v110) and GENCODE (v35) annotations from GRCh38 and performed hand-curation of the ampliconic gene arrays (**Fig. 1** and **Supplementary Table 4-5**). NCBI RefSeq and EBI Ensembl generated additional *de novo* annotations using HG002 Iso-Seq transcriptomes from B-Lymphocyte and induced pluripotent stem cell (iPS) lines as well as tissue-specific expression data from other publicly available sources (**Supplementary Table 1, Supplementary Fig. 6-7**).

Our annotation of T2T-Y totals 693 genes and 883 transcripts, of which 106 genes (488 transcripts) are predicted to be protein-coding (**Table 1** and **Supplementary Table 4**). In addition to containing all genes annotated in GRCh38-Y, T2T-Y contains an additional 110 genes, among which 41 are predicted to be protein coding. The majority of these protein-coding genes (38 of 41) are additional copies of *TSPY*, one of the nine ampliconic gene families, filling the corresponding gap in GRCh38-Y (**Table 1**). The annotated ampliconic gene copies in T2T-Y were largely concordant with copy numbers estimated from Illumina reads and droplet digital PCR (ddPCR)[37], confirming the accurate copy number representation of the ampliconic genes in T2T-Y (**Supplementary Table 6-9**). RNA-Seq data confirmed expression of the annotated ampliconic genes in testis[38]. Only six genes differed in their annotation between GRCh38-Y and T2T-Y, due to presumed Y haplogroup differences (**Supplementary Table 10**).

Repeat annotation

We generated comprehensive repeat annotations, incorporating repeat models previously updated with CHM13[39], as well as 29 previously unknown repeats identified in T2T-Y (**Extended Data Fig. 4a, Supplementary Table 11**). The newly added sequences increased the percentage of identifiable repeats on the Y chromosome from 66.3% to 84.9%, or 17.5 Mb of non-N bases in GRCh38-Y compared to 53 Mb of bases in T2T-Y (**Table 1, Supplementary Tables 12-13** and

Supplementary Fig. 8). While short interspersed nuclear elements (SINEs), specifically *Alu*s, are found embedded as part of the human satellite 1 (HSat1) units across most of the q-arm, other transposable elements (TEs: long-interspersed nuclear elements (LINEs), long-terminal repeats (LTRs), SINE-VNTR-*Alu*s (SVAs), DNA transposons, and Rolling circles) are completely absent (**Fig. 1**). Moreover, TE distribution biases typify different subregions of ChrY, as *Alu*s are enriched in the PAR1 region, while other TEs (particularly L1s) are more abundant in the X-transposed region (XTR)[1] (**Extended Data Fig. 4b-c** and **Supplementary Table 14**). The DYZ19 region is annotated by RepeatMasker entirely as LTRs (**Extended Data Fig. 4c**), but further analyses indicate this is a satellite array spanning 265 kb whose 125 base monomeric consensus is derived from an expanded portion of a LTR12B sequence[40]. Repeat discovery and annotation of T2T-Y also allowed for improved annotation of ChrX in both HG002 and CHM13, particularly in the PAR regions, adding ~33 kb of satellite annotations per ChrX (**Supplementary Table 15**).

In addition, we searched for TE driven transductions mediated by L1s and SVAs. We detected six potential 3' L1 transductions within the T2T-Y, yet no SVA-driven DNA transductions (**Supplementary Table 16**). Despite a genome-wide investigation of both T2T-CHM13+Y and GRCh38, we were not able to locate any potential donor elements, which confirms a prior analysis that found no evidence for DNA transduction between the Y and the rest of the genome[41]. The transduction rate in T2T-Y was also much lower (0.096 per 1 Mb) than the transduction rate observed in the CHM13 autosomes (avg. 6.9 per 1 Mb) and ChrX (10.19 per 1 Mb)[39] (**Supplementary Note 1**).

In the T2T-Y, we identified a total of 825,526 repetitive sequence motifs capable of forming alternative DNA structures (non-B DNA), compared to only 138,640 in GRCh38-Y (**Supplementary Table 17**, **Supplementary Note 2**). This nearly 6-fold increase is largely attributed to our use of novel and improved experimental and computational methodology, as non-B DNA motifs, which might form structures during sequencing, are notoriously difficult to sequence through[42]. We found a particular enrichment of these motifs at the newly sequenced centromeric region (see below) and heterochromatic region on the Yq arm (**Fig. 1**).

Ampliconic genes in composite repeats

Composite repeats are a type of segmental duplication that are typically arranged in tandem arrays, likely derived through unequal crossing over that contributed to their increased copy numbers[1,37]. The *TSPY, RBMY*, and *DAZ* ampliconic gene families are all associated with composite repeats on the Y chromosome, and the T2T-Y assembly provides an opportunity to analyze the complete structure of these arrays (**Fig. 2**).

TSPY contains the largest number of protein-coding copies on the Y chromosome and is only expressed in testis. Expression level of this gene is dosage dependent and the copy number is polymorphic between individuals[43]. In GRCh38-Y, the *TSPY* array includes a 40 kb gap and a limited number of intact protein-coding copies. Our T2T-Y assembly resolved 45 protein-coding *TSPY* copies, including *TSPY2*, which was found downstream of the *TSPY* array in the distal part of the proximal inverted repeat IR3, in contrast to GRCh38-Y where it is located upstream, possibly due to translocations between the IR3 pairs. The distal positioning of *TSPY2* in HG002

was confirmed among all other Y haplogroups except R and Q, which match the proximal positioning of GRCh38-Y[32]. All 44 protein-coding copies in the *TSPY* array are embedded in an array of composite repeat units (~20.2 kb in size, matching prior reports[1,43]), with one composite unit per gene (**Fig. 2a** and **Supplementary Table 18**). Each unit includes five new repeat annotations (fam-*), several retroelements in the LINE, SINE, and LTR classes, and simple repeats. This 931 kb array is the largest gene-containing composite repeat array in the T2T-CHM13+Y assembly outside of the rDNA locus, and the third largest overall (the first being the rDNA arrays followed by an LSAU-BSAT composite array on Chr22[39]).

Data from 187 SGDP samples confirmed high *TSPY* sequence conservation but copy number varied from 10–40 copies (**Fig. 2b**). Phylogenetic analysis using protein-coding *TSPY*s from a Sumatran orangutan (*Pongo abelii*) and a Silvery gibbon (*Hylobates moloch*) as outgroups confirmed that all protein-coding *TSPY* copies (including *TSPY2*) originated from the same branch, which is separated from the majority (all but one) of *TSPY* pseudogenes (**Extended Data Fig. 6**). This result contradicts earlier findings, which concluded that *TSPY2* originated from a different lineage[44].

The composite structure of *RBMY* is similar to that of *TSPY* (one composite unit per gene), is comparable in size (with *RBMY* at 23.6 kb), and includes LINEs, SINEs, simple repeats, and eight new repeat annotations (**Fig. 2c**). In contrast, the *DAZ* locus is structured such that the entire repeat array, consisting of 2.4 kb composite units each containing a new repeat annotation and a fragmented L3, falls within one gene annotation (**Fig. 2d**). Out of the three composite arrays described here, *DAZ* is the only one also found on an autosome (Chr3, *DAZL*)[45], although as a single unit and lacking the young LINE1 (L1PA2) insertion of the ChrY *DAZ* copies.

Centromere

Normal human centromeres are enriched for an AT-rich satellite family (~171 base monomer), known as alpha satellite, typically arranged into higher-order repeat (HOR) structures and surrounded by more diverged alpha and other satellite classes[46]. Each HOR copy is nearly identical and comprises a tandemly arrayed set of monomers. We annotated 366 kb of alpha satellite in T2T-Y, spanning 317 kb of the DYZ3 HOR array. While the individual units within the HOR array are highly similar (99.5–100%), three HOR subtypes were identified from the full-length repeat units based on their monomer structure (red, blue, and green HOR haplotypes in **Fig. 3, Supplementary Figs. 9-13** and **Supplementary Tables 19-20**). The majority of the T2T-Y centromeric array is composed of 34-mer HORs with a small expansion of a 36-mer, and with longer HOR variants observed in the flanking p-arm (42-mer) and q-arm (46-mer). These variants are structurally different from the RP11 centromere, which is the basis for the GRCh38-Y centromere model and was recently finished by ONT sequencing[47] (**Fig. 3**).

Methylated CpG sites called by both HiFi and ONT reads reveal two adjacent regions of hypomethylation (separated by approximately 100 kb) in the centromeric dip region (CDR) (**Fig. 3**), which has been reported to coincide with the CENP-A binding and is the putative site of kinetochore assembly[46]. In the T2T-Y centromere, the presence of two distinct hypomethylated dips per chromatin fiber was confirmed by inspection of single-molecule ONT reads

(**Supplementary Fig. 14**). A similar pattern of multiple methylation dips within a single centromere was observed in other T2T-CHM13 chromosomes such as Chr11 and Chr20[48]. In addition, the HORs contained abundant inverted, A-phased, and mirror repeat motifs, forming a periodic pattern occurring every 5.7 kb (**Fig. 3** and **Supplementary Table 17**). Such non-B DNA motifs, inverted repeats in particular, potentially forming cruciforms, are hypothesized to play a functional role in defining human Y centromeres[49] and their presence is confirmed here at the sequence level.

Sequence classes and palindromes

We annotated sequence classes on the T2T-Y as ampliconic, X-degenerate, X-transposed, pseudoautosomal, heterochromatic, and other, in accordance with Skaletsky *et al*[1]. In addition, we were able to classify a more precise annotation for the satellites (including DYZ17 and DYZ19) and the centromere (**Fig. 1** and **Supplementary Table 21**). The X-degenerate and ampliconic regions were estimated to be 8.67 Mb and 10.08 Mb in length, in concordance with previous findings[1]. The T2T-Y ampliconic region contains eight palindromes, with palindromes P4–P8 highly concordant with GRCh38-Y (i.e. in terms of arm, spacer lengths, and sequence identity). Arm-to-arm identity of these five T2T-Y palindromes nested within X-degenerate regions ranged from 99.84–99.96% (**Supplementary Table 22-23**). Palindromes P1–P3 harbor the AZFc region, which contains genes critical for sperm production[51]. We discovered a large polymorphic inversion (>1.9 Mb) in respect to GRCh38-Y that likely arose from a single non-allelic homologous recombination event. Using Strand-seq, we were able to locate the breakpoints at two "red" amplicons (naming according to Kuroda-Kawaguchi *et al.*[52]): one forming the P2 palindrome and the other inside the P1 palindrome (**Fig. 4**). This rearrangement was previously annotated as the "gr/rg" (green-red/red-green) inversion with variable breakpoints and was confirmed to be present across six Y-chromosome haplogroups out of 44 genealogical branches[53]. Another inversion was detected in P3, which was recently reported as a recurrent variation in human[54] (**Extended Data Fig. 7a**). Although inversions between amplicons are believed to serve as substrates for subsequent AZFc deletions and duplications that might affect sperm production[53,55–57], pinpointing the breakpoints and measuring the frequency of the polymorphic inversions was difficult because of the large size and high identity of the palindromic arms.

Composition of the q-arm heterochromatin

The human Y chromosome contains a large heterochromatic region at the distal end of the q arm (Yq12), which consists almost entirely of two interspersed satellite sequences classically referred to as DYZ1 and DYZ2[58–61]. The single largest gap in GRCh38-Y is at Yq12, with minimal representation of DYZ1 and DYZ2, mostly in unplaced scaffolds. Here, we uncovered the detailed structure of the Yq12 region at single-base resolution, characterizing over 20 Mb of DYZ1 and 14 Mb of DYZ2 repeats. In T2T-Y, DYZ1 and DYZ2 are interspersed in 86 large blocks, with DYZ1 blocks ranging from 80–1,600 kb (median of 370 kb) and DYZ2 blocks ranging from 20–1,200 kb (median of 230 kb). DYZ2 blocks appear more abundant at the distal end of Yq12, and this trend is also visible in metaphase chromosome spreads with fluorescence *in situ* hybridization (FISH) (**Fig. 5a-b** and **Fig. 1** Cen/Sat track). Yq12 is highly variable in size and sequence structure

between individuals[62–64], and the number and size of these satellite blocks is expected to vary considerably.

DYZ1 is composed of a Y-specific subfamily of HSat3 sequences that occurs primarily as ~3.6 kb nested tandem repeats derived from an ancestral tandem repeat of the pentamer CATTC[65]. DYZ2 is composed of an unrelated satellite family, HSat1B, and comprises a ~2.5 kb tandem repeat made up of three parts: an ancient *Alu*Y subunit (20% diverged from the *Alu*Y consensus), an extremely AT-rich region (>85% AT), and a more GC-rich region[61,65,66]. The vast majority of repeat instances were over 98% identical, with slightly higher divergence at the more peripheral satellite blocks (**Fig. 5c**). A detailed comparison of the sequences within T2T-Y revealed recent structural rearrangements including iterative, tandem duplications as large as 5 Mb, which span multiple blocks of DYZ1 and DYZ2 (**Extended Data Fig. 8**). These structural rearrangement patterns are consistent with evolution by unequal exchange mechanisms. In addition, approximately 15% of Strand-seq libraries showed sister chromatid exchanges within the Yq12 heterochromatic region (**Extended Data Fig. 7b**).

While HSat3 is present across multiple chromosomes including the acrocentric short arms, HSat1B is present almost exclusively on the Y and the acrocentric short arms in smaller amounts, with the exception of Chr10[67]. While HSat1B carries an *Alu*Y-derived subunit as part of its composite repeat unit, some HSat3 arrays are tightly associated with *Alu* sequences, with blocks of HSat3 intermingled with *Alu* fragments, including *Alu*Y. Phylogenetic analyses place the ChrY HSat1B *Alu*Y subunits in a cluster with *Alu*Y subunits found in HSat1B sequences on the acrocentric chromosomes, with the highly homogenized ChrY copies appearing as a single cluster. Given the topology of this tree, it appears that the HSat1B sequences found on the acrocentric chromosomes were derived from the Y-linked HSat1B, with seeding events leading to local expansion and homogenization (**Fig. 5d**, upper branches). The *Alu*Y fragments found interspersed with HSat3 on the Y chromosome also phylogenetically cluster with *Alu*Y fragments associated with HSat3 on the acrocentric chromosomes. However, there is no evidence for local homogenization of HSat3-*Alu* fragments; likewise, there is no support for phylogenetic clustering by subgroup nor by chromosome. Based on the deep divide between the HSat1B and HSat3 clades in the tree for both ChrY and the acrocentric chromosomes, we conclude that the initial seeding events that created these arrays were independent of one another, yet were derived from *Alu*Y elements from PAR1 (**Fig. 5d**, lower branches).

Improved variant calling for XY samples

We performed short-read alignment and variant calling for 3,202 samples (1,603 XX; 1,599 XY) from the 1KGP Phase 3, including 1,233 unrelated XY samples averaging at least 30× coverage of 150 bp paired-end reads[27]. This set of 1,233 XY samples spans all 26 geographically diverse 1KGP populations and 35 distinct Y-chromosome haplogroups (**Supplementary Table 24**). To more accurately represent the diploid nature of the PARs, we completely hard-masked ChrY in XX samples and ChrY PARs in XY samples, thereby forcing any reads originating from the ChrY PAR to align to the ChrX PAR (**Supplementary Tables 25-28** and **Extended Data Fig. 9**). Diploid genotypes were then called within the PAR for both XX and XY samples[68] (**Extended Data Fig.**

10a). Aside from this modification, the alignment and variant calling pipeline mirrored our previous analysis based on GRCh38-Y[69].

Across all 1,233 unrelated XY samples, we observed improved alignment to T2T-Y, including a higher number of mapped reads (increase of 1.4 million reads on average, SD: 432,115; **Fig. 6a**), a higher proportion of properly paired reads (increase of 1.4% on average, SD: 1.4%; **Fig. 6b**), and a lower proportion of mis-matched bases (decrease of 0.6% on average, SD: 0.06%; **Fig. 6c**) per sample relative to GRCh38-Y (**Supplementary Table 29**). Within syntenic regions of the two Y chromosome assemblies, the number of variants per sample declined for samples from all Y haplogroups with the exception of haplogroup R (haplogroup of GRCh38-Y) and with the greatest reduction observed for samples of haplogroup J1 (haplogroup of T2T-Y) (**Fig. 6d, Extended Data Fig. 10b-c**). Selecting one individual each from the J1, R1b, and E1b haplogroups, we compared per-variant read depth and allele balance for both references (**Fig. 6e**). In all three samples, we observed more variants with excessive read depth and abnormal allele balance on GRCh38-Y, corresponding with putative collapsed duplications (**Supplementary Table 30**, **Fig. 6f-g**). We replicated these analyses using an additional 279 samples across 142 populations from the SGDP[28] and found similarly improved mappings and variant discovery using T2T-Y (**Extended Data Fig. 10d-e**, **Supplementary Figs. 16-18**).

Due to genomic repeats, accuracy of short-read variant calling is heterogeneous across the genome. One approach for improving reliability is to restrict analysis to "accessible" regions based on various alignment metrics. To this end, we followed published protocols to generate a short-read accessibility mask for T2T-Y based on patterns of normalized read depth, mapping quality, and base-calling quality[70]. Our masks reveal that while the heterochromatic long arm (Yq) remains largely inaccessible to short-read analysis, T2T-Y still adds 578 kb of accessible sequence compared to GRCh38-Y (increase of 4.2%, **Table 1**).

Taken together, these analyses indicate the complete T2T-Y reference improves short-read alignment and variant calling across populations and corrects errors in GRCh38-Y, but acknowledging the rich resources available on GRCh38, we also curated a 1-to-1 whole-genome alignment between each GRCh reference (GRCh37 and GRCh38) and T2T-CHM13+Y to enable lifting annotations in either direction. The vast majority of genetic variants in ClinVar (2022-03-13 release), dbSNP (build 155), and GWAS Catalog (v1.0 release) were successfully lifted to T2T-CHM13+Y (99.2%/97.8%/98.9% overall and 100%/95.0%/100% for ChrY, respectively, **Supplementary Table 31**). Accessibility masks and lifted annotations are provided along with variant calls as a resource for future studies (**Data Availability**).

Contamination of genomic databases

Human DNA sequences can sometimes appear as contaminants in the assembled genomes of other species. In microbial studies, the human reference sequence has been used to screen out contaminating human DNA; however, due to the incomplete nature of the current reference, some human fragments are missed and mistakenly annotated as bacterial proteins, leading to thousands of spurious proteins in public databases[72,73]. For example, a recent analysis of nearly 5,000 human whole-genome data sets found an unexpected linkage between multiple bacterial

species and human samples of XY karyotype, including 77,647 100-mers that were significantly enriched in the XY samples[74]. The authors hypothesized that these bacterial genomes were not actually present in the samples, but rather the effect was caused by real human ChrY sequences matching to contaminated bacterial genome database entries. We compared XY-enriched 100-mers from the Chrisman *et al.* study[74] to the T2T-Y chromosome and found that, as predicted, more than 95% of them had near-perfect matches to the complete T2T-Y sequence.

We further tested the entire NCBI RefSeq bacterial genome database (release 213, July 2022, totalling 69,122 species with 40,758,769 contig or scaffold accessions) and identified all 64-mers that appeared in both the bacterial database and T2T-Y. We found 4,179 and 5,148 potentially contaminated sequences matching GRCh38-Y and T2T-Y, respectively. (**Extended Data Fig. 11a,** top and middle). The sequences were relatively short in length (<1 kb), as is typical of contaminating genomic segments (**Extended Data Fig. 11b**). The vast majority of contaminated sequences found only with T2T-Y localized to the newly added HSat1B and HSat3 repeats (**Extended Data Fig. 11c**, **Supplementary Table 32**). Such repetitive sequences are common sources of contamination because of their high copy-number. We predict this human-derived sequence contamination issue includes sequence from all human chromosomes and extends to all sequence databases, including non-microbial genomes.

Discussion

Owing to its highly repetitive structure, the human Y chromosome is the last of the human chromosomes to be completed from telomere to telomere. Here, we have presented T2T-Y, a complete and gapless assembly of the Y chromosome from the HG002 benchmarking genome, along with a full annotation of its gene, repeat, and organizational structure. We have combined T2T-Y with the prior T2T-CHM13 assembly to construct a new reference, T2T-CHM13+Y, that is inclusive of all human chromosomes. This assembly, along with all of the annotation resources presented here, is available for use as an alternative reference via NCBI and the UCSC Genome Browser[75] (**Data Availability**).

Our analysis of the T2T-CHM13+Y reference assembly reveals a reduction in false-positive variant calls for XY-bearing samples due to the correction of collapsed, incomplete, misassembled, or otherwise inaccurate sequences in GRCh38-Y. Given the history of the GRCh38-Y assembly and its reliance on BAC libraries, we see no feasible means for its completion and suggest T2T-Y as a more suitable ChrY reference going forward. We recommend the use of T2T-CHM13 when mapping reads from XX samples and ChrY-PAR-masked T2T-CHM13+Y when mapping XY samples (**Supplementary Note 3**).

The completion of ampliconic and otherwise highly repetitive regions of ChrY will also require updates to existing gene annotations that are based on the incomplete GRCh38-Y assembly. How to label and refer to genes within variable-size ampliconic arrays, like *TSPY*, is an open question. Moreover, the highly repetitive sequences pose new challenges to computational tools developed on GRCh38. One example is the inconsistent methylation pattern observed in the satellite enriched heterochromatin region, in which both HiFi and ONT are prone to sequencing biases, hindering accurate biological interpretation (**Supplementary Note 4** and **Supplementary**

Fig. 19). Lastly, we have noted the improved detection of human contamination in genomic databases using T2T-CHM13+Y and recommend a full contamination audit of public genome databases using this updated human reference. Taken together, these results illustrate the importance of using a complete human reference genome for bioinformatic analyses.

Construction of the T2T-Y assembly challenged the assembly methods developed for the haploid CHM13 genome and spurred the development of new, automated methods for diploid human genome assembly. In particular, the PARs of the HG002 sex chromosomes required phasing akin to heterozygous, diploid haplotypes, and the palindromic and heterochromatic regions of ChrY required expert curation of the initial assembly string graph. Lessons learned from our assembly of T2T-Y informed the development of the Verkko assembler[76], which automates the integration of HiFi and ONT data for diploid human genome assembly. The companion study of Hallast *et al.*[32] successfully used Verkko to generate 43 near-T2T assemblies from a diverse panel of human Y chromosomes, revealing dynamic structural changes within this chromosome over the past 180,000 years of human evolution. Ultimately, as the complete, accurate, and gapless assembly of diploid human genomes becomes routine, we expect "reference genomes" will become known as simply "genomes".

Projects such as the Human Pangenome Reference Consortium[77] are in the process of generating high-coverage HiFi and ONT sequencing for hundreds of additional human samples, and the assembly of these diverse, complete human genomes, along with similar quality assemblies of the non-human primates, will provide an unparalleled view of human variation and evolution. With the availability of complete, diploid human genome assemblies, association between phenotype and genotype will finally move beyond small variants alone and be made inclusive of all complex, structural genome variation.

References

1. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
2. Miga, K. H. *et al.* Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).
3. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
4. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
5. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
6. Gustafson, M. L., M. D. & Donahoe, P. K., M. D. MALE SEX DETERMINATION: Current Concepts of Male Sexual Differentiation. *Annu. Rev. Med.* **45**, 505–524 (1994).
7. Vog, P. H. *et al.* Human Y Chromosome Azoospermia Factors (AZF) Mapped to Different Subregions in Yq11. *Hum. Mol. Genet.* **5**, 933–943 (1996).
8. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
9. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
10. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
11. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
12. Nurk, S. *et al.* HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* gr.263566.120 (2020) doi:10.1101/gr.263566.120.
13. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
14. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
15. Kirsche, M. *et al.* Jasmine and Iris: population-scale structural variant comparison and analysis. *Nat. Methods* **20**, 408–417 (2023).
16. Jain, C., Rhie, A., Hansen, N. F., Koren, S. & Phillippy, A. M. Long-read mapping to

repetitive reference sequences using Winnowmap2. *Nat. Methods* **19**, 705–710 (2022).
17. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).
18. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
19. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
20. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).
21. Shumate, A. *et al.* Assembly and annotation of an Ashkenazi human reference genome. *Genome Biol.* **21**, 129 (2020).
22. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
23. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
24. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
25. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).
26. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
27. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022).
28. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
29. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
30. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
31. Sanders, A. D. *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).
32. Hallast et al. Assembly of 43 diverse human Y chromosomes reveals extensive complexity and variation. *bioRxiv* (2022).

33. Hammer, M. F. *et al.* Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Hum. Genet.* **126**, 707 (2009).
34. Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–599 (2016).
35. Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.* **18**, 1362–1368 (2008).
36. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* **38**, 2049–2051 (2022).
37. Vegesna, R., Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Dosage regulation, and variation in gene expression and copy number of human Y chromosome ampliconic genes. *PLOS Genet.* **15**, e1008369 (2019).
38. NCBI RefSeq v110 Browser. Homo sapiens isolate NA24385 chromosome Y, alternate assembly T2T-CHM13v2.0. Last accessed: 2022-12-01. (2022).
39. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).
40. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**, 533 (2008).
41. Halabian, R. & Makałowski, W. A Map of 3′ DNA Transduction Variants Mediated by Non-LTR Retroelements on 3202 Human Genomes. *Biology* **11**, 1032 (2022).
42. Weissensteiner, M. H. *et al.* Distinct sequencing success at non-B-DNA motifs. 2022.06.13.495922 Preprint at https://doi.org/10.1101/2022.06.13.495922 (2022).
43. Tyler-Smith, C., Taylor, L. & Müller, U. Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome. *J. Mol. Biol.* **203**, 837–848 (1988).
44. Xue, Y. & Tyler-Smith, C. An Exceptional Gene: Evolution of the TSPY Gene Family in Humans and Other Great Apes. *Genes* **2**, 36–47 (2011).
45. Saxena, R. *et al.* Four DAZ Genes in Two Clusters Found in the AZFc Region of the Human Y Chromosome. *Genomics* **67**, 256–267 (2000).
46. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
47. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).
48. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

49. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
50. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).
51. Nailwal, M. & Chauhan, J. B. Azoospermia factor C subregion of the Y chromosome. *J. Hum. Reprod. Sci.* **10**, 256 (2017).
52. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29**, 279–286 (2001).
53. Repping, S. *et al.* A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics* **83**, 1046–1052 (2004).
54. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986-2005.e26 (2022).
55. Teitz, L. S., Pyntikova, T., Skaletsky, H. & Page, D. C. Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages. *Am. J. Hum. Genet.* **103**, 261–275 (2018).
56. Jobling, M. A. Copy number variation on the human Y chromosome. *Cytogenet. Genome Res.* **123**, 253–262 (2008).
57. Navarro-Costa, P., Plancha, C. E. & Gonçalves, J. Genetic Dissection of the AZF Regions of the Human Y Chromosome: Thriller or Filler for Male (In)fertility? *BioMed Res. Int.* **2010**, e936569 (2010).
58. Evans, H. J., Gosden, J. R., Mitchell, A. R. & Buckland, R. A. Location of human satellite DNAs on the Y chromosome. *Nature* **251**, 346–347 (1974).
59. Schmid, M., Guttenbach, M., Nanda, I., Studer, R. & Epplen, J. T. Organization of DYZ2 repetitive DNA on the human Y chromosome. *Genomics* **6**, 212–218 (1990).
60. Manz, E., Alkan, M., Bühler, E. & Schmidtke, J. Arrangement of DYZ1 and DYZ2 repeats on the human Y-chromosome: a case with presence of DYZ1 and absence of DYZ2. *Mol. Cell. Probes* **6**, 257–259 (1992).
61. Altemose, N. A classical revival: Human satellite DNAs enter the genomics era. *Semin. Cell Dev. Biol.* **128**, 2–14 (2022).
62. Gripenberg, U. Size variation and orientation of the human Y chromosome. *Chromosoma* **15**, 618–629 (1964).
63. Mathias, N., Bayés, M. & Tyler-Smith, C. Highly informative compound haplotypes for the

human Y chromosome. *Hum. Mol. Genet.* **3**, 115–123 (1994).

64. Altemose, N., Miga, K. H., Maggioni, M. & Willard, H. F. Genomic Characterization of Large Heterochromatic Gaps in the Human Genome Assembly. *PLOS Comput. Biol.* **10**, e1003628 (2014).
65. Cooke, H. Repeated sequence specific to human males. *Nature* **262**, 182–186 (1976).
66. Frommer, M., Prosser, J. & Vincent, P. C. Human satellite I sequences include a male specific 2.47 kb tandemly repeated unit containing one Alu family member per repeat. *Nucleic Acids Res.* **12**, 2887–2900 (1984).
67. Babcock, M., Yatsenko, S., Stankiewicz, P., Lupski, J. R. & Morrow, B. E. AT-rich repeats associated with chromosome 22q11.2 rearrangement disorders shape human genome architecture on Yq12. *Genome Res.* **17**, 451–460 (2007).
68. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience* **8**, giz074 (2019).
69. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
70. Bekritsky M.A., Colombo C., Eberle M. A. Identifying Genomic Regions with High Quality Single Nucleotide Variant Calling. *Identifying Genomic Regions with High Quality Single Nucleotide Variant Calling* https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/identifying-genomic-regions-with-high-quality-single-nucleotide-.html.
71. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
72. Breitwieser, F. P., Pertea, M., Zimin, A. V. & Salzberg, S. L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **29**, 954–960 (2019).
73. Steinegger, M. & Salzberg, S. L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* **21**, 115 (2020).
74. Chrisman, B. *et al.* The human “contaminome”: bacterial, viral, and computational contamination in whole genome sequences from 1000 families. *Sci. Rep.* **12**, 9863 (2022).
75. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
76. Rautiainen, M. *et al.* Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* 1–9 (2023) doi:10.1038/s41587-023-01662-6.
77. Liao, W.-W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023).

Table

Table 1 | Comparison of GRCh38-Y and T2T-Y.

		GRCh38-Y	T2T-Y	%Δ
Assembly	**Total bases**	57,264,655	62,460,029	+9.1
	Assigned bases	57,227,415	62,460,029	+9.1
	Unlocalized bases	37,240	0	
	Num. gaps	56	0	
	Num. N-bases	30,812,366	0	
Annotation	**Num. genes**	589	693	+17.7
	Protein coding	66	106	+60.6
	Num. additional genes	6	110	
	Protein coding	1	41	
	Num. transcripts	681	883	+29.7
	Protein coding	372	488	+31.2
	Num. additional transcripts	4	206	
	Protein coding	4	120	
Ampliconic gene copy numbers	***BPY2***	4 (3, 0)	4 (3, 0)	0
	CDY	26 (4, 0)	26 (4, 0)	0
	DAZ	4 (4, 0)	4 (4, 0)	0
	HSFY	8 (2, 0)	8 (2, 0)	0
	PRY	8 (2, 0)	8 (2, 0)	0
	RBMY	32 (6, 4)	34 (6, 4)	+3.3
	TSPY	25 (7, 0)	66 (45, 0)	+164.0
	VCY	2 (2, 0)	2 (2, 0)	0
	XKRY	8 (0, 2)	8 (0, 2)	0
Haplogroup	**Haplogroup**	R-L20 (R1b1a2a1a2b1a1)	J-L816 (J1a2b3a1)	
	Ancestry	European	Ashkenazi Jewish	
Repetitive bases	**SINE**	2,625,350	4,385,917	+67.1
	Retroposon	18,506	18,500	-0.0
	LINE	6,378,323	6,456,888	+1.2
	LTR	4,604,368	4,613,537	+0.2
	DNA/Rolling-circle	2,626,425	4,387,030	+67.0
	Satellite	1,578,773	14,522,636	+819.9
	Simple repeat	1,124,311	21,568,381	+1,818.4
	Other	705,062	972,612	+37.9
	All repeat classes	17,501,283	53,004,524	+202.9
	% repetitive (non-N)	66.3	84.9	+28.1
Accessible with short-reads		13,785,359	14,363,623	+4.2

Annotation statistics for GRCh38-Y are taken from the RefSeq (v110) annotation, and T2T-Y statistics are taken from a lifted and curated combination of RefSeq (v110) and GENCODE (v35) annotations. Num. additional genes/transcripts are those found exclusively in one assembly compared to the other. Ampliconic gene copy numbers are shown as X(Y,Z) where X = total number of annotated genes; Y = protein-coding genes; and Z = transcribed pseudogenes. %Δ is the percent change from GRCh38-Y to T2T-Y. Blank spaces indicate not applicable.

Figures

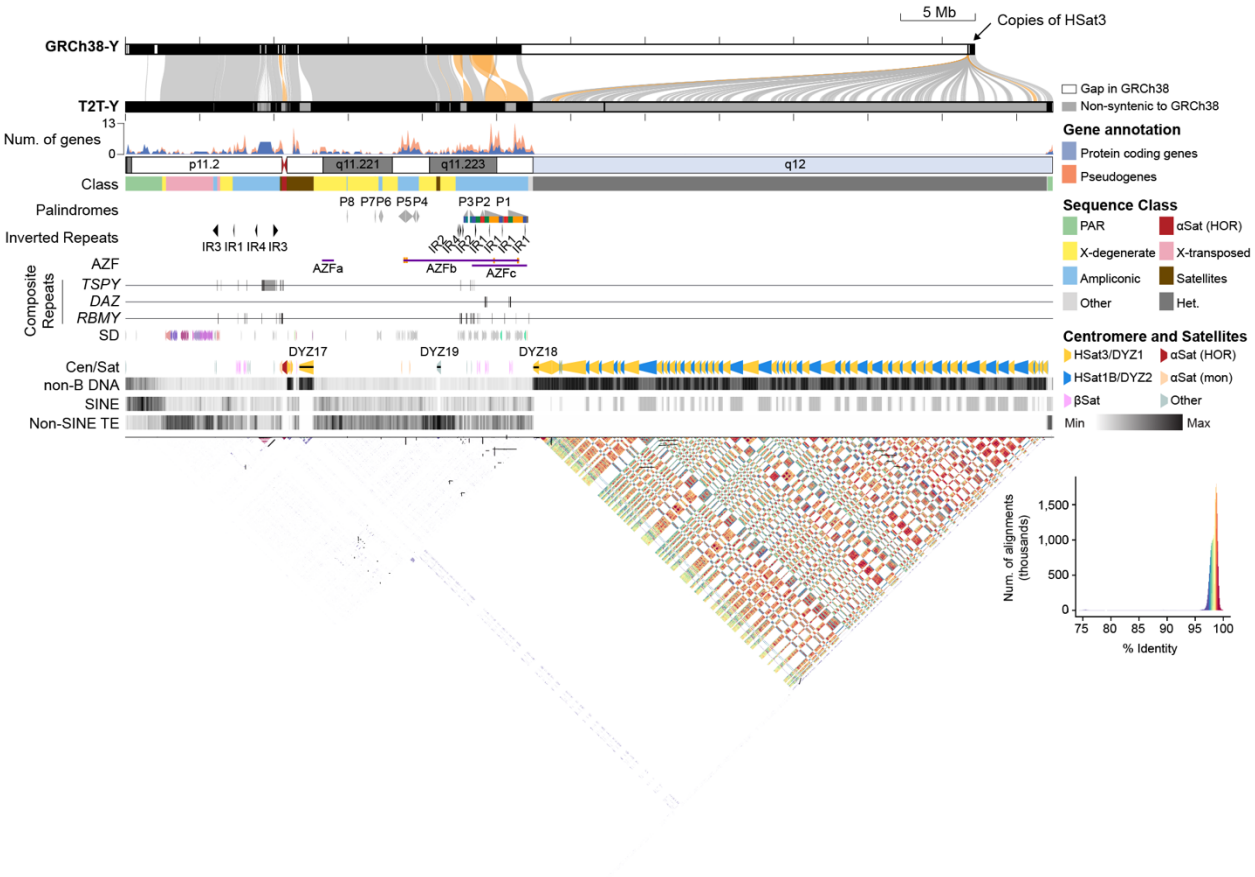


Fig. 1 | The structure of a complete Y chromosome. From top to bottom: Alignment of GRCh38-Y and T2T-Y. Regions with sequence identity over 95% are connected and colored by alignment direction (gray, forward; orange, reverse). Gene density plot shows enriched protein coding genes in ampliconic sequences. Sequence class, palindromes, inverted repeats (IR), and Azoospermia factor (AZF) a-c are annotated. Composite repeat arrays are named after the contained ampliconic genes. Segmental duplications (SDs) are colored by duplication types defined in DupMasker[35]. Centromere and satellite annotations (Cen/Sat) highlight the alternating HSat1 and HSat3 pattern comprising Yq12. Non-B DNA track shows regions forming alternate sequence structures are enriched in centromeric and satellite repeats. Short-interspersed repeat elements (SINE), including *AluY*, are highly enriched in the pseudo autosomal region 1 (PAR1). All other non-SINE transposable elements (TEs) are only found in the euchromatin. All repeats within T2T-Y are visualized by StainedGlass[36] with similar repeats colored by % identity in the style of an alignment dotplot.

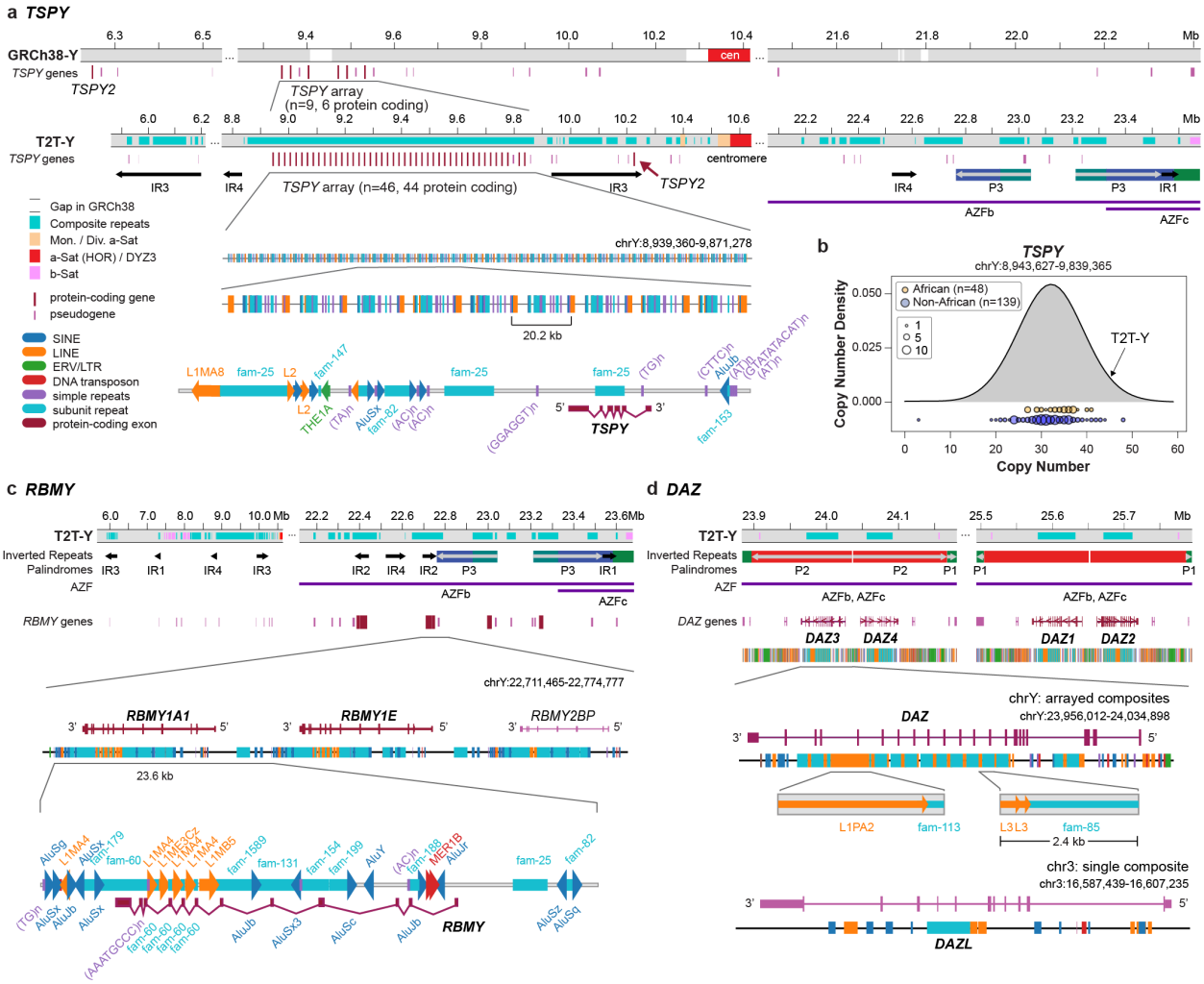


Fig. 2 | Ampliconic genes forming composite repeats. a. T2T-Y has 44 *TSPY* protein-coding genes organized in a single continuous array and a single *TSPY2* copy, compared to GRCh38-Y which has a gap in the *TSPY* array. T2T-Y shows a more regularized array and recovers additional *TSPY* pseudogenes not present in GRCh38-Y. **b**. Copy number differences of the *TSPY* protein-coding copies found in the SGDP. **c**, Repeat composition of the *RBMY* gene family. **d**. Repeat composition of the *DAZ* gene family, with one extra copy annotated on Chr3 that is missing L1PA2. While *TSPY* and *RBMY* genes are found within repeat composites forming arrays, *DAZ*-associated composites are embedded within the introns of the gene.

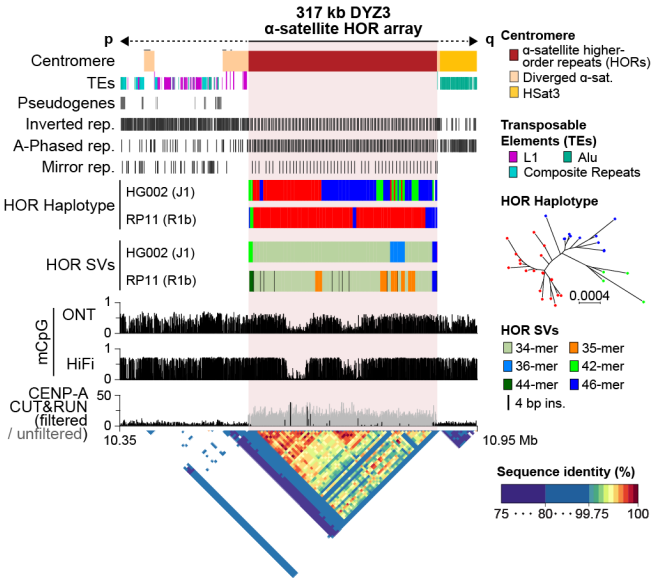


Fig. 3 | The structure of the T2T-Y centromere. No TEs were found within the DYZ3 array, while L1s (proximal) and *Alu*s (distal) were found within the diverged alpha satellites (drawn taller than the other TEs). A periodic non-B DNA motif pattern is shown within the HOR array. The HG002 (T2T-Y) HOR haplotypes and SVs reveal a different long-range structure and organization compared to a previously assembled centromere from RP11[47]. Three major HOR haplotypes were identified in HG002-Y based on their phylogenetic distance (red, blue, and green). RP11-Y has no 36-mer variants, but does have a number of 35-mers containing internal duplications. Histograms show the fraction of methylated CpG sites called by both ONT and HiFi, with two hypo-methylated centromeric dip regions (CDR) supported by CENP-A binding signal from CUT&RUN[50]. A StainedGlass dotplot illustrates high similarity within the HOR array (99.5–100%).

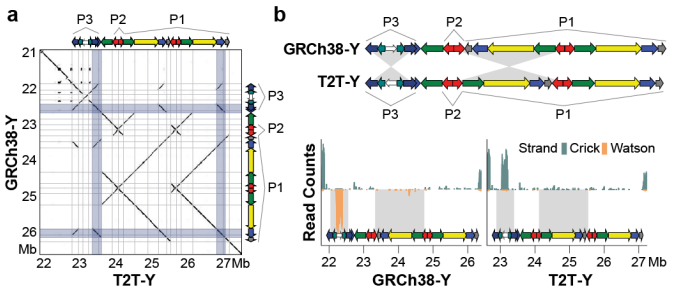


Fig. 4 | Comparison of the palindromic structure of the P1–P3 region. a. GRCh38-Y and T2T-Y alignment dotplot and schematics of the palindromes. Frequently recombining inverted repeats (IRs) in Azoospermia factor c (AZFc) region are highlighted in light blue. Deletion of AZFc between the IRs are known to cause spermatogenic failure[52]. A self-dotplot of the T2T-Y with AZFb and AZFc annotation is available as **Supplementary Fig. 15**. **b**. Top, a schematic of the palindromes. Two inversions are found, one in P3 and one between P1-P2. Below, Strand-seq signal from HG002 confirms the inverted orientation of P3 and P1 in T2T-Y compared to GRCh38-Y.

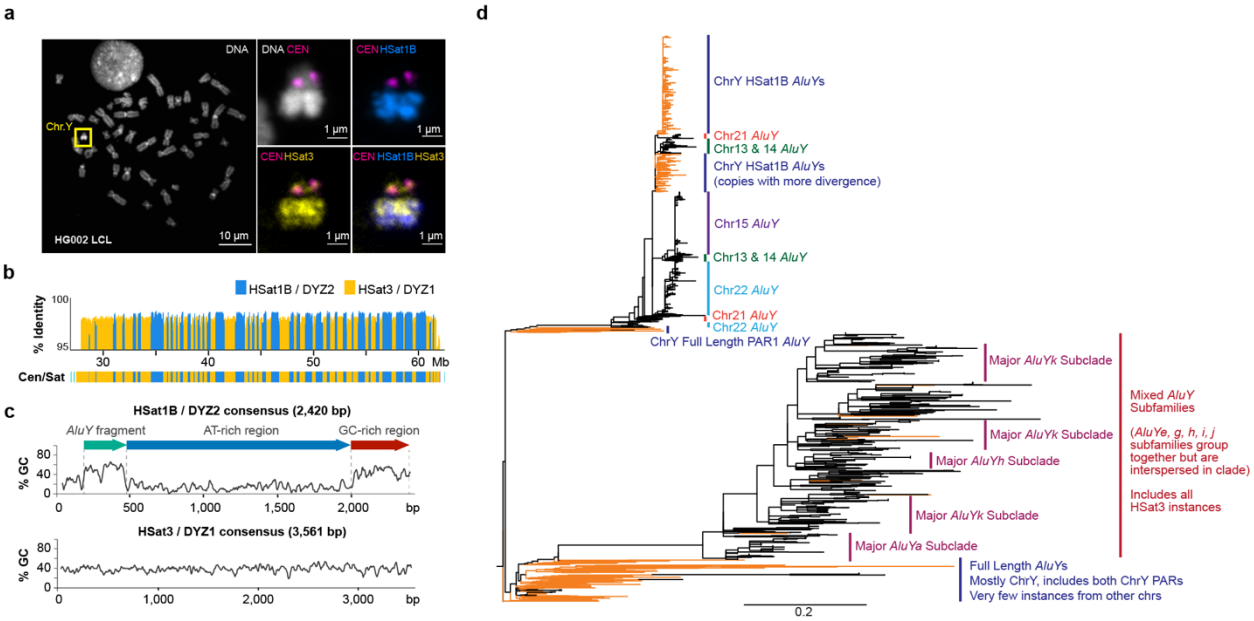


Fig. 5 | Heterochromatic region of the distal Y q-arm (Yq12). **a.** FISH painting of the Y chromosome, centromere/DYZ3 (magenta), HSat1B (blue), and HSat3 (yellow). Top-left, overall chromosome labeling by DNA dye (DAPI) with ChrY highlighted in an HG002-derived lymphoblastoid cell line (GM24385). The right panels show ChrY labeled with FISH probes recognizing centromeric alpha satellite/DYZ3 (magenta), HSat3/DYZ1 (yellow), and HSat1B/DYZ2 (blue). In concordance with the T2T-Y assembly, HSat3 probes indicate the presence at DYZ17 (close to centromere) as well as a slight enrichment to the proximal part of the Yq12 (DYZ1), while HSat1B is only present in the Yq12 and is more enriched towards the distal part (DYZ2). Maximum intensity projections are shown in all panels. The results of this experiment were replicated using two different sets of PCR probes. Fifteen large-field images containing at least 20 spreads were analyzed per condition. **b.** % identity of each DYZ2/DYZ1 repeat unit to its consensus sequence. **c**. % GC sequence composition of the HSat1B/DYZ2 and HSat3/DYZ1 repeat units and the position of an ancient *Alu*Y fragment in DYZ2. **d**. Phylogenetic tree of *Alu*Y sequences associated with HSat1B and HSat3, rooted on *Alu*Sc8. Tree represents subsampling of *Alu*Y elements, both full length (FL) and truncated, including *Alu*Y sequences found within HSat1B units and associated with HSat3 arrays. Elements located on ChrY are denoted with orange branches. The scale bar represents 0.2 substitutions per site on a branch of the same length.

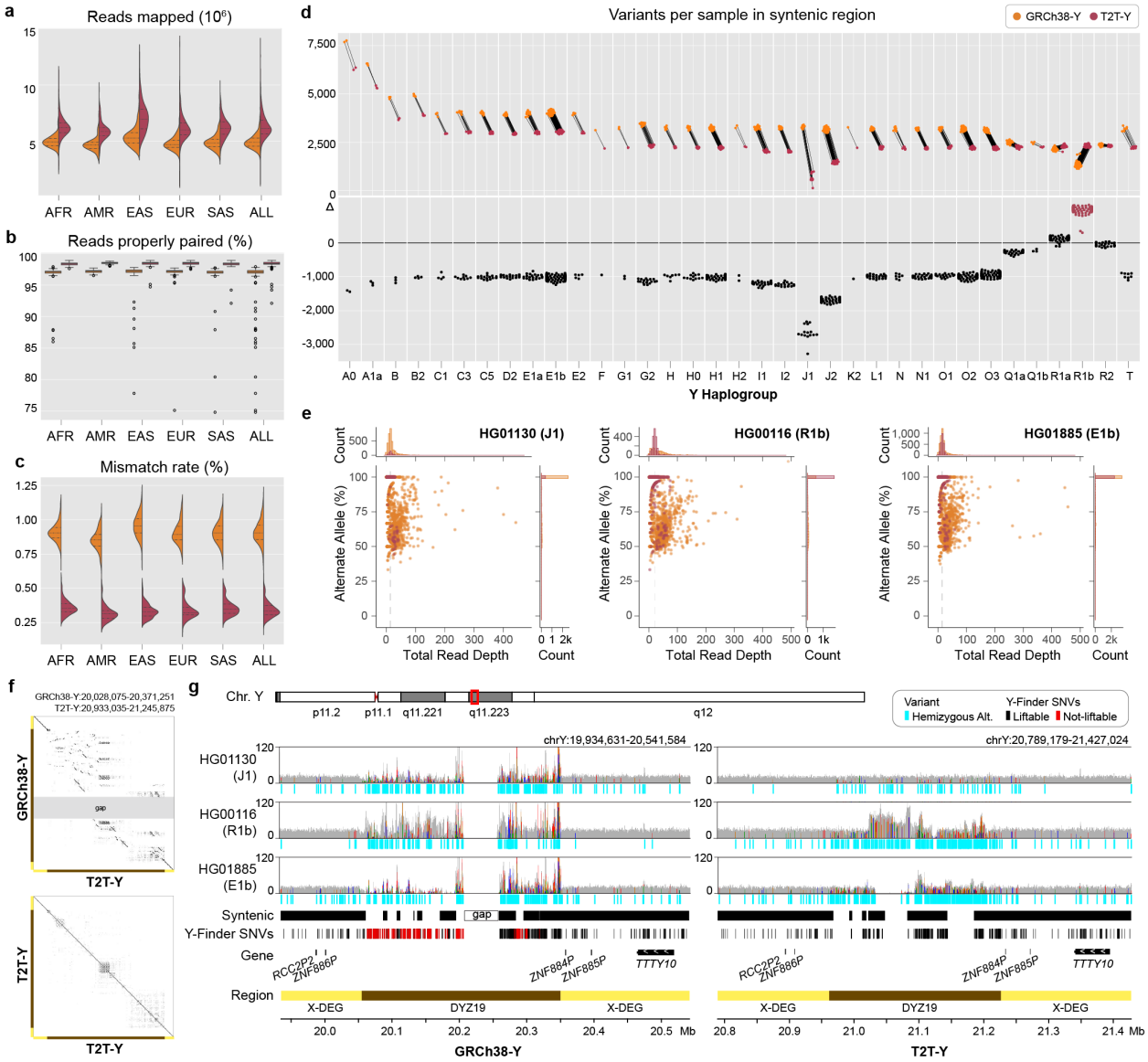


Fig. 6 | Short-read mappability and variant calling improvements on T2T-Y. In all plots, GRCh38-Y is colored orange and T2T-Y is maroon. The complete sequence of T2T-Y improves short-read alignment of the 1KGP dataset by **a.** increased number of reads mapped, **b**. higher portion of reads properly paired, and **c.** lower mismatch rate compared to GRCh38-Y. Bar in the box plot represents the 1st, 2nd (median), and 3rd quartile of the data. Whiskers are bound to the 1.5 × interquartile range. Data outside of the whisker ranges are shown as dots. **d.** The number of called variants within syntenic regions is reduced on T2T-Y for all haplogroups except R1 (haplogroup of GRCh38-Y). **e.** Further investigation on 3 samples (J1, R1b, and E1b) shows a higher number of variants called with excessive read depth and variable alternate allele fractions for GRCh38-Y. Each dot represents a variant, with the % alternate alleles as a function of total read depth. Dotted line represents the median coverage on T2T-Y, close to the expected 1-copy coverage. **f.** Dotplot of the DYZ19 array between GRCh38-Y and T2T-Y and self-dotplot of T2T-Y. Large rearrangements are observed, with multiple inversions proximal to the gap in GRCh38-Y with respect to T2T-Y (top), while more identical, tandem duplications are visible in T2T-Y (bottom). **g.** Read pile-ups and variants on DYZ19 for GRCh38-Y (left) and T2T-Y (right) as shown with IGV[71]. Gray histogram shows the mapped read coverage, with colored lines indicating non-reference bases with >60% allele frequency. Regardless of the haplogroup, the incomplete DYZ19 array in GRCh38-Y hinders interpretation. Syntenic regions between the two Ys are marked, and SNV sites used to identify Y haplogroup lineages in Y-Finder are shown below, with variants liftable from GRCh38-Y to T2T-Y in black, not-liftable in red, respectively.

Online Methods

This section provides a brief summary of the methods. Refer to the **Supplementary Methods** for more details.

Sequencing

Seventeen PacBio HiFi WGS runs were generated from GM24385 using the SMRTbell Express Template Prep Kit 2.0 on the Sequel II platform, after size-selection for 15 to 25 kb fragments. All of the ONT WGS runs are from the Jarvis *et al*.[20] study, which were generated using protocols from Shafin *et al*.[78] and Jain *et al*.[11].

RNA was extracted from three cell lines to generate Iso-Seq reads: EBV-immortalized lymphoblastoid cell line (GM24385), iPSC of the EBV-immortalized lymphoblastoid cell line (GM26105), and iPSC derived directly from Peripheral Blood Mononuclear Cells (GM27730). Iso-Seq data was generated on the Sequel II platform and was processed using Lima and IsoSeq3.

Specific runs used either in assembly or validation along with their accessions are available in **Supplementary Table 1**, and DNA preparation and library generation information is available in **Supplementary Methods**.

Assembly and validation

An assembly string graph was first constructed using PacBio HiFi reads (~60x) and processed using custom pruning procedures. Due to high sequence similarity within PAR1 and PAR2, the HG002 ChrX and ChrY string graph components shared connections to one another in the PARs, but to no other chromosomes in the genome. The remaining tangles in these sex-chromosome subgraphs were resolved using ONT reads longer than 100 kb (~90x). A semi-automated repeat resolution strategy utilized GraphAligner[13] to map the ultra-long ONT data to the HiFi assembly graph and identify the correct traversals. To resolve the PAR regions, ChrX and ChrY chromosomal walks were identified using homopolymer compressed trio-binned k-mers from parental Illumina reads[79], and a consensus sequence was computed for each. Remaining coverage gaps caused by HiFi sequencing biases were patched using a *de novo* Flye assembly of trio-binned paternal ONT reads[14,80]. For new projects, we now recommend the automated Verkko pipeline[76], which is able to replicate the semi-manual T2T-Y assembly presented here.

For polishing, the ChrXY draft assemblies were combined with the T2T-CHM13v1.1 autosomes to prevent mapping biases caused by the incompletely resolved autosomes of HG002 (Hereby T2T-CHM13+XY). For further polishing and validation, we used 66× Illumina, 84× HiFi, and 250× ONT (being haploid, the effective coverage on X and Y is half those depths). Small corrections were identified with DeepVariant[81,82] and filtered with Merfin[14]. Large errors were identified with Sniffles[83], cuteSV[84], and through a comparison to the HPRC-HG002v1 assembly[20]. All of the large errors were patched using marker-assisted HiFi and ONT reads. Assembly issues were identified using repeat-aware long-read alignments from Winnowmap2[16] (filtered with globally unique markers[17]) and VerityMap[85] (guided by locally unique markers). Coverage summaries were obtained using scripts from the T2T-CHM13 assembly evaluation[17]. Putative collapses and

inversion errors were identified using Strand-seq data. Raw sequencing reads from 65 Strand-seq libraries[30,31] were aligned to both GRCh38 and T2T-CHM13+XY with BWA[86], then processed with breakpointR[30,87] to identify inversion errors. Recurrent inversions were identified by comparing to results from Porubsky *et al.*[54]. To further confirm integrity of ChrY in the HG002 cell-line, we aligned publicly available GIAB[22] HiFi and ONT reads from the paternal HG003 genome (including from the PacBio Revio platform[88]), and performed the same coverage analysis. Base error rate was measured by Merqury using a hybrid k-mer set from Illumina and HiFi reads[17,18] (**Supplementary Table 3**).

Comparison to GRCh38-Y

Y haplogroup identification

The Y-chromosome haplogroup of HG002 and GRCh38 was identified using yhaplo[89], which builds a tree from phylogenetically informative SNPs that accumulate in the non-recombining portion of the Y. The Y-haplogroups of the 1KGP samples were identified using Y-Finder[90], using SNP calls on GRCh38 from Aganezov et al[69].

Alignments between GRCh38 and HG002 Y assemblies

Alignments between the GRCh38-Y and T2T-Y assemblies for the purposes of visualization with SafFire were generated with minimap2[91]. The PAF was then processed with rustybam and visualized with SafFire. DupMaske[35] and dna-brnn[92] annotations were generated using Rhodonite (10.5281/zenodo.6036498).

A complementary alignment was generated with LASTZ[93] after softmasking repeats from WindowMasker[94]. The alignment dotplot and best identity were plotted using R (https://github.com/arangrhie/T2T-HG002Y/tree/main/alignments/lastz). Regions along T2T-Y were colored according to their class.

To visualize three big structural differences of the three ChrY assemblies (GRCh37-Y, GRCh38-Y, and T2T-Y), we used the Pangenomics Research Tool Kit[95] to construct principal bundles representing contiguous and conserved sequences among the pangenome contigs.

Gene annotation

GENCODEv35 CAT/Liftoff annotation

Preliminary gene annotation was performed by mapping GENCODEv35[96] annotations from GRCh38-Y to T2T-Y using a Cactus[97] alignment with Chimp as an outgroup. Iso-Seq reads were aligned and assembled with Stringtie2[98], aligned to the assembly with TransMap[99], and used as input for CAT[100] along with the GENCODEv35 annotation. GENCODEv35 Y annotations were mapped with Liftoff[101], then intersected with Bedtools[102] to isolate genes that Liftoff mapped to ChrY that were not in the CAT annotation.

De novo RefSeqv110 and GENCODEv38 annotation

In the meantime, a *de novo* RefSeq annotation was performed on both GRCh38 and T2T-CHM13+Y and released (v110) as previously described for other vertebrate genomes[103,104]. A total of 82,862 curated RefSeq transcripts, 345,700 cDNAs, 8.65 million ESTs, 9.7 billion RNA-Seq reads, and 83 million PacBio IsoSeq and Oxford Nanopore reads from over 30 distinct tissues were retrieved from SRA and tentatively aligned to the assembly using Splign[105] or minimap2[91].

Simultaneously, an Ensembl gene annotation was performed by a mapping subset of the genes from GENCODEv38[96] using minimap2[91] and MAFFT[106]. Transcripts with low coverage or identity (<98%) were re-aligned using Exonerate[107]. Genes in potential recent duplications or collapsed paralogues were adjusted accordingly.

RefSeq/Liftoff, curated ampliconic gene annotation

Because the additional copies of the ampliconic genes hindered comparison to known genes in GRCh38-Y with differing gene IDs and names, we performed one more annotation by mapping GRCh38 RefSeq v110 annotations with Liftoff to T2T-CHM13+Y. We compared ampliconic gene family annotation results from that of GENCODE CAT/Liftoff and assigned gene names followed by best gene coverage and identity, including introns. Later, based on discussions with authors of a companion paper[32], we adjusted the gene names for 3 protein-coding annotations based on exon sequence identity (**Supplementary Table 5**).

Validation of the ampliconic protein-coding genes

Copy numbers for each ampliconic gene family in both the GRCh38-Y and T2T-Y assemblies were estimated using an adapted application of AmpliCoNE[37]. Copy numbers of these gene families were previously estimated for HG002 using Illumina reads from GIAB[108] and droplet digital PCR (ddPCR)[37]. The only notable difference was in the *TSPY* copy number, in which we identified 45 intact protein-coding copies. The copy number was slightly higher in the assembly than the estimates derived from Illumina reads and ddPCR (45 vs. 40 and 42, respectively). The *in silico* PCR primer search matched all 44 protein coding copies in the *TSPY* gene array and *TSPY2*, and two pseudogenes at the 3' end of the *TSPY* array which we were unable to avoid in the ddPCR primer design. We conclude that our AmpliCoNE, ddPCR, and *in silico* PCR estimates agree with the ampliconic gene annotations in the T2T-Y assembly (**Table 1**).

Repeat annotation

Segmental duplications

Segmental duplication (SD) annotations were created using the same methods as in Vollger *et al.* without modification[3]. In brief, SDs in T2T-CHM13+Y were identified using SEDEF[109] after repeat masking with Tandem Repeats Finder[110] and RepeatMasker[111].

Repeat model discovery and annotation

A three-step repeat annotation was performed to annotate new repeat models on ChrY. First, RepeatMasker was performed on the T2T-Y assembly using the Dfam 3.3 library[112], hard-masking previously annotated repeats. Second, RepeatModeler analysis was performed on the remaining unmasked regions to identify new repeat model consensuses, which were subjected to extension and filtering, and used as a library for a secondary RepeatMasker run. Two methods were primarily used to identify new satellite repeats: ULTRA[113] and NTRprism[46]. Unannotated regions >5 kb were identified via Bedtools[114] by subtracting repeat annotations from first and last steps above. The regions were manually curated in UCSC Genome Browser to check for any feature overlap (e.g. gene annotations). Tandemly repeated sequences were detected and assessed with a combination of ULTRA, NTRprism, and the TRF GUI version[110] to determine the best monomer consensus for a given satellite model. Lastly, the compilation pipeline laid out in Hoyt *et al.*[39] was followed to avoid potential false positives by simply masking with a combined library of new repeat models and known repeat models (Dfam library). The same three-step repeat annotation pipeline was applied to GRCh38-Y as well. Repeats were summarized using buildSummary.pl[115] at the class and family level (**Table 1**, **Supplementary Table 12**) and at the subfamily level for new repeats (**Supplementary Table 11**) in both T2T-Y and GRCh38-Y.

Composite repeats

Composite elements were defined and characterized as described in Hoyt *et al.*[39] as repeating units consisting of three or more repeated sequences, including TEs, simple repeats, composite subunits, and/or satellites, that are found as a tandem array in at least one location in the genome. BLAT[116] was used to locate other composite unit copies across T2T-Y and cross-reference them with their associated gene annotations (CAT/liftoff). Identification of the potentially active, full length TEs (SINEs, LINEs, and retroposons are AluY, L1Hs, and SVA_E/F) across T2T-Y and GRCh38-Y was done by following the methods of Hoyt *et al.*[39].

Satellite annotation

Centromeric Satellite (Cen/Sat) annotations were generated as in Altemose *et al.*[46], with a few refinements tailored to include annotations of the entire ChrY. Major satellite types were extracted from the RepeatMasker track, with features merged for the same satellite type within 10 kb of each other. For HSat2 and HSat3, a specialized annotation tool was used (https://github.com/altemose/chm13_hsat)[46]. DYZ19 and HSat1B were annotated using RepeatMasker annotations. Exact boundaries between HSat3 and HSat1B (aka DYZ1 and DYZ2) were manually refined.

Transduction analysis

We utilized the same approach as Hoyt *et al.*[39] to identify putative DNA transductions mediated by retroelements. Briefly, L1s and SVAs were identified in T2T-Y to detect the target site duplications and 3' transduction signatures using a modified version of TSDfinder[117]. Then, we

removed transductions residing in SDs and masked the transduced sequences using RepeatMasker[111]. To find the potential progenitor of each transduction within T2T-CHM13+Y and GRCh38, the offspring sequences were aligned to the corresponding databases of full-length L1s and SVAs using BLAST[118].

Non-B DNA motif annotation

To predict sequence motifs with the potential to form alternative DNA structures (non-B DNA), we used nBMST[119] for repeat motifs (A-phased, direct, inverted, and mirror repeats and STRs) and Z-DNA motifs[120,121]. G4-motifs were detected with Quadron[122], which also yields a score that quantifies the stability of a predicted G4 structure based on a machine-learning algorithm trained on empirical datasets. Motifs with a Quadron score ≥19 were considered stable, and thus used throughout our analysis. The non-B motifs were intersected with other existing annotations of T2T-Y (gene annotations, satellite repeats, and CpG islands) using Bedtools[114]. Rideogram[123] was used to generate these visualized tracks as well as the three composite repeat tracks. GraphPad Prism[124] was used to generate the TE composition per sequence class plots.

Data visualization

For **Fig. 1** and **Fig. 3**, the alignment of GRCh38-Y and T2T-Y was visualized with SafFire[125]. Segmental duplications (SDs) are colored by duplication types defined in DupMasker[35]. IGV[71] was used to draw ideograms, sequence classes, palindromes, inverted repeats, and AZF. Bedtools[114] was used to calculate density (across each gene type), bp coverage (across each repeat class) and average CpG methylation frequency per 100 kb window. Dotplots colored by identity were generated with StainedGlass[36].

TSPY gene family analysis

TSPY copy number estimation from SGDP

Copy number of the TSPY gene was estimated as in Vollger *et al*.[3]. In brief, we applied the fastCN pipeline[126], which uses read-depth as a proxy. Short-read sequence data were processed into 36 bp non-overlapping fragments and mapped using mrsFAST[127] to a T2T-CHM13+Y reference masked with 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 for *TSPY* were generated by taking a weighted average of the copy number estimates from windows overlapping the locus.

Phylogenetic tree analysis of the TSPY genes

All curated protein-coding and pseudogene *TSPY* copies (including introns) from the CAT/Liftoff and RefSeq/Liftoff annotations were used. For outgroup rooting of the tree, *TSPY* sequences were used from *Hylobates moloch* (NW_022611649.1)[128] and *Pongo abelii* (KP141780.1)[129]. Alignment was carried out in MAFFT[106]. Phylogenetic analysis was run in RAxML-NG[130] with 200 bootstrap replicates with rapid bootstrap approximation. Consensus bootstrap values were then mapped to the highest likelihood phylogeny in Geneious[131] and visualized in FigTree[132].

Centromere analysis

The T2T-Y was processed using the standard alpha-satellite (AS) tools as described in Altemose *et al.*[46]. The S4CYH1L (DYZ3) AS HOR was re-examined and re-defined for this paper to take into account its polymorphic variants both known from the old literature[133,134] and revealed by the recent complete centromere assembly of RP11[47].

The CENP-A CUT&RUN data was aligned to the T2T-CHM13+Y assembly as previously described in Altemose *et al.* [46]. The alignments were filtered using the single-copy k-mer locus filtering method as described in Hoyt *et al.* [39] through the use of the UCSC GenomeBrowser tool overlapSelect.

HG002 ONT UL data was re-basecalled using Guppy v6.1.2, Remora to obtain CpG methylation data (Supplementary Table 1). Modbams were converted to FASTQ files and aligned with Winnowmap2[16]. HG002 ONT nanoNOMe data was generated in Gershman *et al*.[48] and analyzed with nanopolish[135]. The probability of methylation for each CpG site in PacBio HiFi reads was assigned using primrose (https://github.com/PacificBiosciences/primrose). Reads were aligned with pbmm2 (https://github.com/PacificBiosciences/pbmm2). The percent of methylated reads at each reference genome position was calculated using pb-CpG-tools (https://github.com/PacificBiosciences/pb-CpG-tools). Resulting modbams were re-processed identically to Remora-called ONT data to collect comparable aggregated native CpG methylation data. The CDR was manually annotated as the area where CpG methylation is lower than the flanking, active, alpha satellite (**Supplementary Fig. 14)**.

Sequence classes on the Y chromosome

The X-degenerate and ampliconic regions were annotated using either exact boundaries of palindromes or intrachromosomal identity as defined in Skaletsky *et al.*[1] with adjusted borders based on the gene annotations. T2T-Y was split into 5 kb sliding windows (step size 1 kb) and these sequences were mapped back to T2T-Y using Winnowmap2[16]. After excluding self-alignments, windows with identity >50% were considered indicative of ampliconic regions if present consecutively.

For the schematic representations in **Fig. 4**, amplicons from Teitz *et al.*[55] were mapped to GRCh38-Y and T2T-Y assemblies with Winnowmap2[16] to identify homologous regions.

Approximate boundaries of the palindrome P4–P8 arms were manually selected using Gepard[136] and further refined based on aligning palindromic arms and adjacent flanks against each other (arm1 to the reverse complement of arm2) using global alignment with Stretcher[137].

For AZFa, sequences between two HERV15 genes (including genes USP9Y and DDX3Y) were used to determine the AZFa boundaries[138]. Boundaries of AZFb and AZFc were defined by the amplicon units P5/proximal P1 deletion (yel3/yel1) and by the b2/b4 deletion. A self-dotplot of the T2T-Y assembly was used with word size of 100 in Gepard[136]. Breakpoints were identified as illustrated in Fig. 2 of Navarro-Costa *et al.*[57] as shown in **Supplementary Fig. 15**.

The PAR and X-transposed regions were initially identified using LASTZ[93] alignments between HG002-X, HG002-Y and CHM13-X. Exact boundaries were later refined using Minimap2[91] alignments.

Yq12 heterochromatin region

DYZ1 and DYZ2

DYZ1 and DYZ2 consensus sequences were generated using multiple sequence alignment using kalign[139] and converted to a profile HMM using HMMER[140]. Dotplots in the **Extended Data Fig. 8** were produced using dottup of the EMBOSS software package[137].

Phylogenetic tree analysis of the AluY

The AluY tree was rooted on the RepeatMasker/Dfam derived consensus sequence for AluSc8. Analysis was run on a MAFFT[106] derived alignment using RAxML-NG[130] with 100 non-parametric bootstrap replications. Note that in the AluY subfamily clade ("Mixed AluY Subfamilies") there are scattered elements across the group even though the majority are represented in the labeled subclades.

Short-read variant calling on T2T-CHM13+Y

Impact of masking PAR and XTR in variant calling

Simulated paired-end sequence reads were generated using NEAT[141]. Variants from 10 XY and 10 XX European individuals were collected from high coverage variant calls of 1KGP[27] and used for benchmarking. Reads were processed with bbduk[142] and mapped using BWA[86] to two versions of GRCh38: X and Y both unmasked (default), and sex chromosome aware (SCC aware[68]). Masking was performed on PAR[143] or both PAR and XTR[68,144]. Mapping quality (MAPQ) was assessed on ChrX in each 50 kb windows, sliding 10 kb using Bedtools[114]. Variant calling was performed with GATK[145] and compared against the chosen variants used in simulating the reads.

Mappability comparison and variant calling in 1KGP samples

Using the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-Space (AnVIL)[146], we performed short-read alignment and variant calling for the 3,202 samples in 1KGP[27] using the T2T-CHM13+Y assembly as a reference. These samples were sequenced to at least 30× coverage by the New York Genome Center (NYGC), and alignment and variant calling was previously performed on the GRCh38 reference. We largely followed the short-read alignment and variant calling pipeline used for analysis of T2T-CHM13v1.0[69], except that we used SCC references for each XX and XY individuals using XYalign[68]. In the XX-specific reference, the entire Y chromosome is masked, whereas in the XY-specific reference, only the Y-PARs are masked. For all analyses, measures of mappability (reads mapped, reads properly paired, mismatch rate) were assessed with Samtools[147], and variant counts and allele frequencies were assessed with bcftools[147]. Variants in syntenic regions between GRCh38-Y and T2T-Y were further subsetted with Bedtools[114].

Putative collapsed regions in GRCh38-Y

Three individual's variant calls and the corresponding bam files from the 1KGP dataset were downloaded from AnVIL: one individual each from the J1, R1b and E1b haplogroups (HG01130, HG00116 and HG01885, respectively). Variant calls on ChrY syntenic region were subsetted using bcftools[147]. From the VCF file, allelic read depth (defined as AD field) and reference allele depth (1st value in the DP field) were extracted using a custom script along with each variant's chromosomal position and visualized with R. Coverage tracks of the bam files were collected with IGVtoolkit[71] and samtools[147]. Variants from HG00116 on GRCh38-Y (R1b, thus least structural variations expected) were further aggregated as "excessive variant region" when non-reference alleles were present, merged within 50 kb. Coverage, variant calls, and the excessive variant regions were manually inspected on GRCh38-Y, and marked as a "putative collapse" if the region: 1) had an excessive number of variants called for all three samples, 2) overlapped with a known gap in GRCh38, and 3) did not overlap the palindromic region (where there were substantial rearrangements between the GRCh38-Y and T2T-Y).

Mapping and variant calling of the SGDP samples

The SGDP includes 279 open-access, high-coverage genomes from 130 diverse populations[28]. Compared to 1KGP, SGDP includes 118 additional populations with samples sequenced to an average of 43× coverage using a shared PCR-free Illumina library. The SGDP samples were aligned and genotyped to T2T-CHM13+Y and GRCh38 on AnVIL[146] following the same pipeline as the analysis of 1KGP samples.

Curated syntenic region and liftover chains

The initial chain file was generated using nf-LO[148] with minimap2[91] alignments. The alignments were filtered and converted to PAF using chaintools. Alignments of nonhomologous chromosomes were removed. Overlapping alignments in the query sequence was removed with rustybam to create 1:1 alignments. PAF alignments were converted back to chain format.

In addition to the minimap2-based whole genome alignment, we applied a wfmash-based pipeline[149] to validate the chain file. This pipeline starts with a wfmash[149] whole-genome alignment of T2T-CHM13+Y and the masked and filtered GRCh38 assembly, and identifies 1-to-1 homologous regions at least 5 kb long with a nucleotide identity of at least 95%. Similarly, the resulting chain was post-processed to obtain 1:1 alignments using rustybam and the paf2chain tool. All PAF files with full CIGAR strings were then inspected with SafFire for quality investigation. The minimap2- and wfmash-based chains showed high consistency over the genomes.

Datasets and resources for T2T-CHM13+Y

Lifting over resources from GRCh38 to CHM13+Y

Using the curated chain file, we lifted over dbSNP build 155[150], the March 13, 2022 release of Clinvar[23,151], and GWAS Catalog v1.0[24,152] from the GRCh38 primary assembly to T2T-CHM13v2.0 (T2T-CHM13+Y). Liftover was performed as previously described[69] using GATK Picard[153] LiftoverVcf and the alignment chain described above.

ENCODE

Reads were obtained from the ENCODE dataset[29] and mapped with Bowtie2[154]. Alignments were filtered using Samtools to remove unmapped or single end mapped reads and those with a mapping quality score <2. PCR duplicates were identified and removed with the Picard tools "mark duplicates". Alignments were then filtered for the presence of unique k-mers. Bigwig coverage tracks and enrichment tracks were created using deepTools2 bamCoverage[155].

gnomAD

Genome wide variant data from the Genome Aggregation Database (gnomAD) release v3.1.2 was lifted over from GRCh38 to each assembly using CrossMap[156]. The chain files used were created from the GRCh38-based HAL file, downloaded from the cactus-minigraph alignment of Liao *et al.*[77]. The resulting VCFs were annotated with predicted molecular consequence and transcript-specific variant deleteriousness scores from PolyPhen-2 and SIFT using Ensembl Variant Effect Predictor.

Human Y chromosome contamination in bacterial genomes

Screening against Chrisman *et al.* study

We used the MUMmer[157] to compare 73,691 bacterial 100-mers reported as enriched in human males by Chrisman *et al.*[74] to the T2T-Y assembly. We found that, as predicted, more than 95% of the 100-mers had near-perfect matches, defined as an exact match of 50 bp or longer, to the complete T2T-Y sequence. The nucmer program from MUMmer was run with default options,

except to specify -l 50 for an exact match length of 50 or more, and -c 50 so that it reported matches as short as 50 bp.

Screening with 64-mers

Meryl[18] was used to compare 64-mers from NCBI RefSeq release 213 (July 2022) to T2T-Y and GRCh38-Y. Each bacterial contig was annotated with the number of matching k-mers in T2T-Y, GRCh38-Y, and the number of k-mers in the contig with a match. Each position in the reference chromosomes was annotated with the multiplicity of the k-mer at that position in the RefSeq contigs, and with the number of contigs containing the k-mer. Hits per query were filtered to retain only contigs with more than 20 k-mer matches or with more than 10% of the contig sequence covered by k-mer matches. The queries at each reference position were combined and accumulated into 10 kb windows and converted to an interval wiggle file for visualization. RefSeq sequence entries with hits were retrieved using seqrequester and categorized using 64-mers built from HSat1B and HSat3 annotations. The first and second words in the sequence entry names were extracted to visualize the taxonomic abundance of the microbial genomes in a pie chart using Kronatools[158] (**Extended Data Fig. 11c)**.

Methods References

78. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
79. Koren, S. *et al.* De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
80. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
81. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
82. Shafin, K. *et al.* Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nat. Methods* **18**, 1322–1332 (2021).
83. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
84. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
85. Bzikadze, A. V., Mikheenko, A. & Pevzner, P. A. Fast and accurate mapping of long reads to complete genome assemblies with VerityMap. *Genome Res.* gr.276871.122 (2022) doi:10.1101/gr.276871.122.
86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
87. Porubsky, D. *et al.* breakpointR: an R/Bioconductor package to localize strand state changes in Strand-seq data. *Bioinformatics* **36**, 1260–1261 (2020).
88. Index of /public/revio/2022Q4. https://downloads.pacbcloud.com/public/revio/2022Q4/.
89. David Poznik. yhaplo | Identifying Y-Chromosome Haplogroups. Last accessed: 2022-11-29. https://github.com/23andMe/yhaplo (2022).

90. Tseng, B. *et al.* Y-SNP Haplogroup Hierarchy Finder: a web tool for Y-SNP haplogroup assignment. *J. Hum. Genet.* **67**, 487–493 (2022).
91. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
92. Li, H. Identifying centromeric satellites with dna-brnn. *Bioinformatics* **35**, 4408–4410 (2019).
93. Harris, Robert S. Improved Pairwise Alignmnet of Genomic DNA. (Penn State, 2007).
94. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
95. Chin, C.-S. *et al.* Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* 1–9 (2023) doi:10.1038/s41592-023-01914-y.
96. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
97. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
98. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278 (2019).
99. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
100. Fiddes, I. T. *et al.* Comparative Annotation Toolkit (CAT)—simultaneous clade and personal genome annotation. *Genome Res.* **28**, 1029–1038 (2018).
101. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
102. Dale, R. K., Pedersen, B. S. & Quinlan, A. R. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
103. Rhie, A. *et al.* Towards complete and error-free genome assemblies of all vertebrate species. *Nature* **592**, 737–746 (2021).
104. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756-763 (2014).
105. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
106. Katoh, K. & Standley, D. M. MAFFT: Iterative Refinement and Additional Methods. in *Multiple Sequence Alignment Methods* (ed. Russell, D. J.) 131–146 (Humana Press, 2014). doi:10.1007/978-1-62703-646-7_8.
107. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
108. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
109. Numanagić, I. *et al.* Fast characterization of segmental duplications in genome assemblies. *Bioinformatics* **34**, i706–i714 (2018).
110. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
111. Arian FA, S., Hubley, Robert, & Green, P. RepeatMasker Open-4.0 2013-2015. (2015).

112. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
113. Olson, D. & Wheeler, T. ULTRA: A Model Based Tool to Detect Tandem Repeats. in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* 37–46 (Association for Computing Machinery, 2018). doi:10.1145/3233547.3233604.
114. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
115. Storer, J. M., Hubley, R., Rosen, J. & Smit, A. F. A. Curation Guidelines for de novo Generated Transposable Element Families. *Curr. Protoc.* **1**, e154 (2021).
116. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
117. Szak, S. T. *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biol.* **3**, research0052.1 (2002).
118. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
119. Cer, R. z. *et al.* Searching for Non-B DNA-Forming Motifs Using nBMST (Non-B DNA Motif Search Tool). *Curr. Protoc. Hum. Genet.* **73**, 18.7.1-18.7.22 (2012).
120. Zou, X. *et al.* Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res.* **45**, 11213–11221 (2017).
121. Svetec Miklenić, M. *et al.* Size-dependent antirecombinogenic effect of short spacers on palindrome recombinogenicity. *DNA Repair* **90**, 102848 (2020).
122. Sahakyan, A. B. *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).
123. Hao, Z. *et al.* RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251 (2020).
124. GraphPad Prism version v9.1.0 for Windows, GraphPad Software, San Diego, California, USA. Last accessed: 2022-11-28. https://www.graphpad.com.
125. Vollger, M. R. SafFire. Last accessed: 2022-11-29. https://github.com/mrvollger/SafFire (2022).
126. Pendleton, A. L. *et al.* Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol.* **16**, 64 (2018).
127. Hach, F. *et al.* mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat. Methods* **7**, 576–577 (2010).
128. Escalona, M. *et al.* Whole-genome sequence and assembly of the Javan gibbon (Hylobates moloch). *J. Hered.* **114**, 35–43 (2023).
129. Cortez, D. *et al.* Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**, 488–493 (2014).
130. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
131. Geneious. Bioinformatics Software for Sequence Data Analysis. Last Accessed: 2022-11-28. *Geneious* https://www.geneious.com/.
132. FigTree. Last accessed: 2022-11-28. http://tree.bio.ed.ac.uk/software/figtree/.
133. Tyler-Smith, C. & Brown, W. R. A. Structure of the major block of alphoid satellite DNA on

the human Y chromosome. *J. Mol. Biol.* **195**, 457–470 (1987).
134. Shepelev, V. A. *et al.* Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genomics Data* **5**, 139–146 (2015).
135. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).
136. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).
137. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
138. Sun, C. *et al.* Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum. Mol. Genet.* **9**, 2291–2296 (2000).
139. Lassmann, T. Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* **36**, 1928–1929 (2020).
140. Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487–2489 (2013).
141. Stephens, Z. D. *et al.* Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLOS ONE* **11**, e0167047 (2016).
142. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. https://www.osti.gov/biblio/1241166 (2014).
143. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).
144. Poznik, G. D. *et al.* Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females. *Science* **341**, 562–565 (2013).
145. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
146. Schatz, M. C. *et al.* Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genomics* **2**, 100085 (2022).
147. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
148. Talenti, A. & Prendergast, J. nf-LO: A Scalable, Containerized Workflow for Genome-to-Genome Lift Over. *Genome Biol. Evol.* **13**, evab183 (2021).
149. Marco-Sola, S., Moure, J. C., Moreto, M. & Espinosa, A. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**, 456–463 (2021).
150. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **9**, 677–679 (1999).
151. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
152. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
153. Van der Auwera GA & O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and*

WDL in Terra. (O'Reilly Media, 2020).
154. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
155. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
156. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
157. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* **14**, e1005944 (2018).
158. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).
159. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
160. Arang Rhie. (2023). Repositories for the analysis of T2T-Y and T2T-CHM13v2.0. *Zenodo*. https://doi.org/10.5281/zenodo.8136598

Data Availability

The T2T-CHM13v2.0 (T2T-CHM13+Y) assembly, reference analysis set, complete list of resources including gene annotation, repeat annotation, epigenetic profiles, variant calling results from 1KGP and SGDP, gnomAD, ClinVar, GWAS, and dbSNP datasets are available for download at https://github.com/marbl/CHM13. The assembly is also available from NCBI and EBI with GenBank accession GCA_009914755.4. Annotation and associated resources are also browsable as "hs1" from the UCSC Genome Browser http://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_3671779_hs1, the Ensembl Genome Browser https://projects.ensembl.org/hprc/ (assembly name T2T-CHM13v2.0) and NCBI data-hub https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_009914755.1/ . Potential assembly issues are listed and tracked at https://github.com/marbl/CHM13-issues. 1KGP and SGDP short read alignments and variant calls are available within AnVIL at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T_CHRY. Sequencing data used in this study is listed in **Supplementary Table 1**.

Code Availability

Custom codes developed for data analysis and visualization are available at https://github.com/arangrhie/T2T-HG002Y, https://github.com/snurk/sg_sandbox, and https://github.com/schatzlab/t2t-chm13-chry and deposited on Zenodo along with https://github.com/marbl/CHM13 and https://github.com/marbl/CHM13-issues[160]. Software and parameters used are stated in the **Supplementary Methods** with more details.

Acknowledgements

We thank P. Hallast, M. C. Loftus, M. K. Konkel, P. Ebert, T. Marschall, and C. Lee for coordination and discussions, J.C.-I. Lee for sharing the GRCh38-Y coordinates used in Y-Finder, and members of the Telomere-to-Telomere consortium and Human Pangenome Reference Consortium for constructive feedback. This work utilized the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov). Computational resources were partially provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic and Computational Biology Core, Institute for Systems Genomics, University of Connecticut. Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Funding support: Intramural Research Program of the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH) HG200398 (A.R., S.N., S.K., M.R., A.M.M., B.P.W., A.M.P); NIH GM123312 (S.J.H., P.G.S.G., G.A.H., R.O.); NIH GM130691 (P.M., M.H.W., K.D.M.); HHMI Hanna Gray Fellowship (N.A.); NIH GM147352 (G.A.L.); NIH HG002939, HG010136 (R.Hu., J.M.S.); NIH HG009190 (P.W.H., A.Ge., W.T.); NIH HG010263, HG006620, CA253481, and NSF DBI-1627442 (M.C.S.); NIH GM136684 (K.D.M.); NIH HG011274, HG010548 (K.H.M.); NIH HG010961, HG010040 (H.L.); NIH HG007234 (M.D.); NIH HG011758 (F.J.S.); NIH DA047638 (E.G.); NIH GM124827 (M.A.W.); NIH GM133747 (R.C.M.); NIH CA240199 (R.O.); NIH HG002385, HG010169, HG010971 (E.E.E.); Stowers Institute for Medical Research (J.G., T.P.); National Center for Biotechnology Information of the National Library of Medicine (NLM), National Institutes of Health (F.T.-N., T.D.M.); Intramural funding at the National Institute of Standards and Technology (NIST) (J.M.Z.); NIST 70NANB20H206 (M.J.); NIH HG010972, WT222155/Z/20/Z, and the European Molecular Biology Laboratory (J.A., P.F., C.G.G., L.H., T.H., S.E.H., F.J.M., L.S.); RNA generation was supported by NIST 70NANB21H101 and NIH 1S10OD028587. Ministry of Science and Higher Education of the Russian Federation, St. Petersburg State University, PURE 73023672 (I.A.A.); The Computation, Bioinformatics, and Statistics (CBIOS) Predoctoral Training Program awarded to Penn State by the NIH (A.W.); Achievement Rewards for College Scientists Foundation, The Graduate College at Arizona State University (A.M.T.O.); E.E.E. is an investigator of the HHMI.

Author contributions

Assembly: S.N., S.K., M.R. Validation: A.R., S.K., M.A., A.V.B., G.F., A.F., A.M.M., J.M., A.M., L.F.P., D.P., F.J.S., K.S., P.M., J.M.Z., K.D.M. ChrY haplogroups: A.R., A.C.W. Alignment: C.-S.C., M.D., R.Har., M.R.V., K.D.M. Satellite annotation: N.A., I.A.A., G.A.L., F.R., V.A.S., K.H.M. FISH: N.A., J.G., T.P. Repeat annotation: S.J.H., P.G.S.G., G.A.H., R.Hu., J.M.S., R.O. Retroelements: R.Hal., W.M. Non-B DNA: M.H.W., K.D.M. Gene annotation: A.R., M.D., P.F., C.G.G., L.H., M.H., J.H., T.H., F.J.M., T.D.M., S.L.S., A.S., F.T.-N. Ampliconic genes: A.R., R.Har., W.T.H., P.M., M.T., K.D.M. Structural annotation: A.R., M.C., H.L., P.M., K.D.M. Epigenetics: A.R., P.W.H., A.Ge., W.T., A.M.W. Mappability: A.M.T.O., M.A.W., J.M.Z. Non-B DNA: M.H.W., K.D.M. Variants and Liftover: A.R., D.J.T., S.K., J.A., N.-C.C., M.D., E.G., A.Gu., N.F.H., W.T.H., S.E.H., S.H., R.C.M., N.D.O., M.E.G.S., L.S., M.R.V., S.Z., J.M.Z., E.E.E., A.M.P. Contamination: A.R., S.L.S., B.P.W., A.M.P. Data generation: M.J., R.K.K., A.P.L., J.K.L., C.M., B.M.M., K.M.M., H.E.O., F.J.S., Y.Z. Data management: A.R., M.D., M.J., J.K.L. Computational resources: R.O., M.C.S., A.M.P. Manuscript draft: A.R., S.N., M.C., S.J.H., D.J.T., N.A., I.A.A., N.-C.C., E.G., J.G., P.G.S.G., A.Gu., R.Hal., W.M., J.M., T.P., F.R., S.L.S., J.M.S., A.M.T.O., A.C.W., M.A.W., S.Z., J.M.Z., E.E.E., R.O., M.C.S., K.H.M., K.D.M., A.M.P. Editing: A.R., A.M.P., with the assistance of all authors. Supervision: J.M.Z., E.E.E., R.O., M.C.S., K.H.M., K.D.M., A.M.P. Conceptualization: A.R., S.N., M.C., E.E.E., K.H.M., K.D.M., A.M.P.

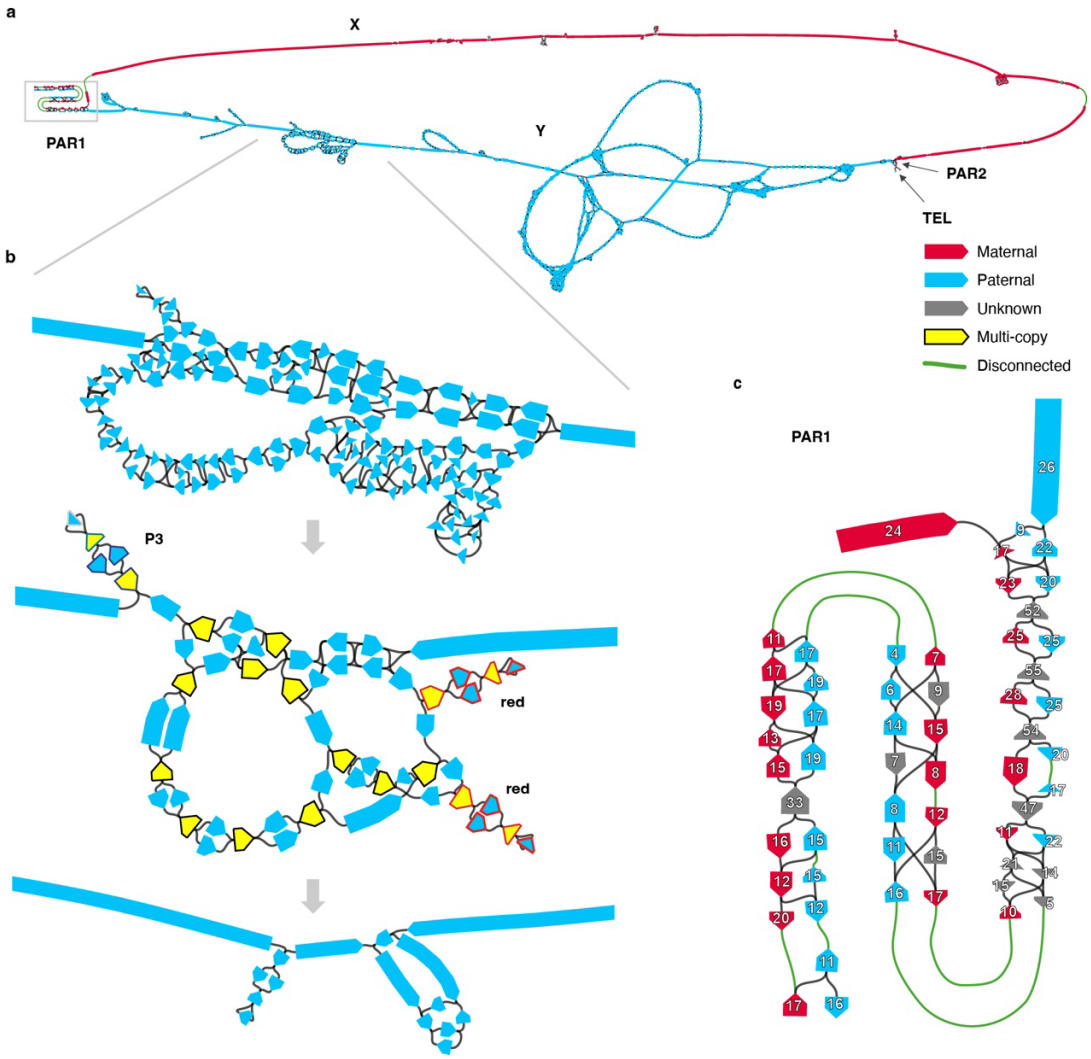
Competing interests

S.N. is now an employee of Oxford Nanopore Technologies; S.K. has received travel funds to speak at events hosted by Oxford Nanopore Technologies; A.F. is an employee of DNAnexus; C.-S.C. is an employee of GeneDX Holdings Corp.; N.-C.C. is an employee of Exai Bio; L.F.P. receives research support from Genetech; F.J.S. receives research support from Pacific Biosciences, Oxford Nanopore Technologies, Illumina, and Genetech; K.S. is an employee of Google LLC and owns Alphabet stock as part of the standard compensation package; W.T. has two patents (8,748,091 and 8,394,584) licensed to Oxford Nanopore Technologies; E.E.E. is a scientific advisory board member of Variant Bio, Inc. All other authors declare no competing interests.

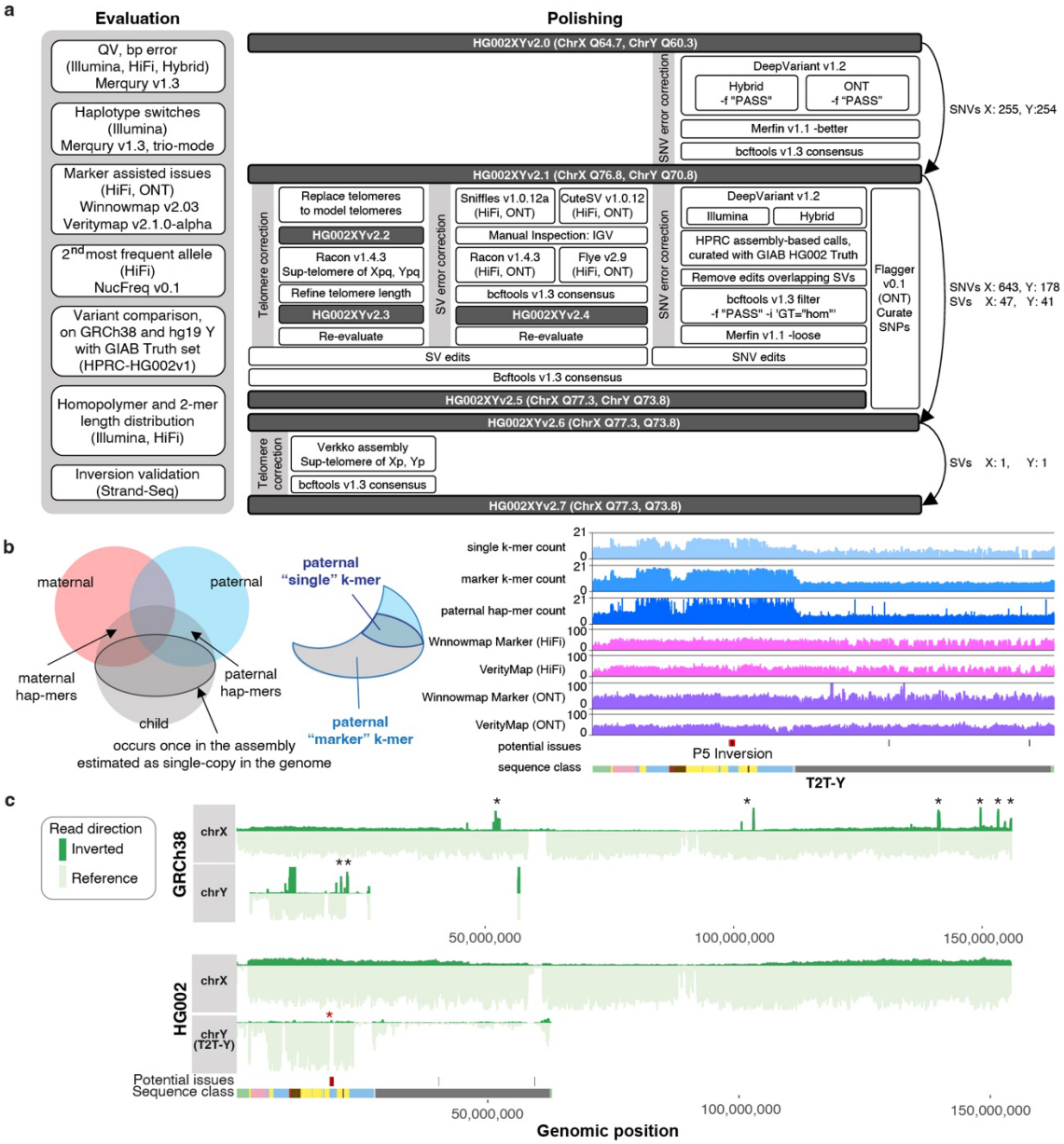
Additional information

SupplementaryTables.xlsx file contains **Table 1** and **Supplementary Tables 1-32**. **SupplementaryInformation.pdf** file contains **Supplementary Methods, Supplementary Figs. 1-19**, and **Supplementary Notes 1-4**. **SupplementaryFiles.zip** contains **Supplementary Files 1-3**.

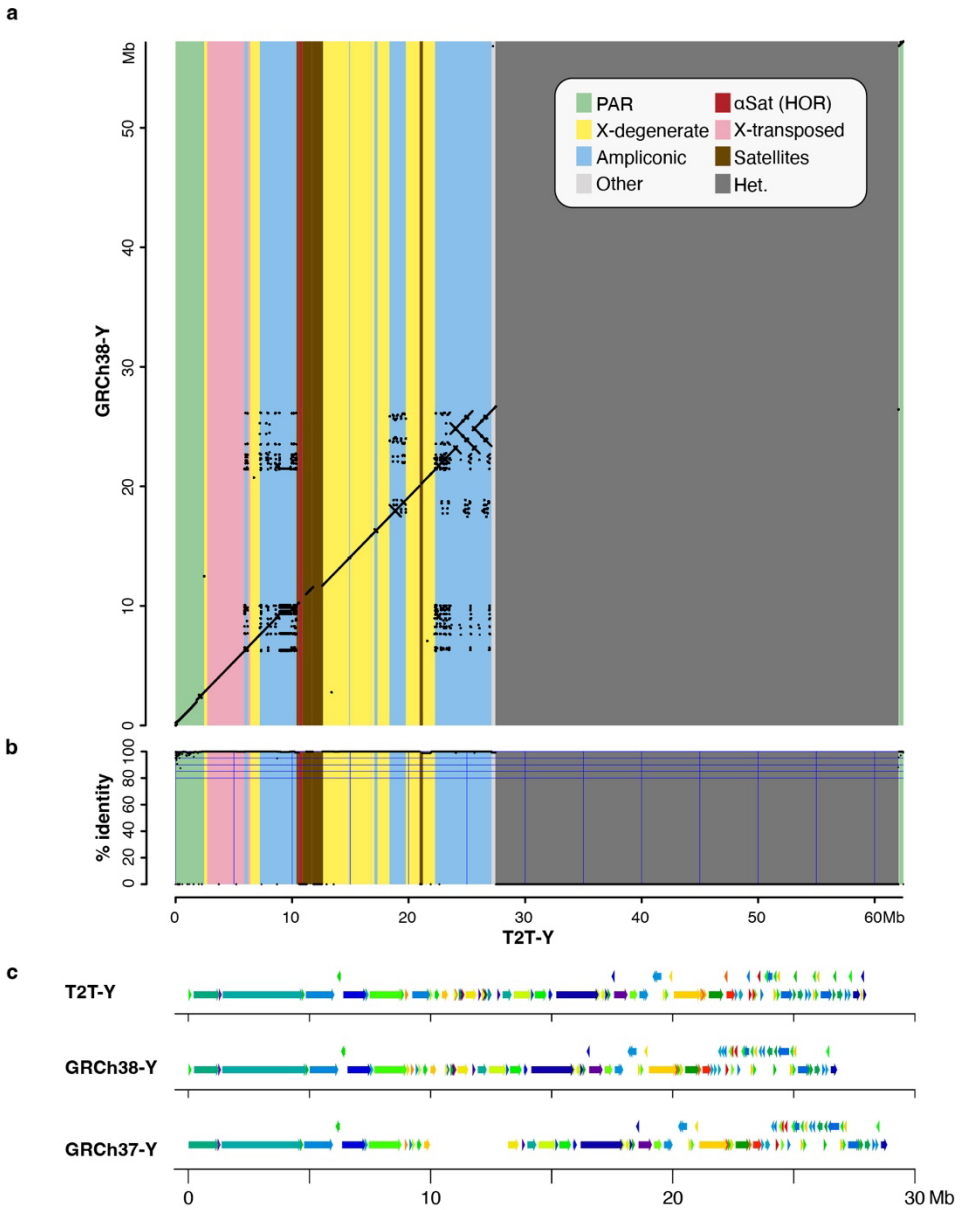
Extended Data Figures



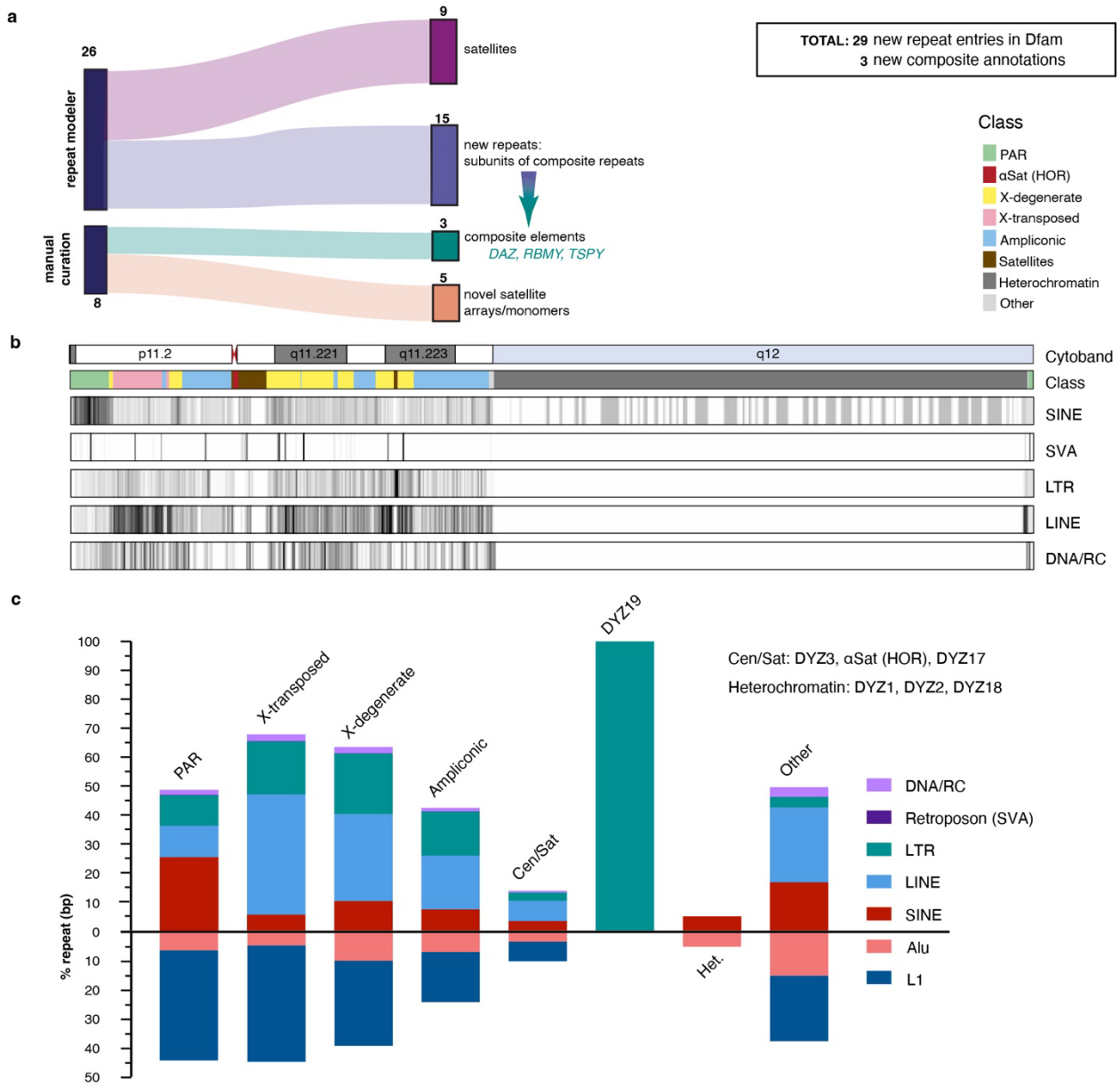
Extended Data Fig. 1 | Assembling the X and Y chromosomes of HG002. a. Chromosome X and Y components of the assembly string graph built from HiFi reads, detected based on node sequence alignments to T2T-CHM13 and GRCh38 references. Each node is colored according to the excess of paternal-specific (blue) and maternal-specific (red) k-mers, obtained from parental Illumina reads, indicating if they exclusively belong to chromosome Y or X, respectively. Most complicated tangles are localized within the heterochromatic satellite region on the Y q-arm. The X and Y subgraphs are connected in PAR1 and PAR2. Graph discontinuities are due to a lack of HiFi sequence coverage in these regions caused by contextual sequencing bias, with 9 out of 11 observed breaks falling within PAR1 on either chromosome (5 out of 5 for chromosome Y). Note that for visualization purposes the length of shorter nodes is artificially increased making the extent of the tangles appear larger than reality. **b**. The effects of manual pruning and semi-automated ONT read integration is illustrated from top to bottom. Top, zoomed in view of a tangle encoding the P1–P3 palindromic region in Y (approx. 22.86–27.08 Mb, see **Fig. 4**). Middle, corresponding subgraph following the manual pruning and recompaction. Nodes excluded from the curated "single-copy" list for automated ONT-based repeat resolution are shown in yellow. Three hairpin structures are highlighted, which form almost-perfect inverted tandem repeats encompassing the entire P3 and two P2 (red) palindromes. Node outlines in the palindromes are colored according to the palindromic arms as in **Fig. 4**. Bottom, corresponding subgraph following the repeat resolution using ONT read-to-graph alignments. Remaining ambiguities were resolved by evaluating ONT read alignments to all candidate reconstructions of the corresponding sub-regions. **c**. PAR1 subgraph labeled with HiFi read coverage on each node. Gaps (green edges) and uneven node coverage estimates indicate biases in HiFi sequencing across the region. **Fig. 1** shows an enrichment of SINE repeats and non-B DNA motifs in PAR1 that may underlie the sequencing gaps in this region.



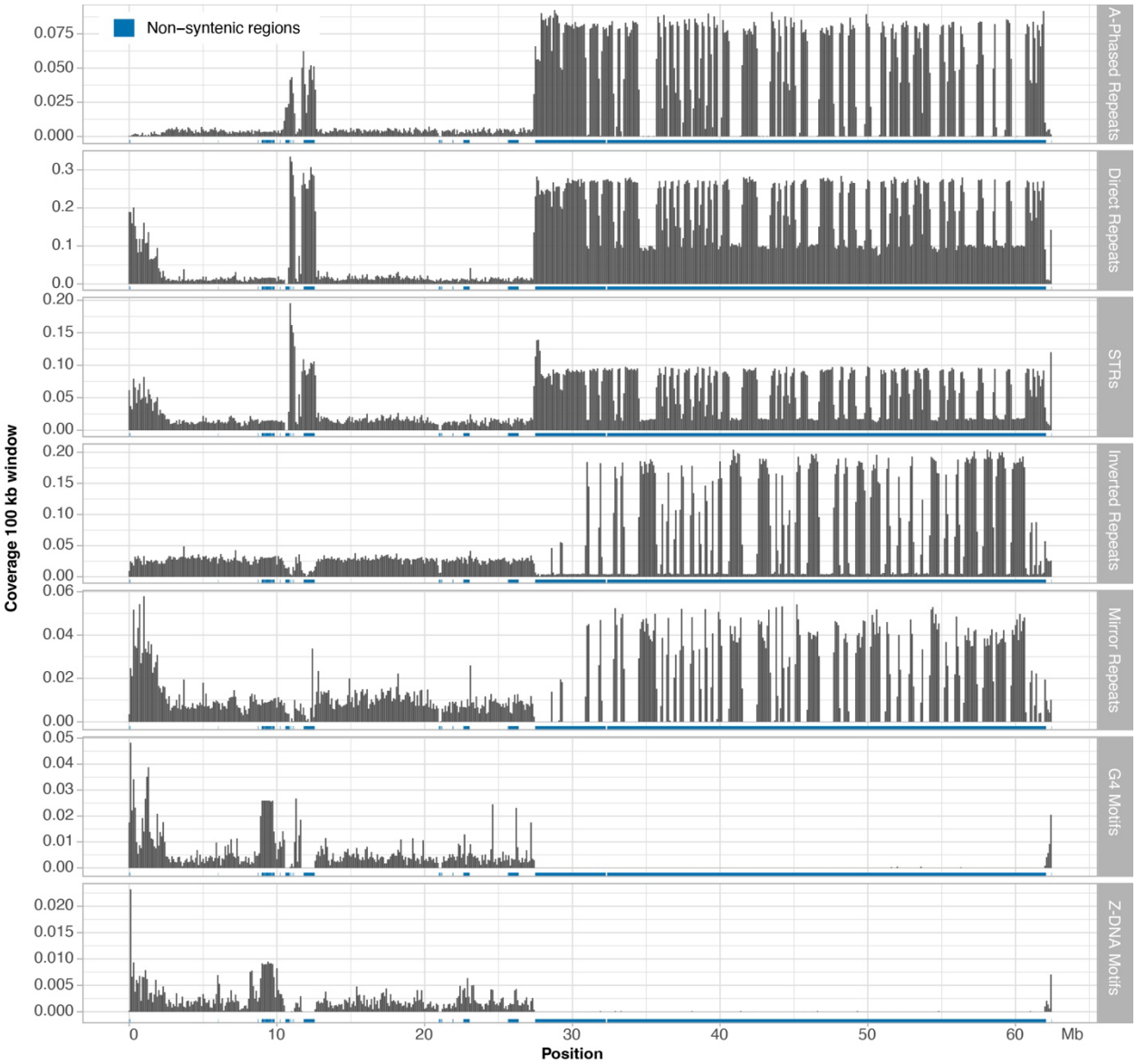
Extended Data Fig. 2 | Validation and polishing of the T2T-Y. a. Evaluation and polishing workflow performed on T2T-CHM13v1.1 autosomes + HG002 XY assemblies. **b**. Venn diagram of the k-mers from the parents and child. On the left, hap-mers[18] represent haplotype specific k-mers inherited by the child. The darker outlined circle inside the child k-mers represent single-copy k-mers (k-mers occurring once in the assembly and single-copy in the child's genome). Right figure shows an example of the paternal specific, "single-copy" and "marker" k-mers. The marker set includes both multi-copy and single-copy k-mers specific to the paternal haplotype that were inherited by the child. Unlike polishing the nearly haploid CHM13 assembly[17], both single-copy k-mers and marker k-mers were used for the marker-assisted alignments to HG002 XY. This helped align more reads within repetitive regions to the correct chromosome for evaluation during polishing. Right panel shows counts of the k-mers and coverage of HiFi and ONT reads using the marker-assisted Winnowmap2 alignment, in addition to alignments from VerityMap, which uses locally unique k-mers for anchoring the reads. **c.** Aggregated Strand-seq coverage profile across all 65 libraries on GRCh38-Y (top) and T2T-Y (bottom). Each bar represents read counts in every 20 kb bin supporting the reference in forward direction (light green) or reverse direction (dark green). Multiple spikes in reverse direction (black asterisks) in GRCh38-Y indicate inversion polymorphisms relative to HG002, likely due to differences between the haplogroups. Such spikes in coverage are not observed on T2T-X and T2T-Y, which confirm the structural and directional accuracy of the HG002 assemblies. A 3 kb inversion of the unique sequence between the P5 palindromic arms was identified as erroneous in T2T-Y (red asterisk), but was confirmed to be polymorphic in the population and left uncorrected in this version of the assembly.



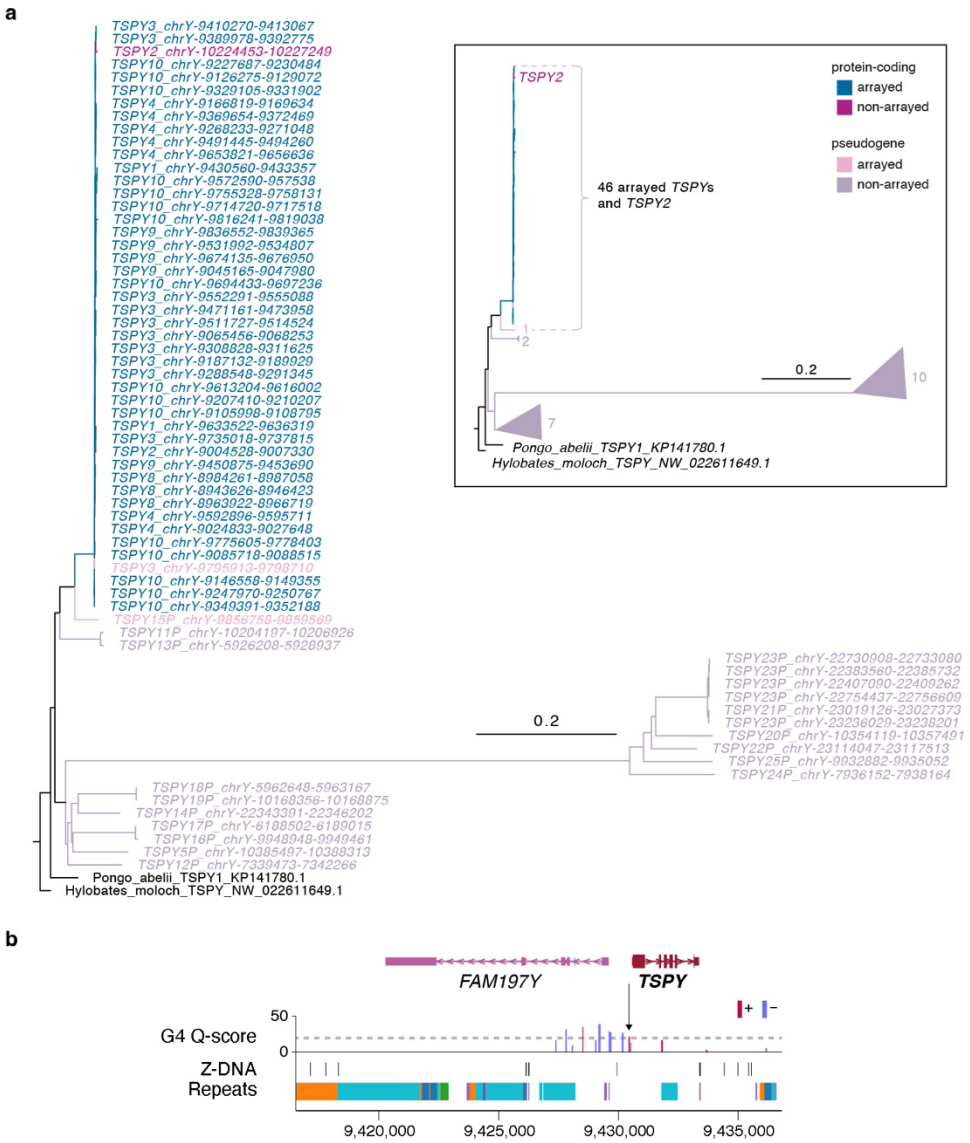
Extended Data Fig. 3 | Large structural differences between T2T-Y and previous GRCh Y assemblies. **a-b.** Ampliconic genes and X-degenerate sequences revealed from alignments between GRCh38-Y (Y-axis) and T2T-Y (X-axis). **a.** Dotplot generated using LastZ[93] after softmasking with WindowMasker[94]. **b.** Identity was computed from matches and mismatches over positions with alignments, excluding gaps. **c.** Structural differences revealed using PRG-TK[95] against GRCh38-Y and GRCh37-Y in the euchromatic region of the Y chromosome.



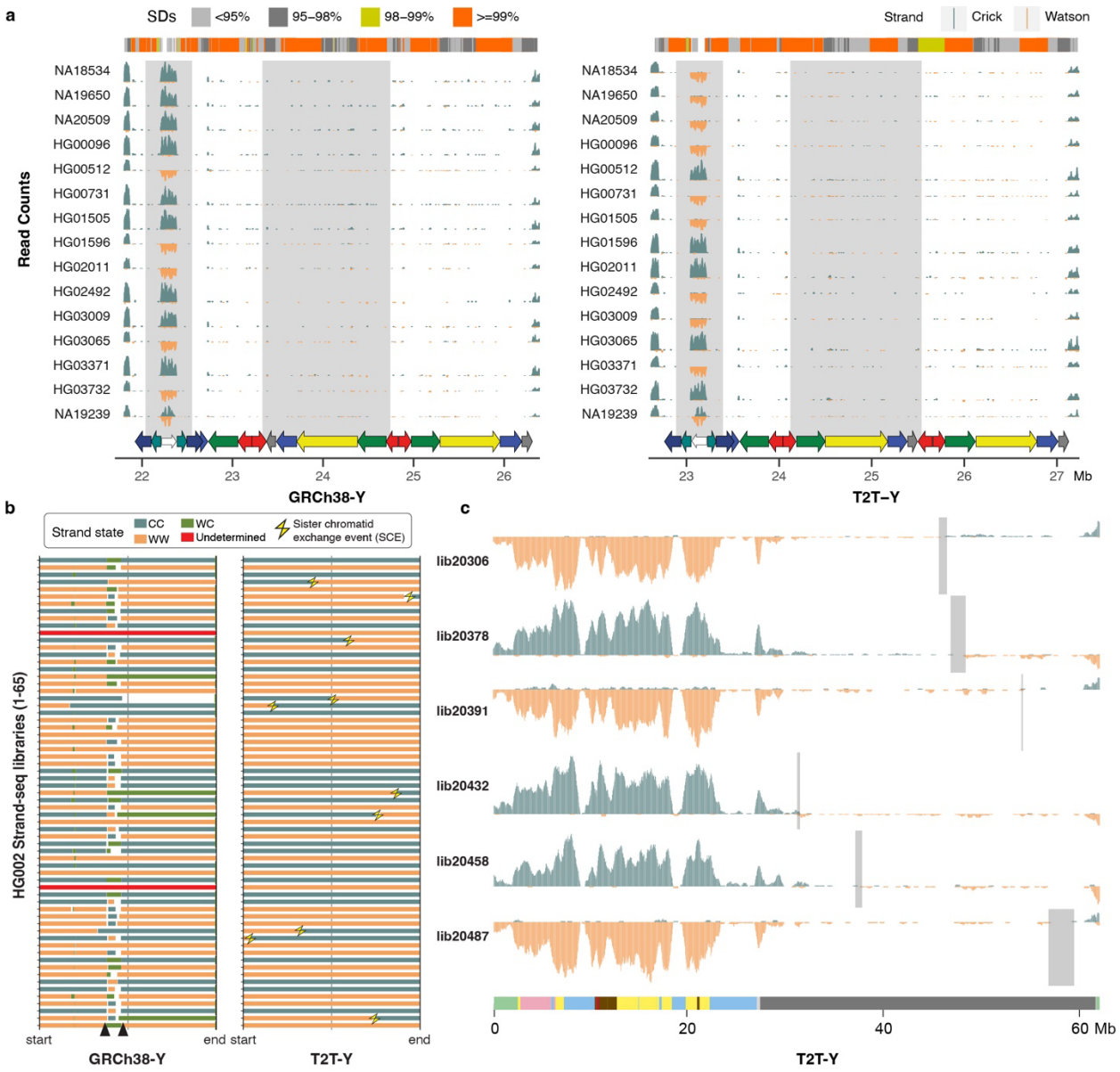
Extended Data Fig. 4 | Repeat discovery and annotation of T2T-Y. a. Assembly completion allowed for a full assessment of repeats and resulted in the identification of previously unknown satellite arrays (predominantly in the PAR1) and subunit repeats that fall within one of three composite repeat units (*TSPY*, *RBMY*, *DAZ*). **b.** Ideogram of TE density (per 100 kb bin). This is an extension of **Fig. 1** with non-SINEs expanded into separate TE classes (SVA, LTR, LINE, DNA/RC). Density scale ranges from low (white, zero) to high (black, relative to total density) and sequence classes are denoted by color. **c.** Summary (in terms of base coverage per region) across all five TE classes and two specific families: *Alu*/SINE and L1/LINE. The satellites in (**b**) were kept separate as two categories; Cen/Sat as the left satellite block including alpha satellites and DYZ19, while all other categories were combined per sequence classes.



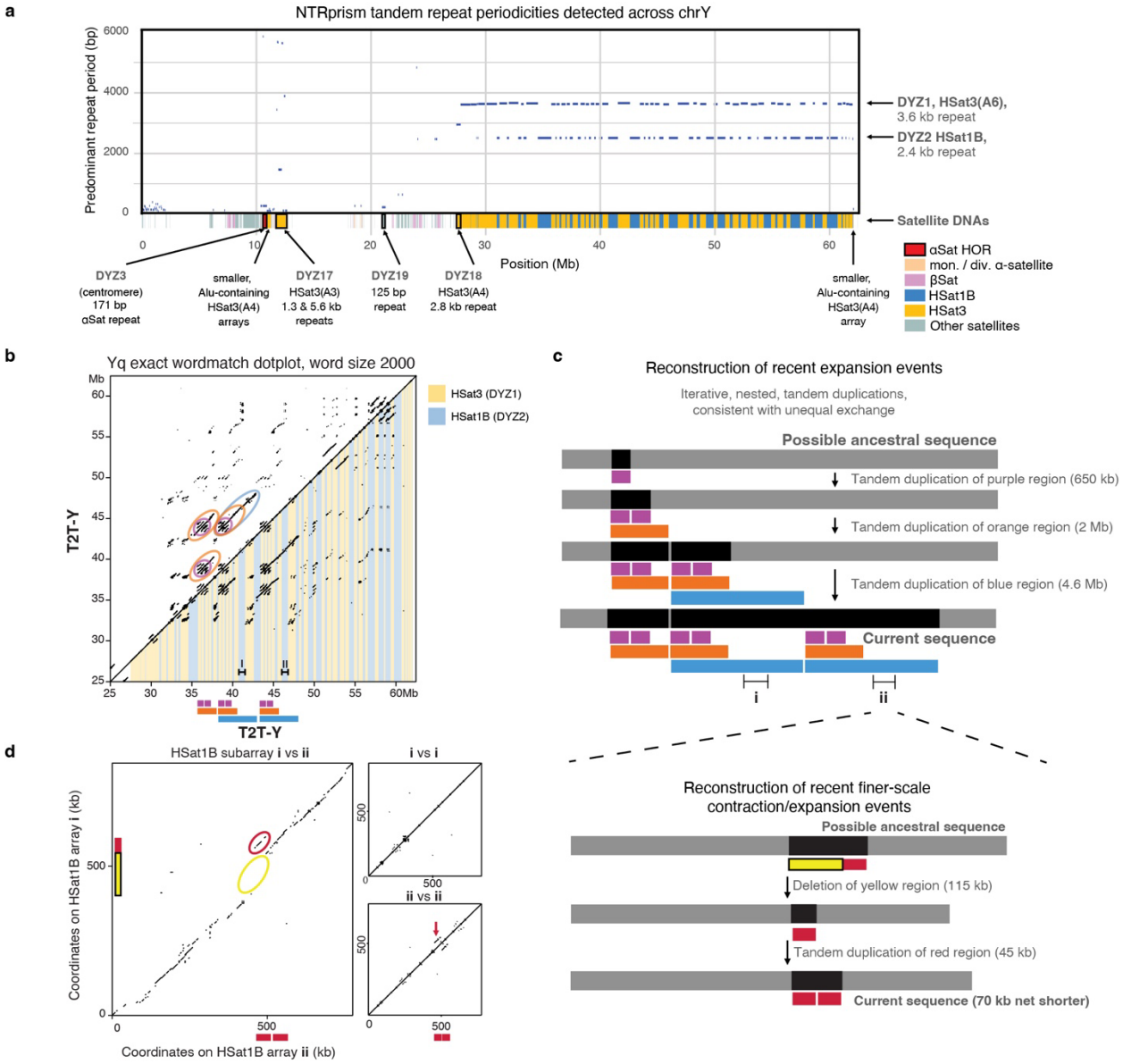
Extended Data Fig. 5 | Non-B DNA motifs along the T2T-Y. HSat3 on the Yq and satellite sequences around the centromere are more enriched with A-phased repeats, direct repeats and STRs, while HSat1B is more enriched with inverted repeats and mirror repeats. Enrichment of non-B DNA sequences were also observed in the PAR region. Notably, the *TSPY* gene array is enriched for G4 and Z-DNA motifs, as shown in **Extended Data Fig. 6b**.



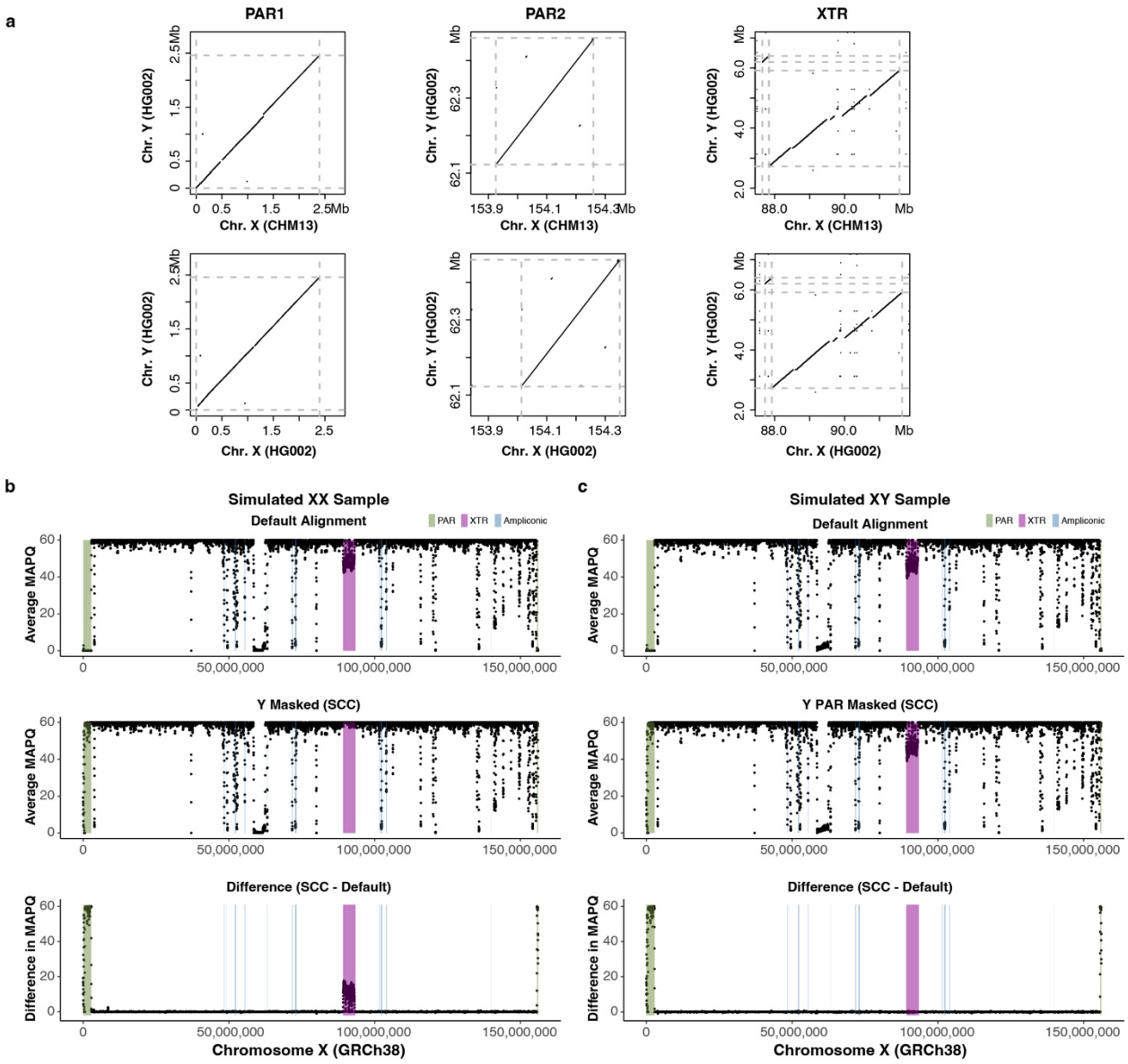
Extended Data Fig. 6 | Phylogenetic tree analysis of the ampliconic *TSPY* gene family and pattern of non-B DNA structure. **a.** Phylogenetic tree analysis using protein-coding *TSPY*s from a Sumatran Orangutan (*Pongo abelii*) and a Silvery gibbon (*Hylobates moloch*) as outgroups confirmed *TSPY2* (distal to the array) and *TSPY* copies within the array originated from the same branch, distinguished from the rest of the *TSPY* pseudogenes. Rectangular inset shows a cartoon representation of the simplified tree. Numbers next to the triangles indicate the number of *TSPY* genes in the same branch. **b.** G4 and Z-DNA structures predicted for a typical *TSPY* copy inside the *TSPY* array. All *TSPY* copies in the array have the same signature, with one G4 peak present ~500 bases upstream of the *TSPY* (arrow). Higher Quadron score[122] (Q-score) indicates a more stable G4 structure, with scores over 19 considered stable (dotted line).



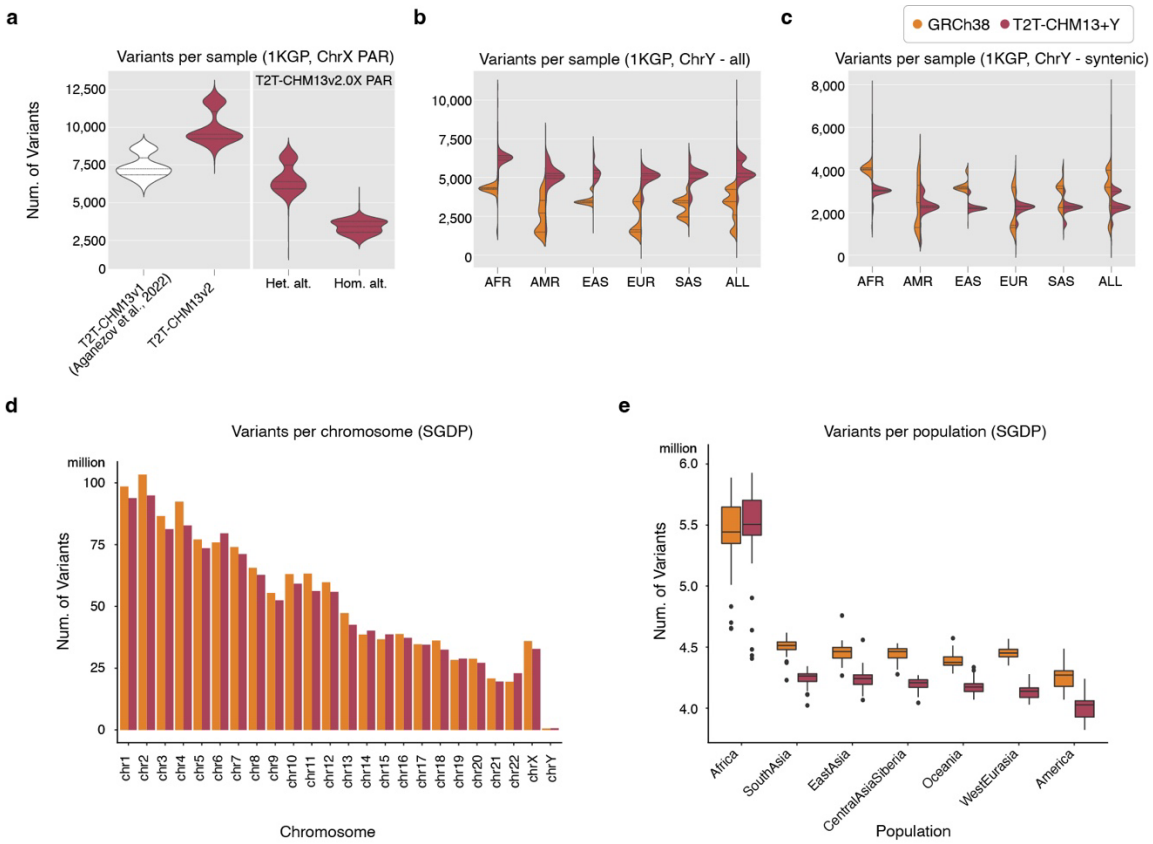
Extended Data Fig. 7 | Recurrent inversions identified with Strand-seq. a. Five out of 15 individuals have the inverted variant as present in HG002 at the P3 palindrome (white arrow). Although inversions across P1–P2 (yellow and red arrows) are difficult to confirm with Strand-seq because of the high sequence similarity between the palindromic arms, different orientations are observable in these samples. **b.** Strand states for 65 Strand-seq libraries of HG002. Depending on the mappings of directional Strand-seq reads (+ reads: 'Crick', C, - reads: 'Watson', W), reference sequence was assigned in three states: WC, WW, and CC. WC, roughly equal mixture of plus and minus reads; WW, all reads mapped in minus orientation; CC, all reads mapped in plus orientation. Changes in strand state along a single chromosome are normally caused by a double-strand-break (DSBs) that occurred during DNA replication[159] in a random fashion and we refer to them as sister-chromatid-exchanges (SCEs, yellow thunderbolts). Recurrent change in strand state over the same region in multiple Strand-seq cells indicates misassembly. Similarly, collapsed or incomplete assembly of a certain genomic region will result in a recurrent strand state change as observed for GRCh38-Y (black arrowheads). In contrast, T2T-Y shows strand state changes randomly distributed along each Strand-seq library with no evidence of misassembly or collapse. **c.** Strand-seq profile of selected libraries over T2T-Y summarized in bins (bin size: 500 kb, step size: 50 kb). Teal, Crick read counts; orange, Watson read counts. As ChrY is haploid, reads are expected to map only in Watson or Crick orientation. Light gray rectangles highlight regions where SCEs were detected in the heterochromatic Yq12 despite a lower coverage of Strand-seq reads. A modified breakpointR parameter was used (windowsize = 500000 minReads = 20) in order to refine detected SCEs presented in panel **b** and **c**.



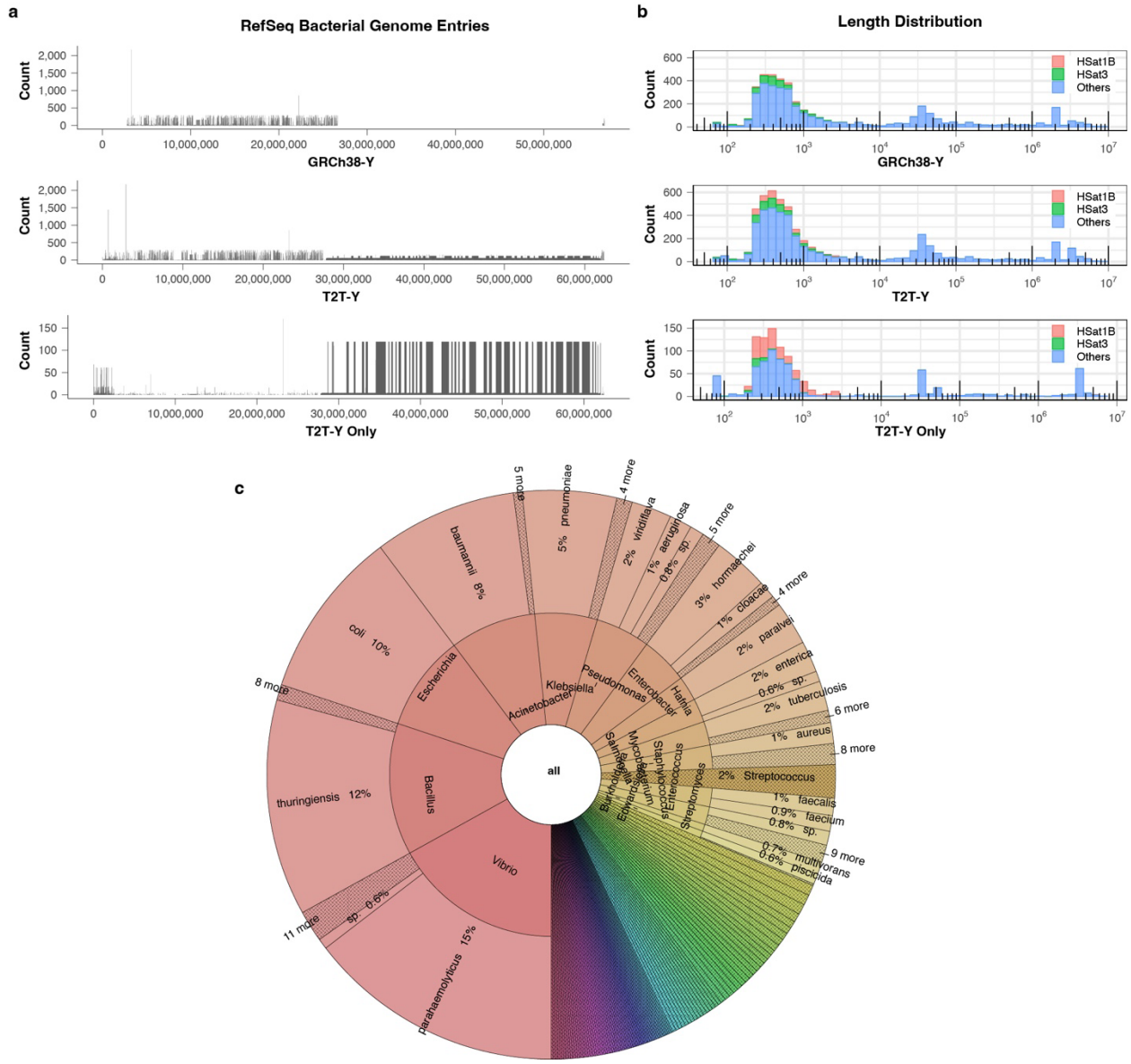
Extended Data Fig. 8 | Satellite annotation and recent expansion events in the Yq heterochromatin. **a.** A plot showing the top repeat periodicities detected by NTRprism[46] in 50 kb blocks tiled across T2T-Y, with centromeric satellite annotations overlaid on the X axis. Large arrays are labeled with their historic nomenclature[1], HSat subfamilies[64], and predominant repeat periodicities. **b.** An exact 2000-mer match dotplot of the Yq region (a dot is plotted when an identical 2000 base sequence is found at positions X and Y). The lower triangle has DYZ1/DYZ2 annotations overlaid as yellow and blue bars, respectively. Circled patterns in the upper triangle correspond to recent iterative duplication events, which are illustrated below the X axis. **c.** A reconstruction of a possible sequence of recent iterative duplications that could explain the observed dotplot patterns. **d**. A 2000-mer dotplot comparison of two ~800 kb HSat1B sub-arrays that were part of a recent large duplication event, along with self-self comparisons of the same arrays, revealing sites of more recent and smaller-scale deletions and expansions (annotated in yellow and red, with a possible sequence of events illustrated by the schematic on the right).



Extended Data Fig. 9 | Genomic similarity in PARs and XTR and improved MAPQ of the PARs through informed sex chromosome complement reference. a. Dotplots from LASTZ alignments of the CHM13-X, HG002-X, and HG002-Y (T2T-Y) over 96% sequence identity. Dashed gray lines represent the start and end of the approximate PARs or XTR boundaries. Disconnected diagonal lines indicate the presence of genomic diversity between each paired region. More genomic differences are observed in the PAR1 between the HG002-Y and CHM13-X. **b-c.** Average mapping quality (MAPQ) across GRCh38-X from simulated reads of an XX (**b**) and XY (**c**) sample. Top, a default version of GRCh38 (with two copies of identical PARs on XY)**.** Middle, a version of GRCh38 informed on the sex chromosome complement (SCC) of the sample (entire Y hard-masked for the XX sample vs. only PARs on the Y hard-masked for the XY sample). Bottom, the difference in average MAPQ between the SCC and default approaches. MAPQ was averaged in 50 kb windows, sliding 10 kb across the chromosome. A positive value means MAPQ score is higher with SCC reference alignment compared to default alignment.



Extended Data Fig. 10 | Number of variants called from 1KGP and SGDP individuals. a. More variants are called on the X-PARs when using the sex chromosome complement reference approach (calling variants in diploid mode on PARs) than the non-masked approach (calling variants in haploid mode on PARs). The 1KGP results for GRCh38-Y are from Aganezov et al.[69], which was performed on CHM13v1.0+GRCh38-Y. **b.** Num. of variants called from each 1KGP XY sample on chromosome GRCh38-Y and T2T-Y **c.** Num. of variants called in the syntenic region between the two Ys. A large num. of additional variants are called on each sample attributed to the newly added, non-syntenic sequences on T2T-Y. Within the syntenic regions, a reduction in the number of variants is observed for each population except for samples from R1 haplogroups as shown in **Fig. 6c**. **d.** Aggregated total number of variants for the 279 SGDP samples per chromosome. **e.** SGDP genome-wide counts of variants per-sample (n=279) demonstrate increased variation in African samples regardless of reference. Each bar in the box plot represents the 1st, 2nd (median), and 3rd quartile of the number of variants in each population. Whiskers are bound to the 1.5 × interquartile range. Data outside of the whisker ranges are shown as dots. For the SGDP samples, variants were called using T2T-CHM13+Y or GRCh38 as the reference. All variants shown in this figure were filtered for “high quality (PASS)”.



Extended Data Fig. 11 | Human contaminants in bacterial reference genomes. a. Number of distinct RefSeq accessions in every 10 kb window containing 64-mers of GRCh38-Y (top), T2T-Y (middle), and in T2T-Y only (bottom). Here, RefSeq sequences with more than 20 64-mers or matching over 10% of the Y chromosome are included. **b.** Length distribution of the sequences from (**a**) in log scale. Majority of the shorter (<1 kb) sequences contain 64-mers found in HSat1B or HSat3. **c.** Number of bacterial RefSeq entries by strain identified to contain sequences of T2T-Y and not GRCh38-Y, visualized with Krona[158].