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 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 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, 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. Sema4, OpCo, Inc, 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, 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
[#] Retired
[*] Corresponding author. Email: adam.phillippy@nih.gov

Abstract

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. 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. 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*; 42 additional protein-coding genes, mostly from the *TSPY* gene 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 and mapped available population variation, clinical variants, and functional genomics data to produce a complete and comprehensive reference sequence for all 24 human chromosomes.

Introduction

The human Y chromosome plays critical roles in fertility and hosts genes important for spermatogenesis in males, as well as *SRY*, the mammalian sex determining locus[1]. However, in the human reference genome, GRCh38, the Y chromosome remains the most incomplete chromosome, with >50% of bases represented by gaps. Most of these multi-megabase gaps represent sequences flanking the endogenous model centromere, parts of the ampliconic regions, and large heterochromatic regions which have been challenging to resolve for decades. The architecture of the Y chromosome, specifically the presence of tandemly arrayed and large, inverted repeats with nearly identical arms (i.e. palindromes)[2], 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[3].

Following the first complete assemblies of chromosomes X[4] and 8[5], the Telomere-to-Telomere (T2T) consortium successfully assembled all chromosomes from a complete hydatidiform mole with a 46,XX karyotype (CHM13)[6]. This first complete human genome assembly was enabled by innovative technological improvements in generating PacBio high-fidelity reads (HiFi)[7] and Oxford Nanopore ultra-long reads (ONT)[8], the development of better assembly algorithms for utilizing HiFi reads and generating assembly graphs[9], the use of ONT reads for better resolving the graph[10], new methods for validating and polishing[11–15], and a coordinated curation effort to finish the assembly.

In parallel, with the goal of including broader genomic diversity across populations[16], the Human Pangenome Reference Consortium (HPRC) has evaluated various methods for generating high-quality diploid genome assemblies[17] using a well characterized human genome, HG002, which is commonly used for benchmarking by the Genome in a Bottle consortium[18]. Using this rich set of data, and integrating the lessons learned from previously assembling CHM13, we successfully reconstructed the complete sequence of the HG002 Y chromosome, hereafter referred to as T2T-Y.

Using the complete sequence of T2T-Y, we investigated the newly assembled pseudoautosomal regions (PARs), X-degenerate and ampliconic sequences, and the heterochromatic satellite and repeat sequence compositions. We have fully 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 generated a 1-to-1 alignment of orthologous blocks between T2T-CHM13+Y and GRCh38, and lifted over available variation datasets from ClinVar[19], GWAS[20], dbSNP[21] and gnomAD[22]. In addition, we have recalled variants from 1000 Genomes Project (1KGP)[23] and Simon Genomes Diversity Panel (SGDP)[24] data, as well as epigenetic profiles from ENCODE data[25], using the new assembly as a reference. These experiments demonstrate improved mappability and variant calling for XY individuals and reveal the complete sequence of a human Y chromosome for the first time.

Results

Assembly, validation, and annotation of T2T-Y

The Y chromosome assembly process generally followed the strategy used for the previous T2T-CHM13 assembly[6] (**Supplementary Table 1** and **Supplementary Fig. 1**). Briefly, an assembly string graph was first constructed using PacBio HiFi reads at ~30x depth of coverage and processed using custom pruning procedures (note: all coverage statistics are given relative to the haploid sex chromosomes). Due to high sequence similarity within PAR1 and PAR2, the HG002 ChrX and ChrY string graph components shared connections to one another, but to no other chromosomes in the genome (**Extended Data Fig. 1a**). The remaining tangles in these subgraphs were resolved using ~45x coverage of ONT reads longer than 100 kb. To facilitate this step, the pipeline was enhanced by a semi-automated repeat resolution strategy, utilizing the read-to-graph alignment paths (**Extended Data Fig. 1b**). PAR1 was strongly affected by HiFi sequencing biases leading to a less resolved and more fragmented graph due to reduced read quality and depth of coverage. 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. Coverage gaps caused by HiFi sequencing biases were patched using a *de novo* assembly of trio-binned paternal ONT reads[11].

The ChrY draft assembly was further polished and validated using sequencing reads from Illumina (33x), HiFi (42x), and ONT (125x). During four rounds of polishing, 1,520 small and 10 large (>50 base) errors were detected and corrected (**Extended Data Fig. 2a**). Small corrections were identified with DeepVariant[26,27] and filtered with Merfin[11]. Large errors were identified with Sniffles[28], cuteSV[29], and through a comparison to the HPRC-HG002v1 assembly[17]. All of the large errors localized to the PAR1 and telomeric regions, and were patched using selected HiFi and ONT reads. Long-read alignments filtered with globally unique[14] and locally unique markers[30] identified three potential assembly issues remaining in the HSat arrays around positions 40 Mb, 53.1 Mb, and 59.1 Mb (**Extended Data Fig. 2b**, **Supplementary Table 2, Supplementary Figs. 2-4**), and Strand-seq[31,32] identified only one inversion error within a palindromic sequence (P5) around position 18.8 Mb (**Extended Data Fig. 2c**). The remaining sequences showed no signs of collapse or false duplication, with even HiFi coverage (mean 39.3 ± 12.5) except for regions associated with known sequencing biases[14], all of which had supporting ONT coverage (reads over 25 kb, mean 78.1 ± 13.6). Because the validation signal at the three HSat positions was ambiguous, these regions were noted but left unchanged. The P5 inversion error was discovered only after the T2T-Y assembly had been fully annotated and released, and because this inversion appears as a true recurrent inversion in other individuals[33], it was noted but left uncorrected in this release. The described T2T-Y assembly is 62,460,029 bases in length with no gaps or model sequences and an estimated error rate of less than 1 error per 10 Mb (Phred Q73.8), as measured by Merqury using a hybrid k-mer set from Illumina and HiFi reads[14,15] (**Table 1**, **Supplementary Table 3**). This T2T-Y assembly (derived from HG002) was combined with the T2T-CHM13v1.1 assembly to create a new Y-bearing reference assembly, T2T-CHM13v2.0, referred to here as T2T-CHM13+Y.

Table 1 | Comparison of GRCh38-Y and T2T-Y. 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. 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. Num. exclusive genes/transcripts are those found in one assembly but not the other. Blank spaces indicate not applicable.

		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	65	107	+64.6
	Num. exclusive genes	0	110	
	Protein coding	0	42	
	Num. transcripts	681	888	+30.4
	Protein coding	372	493	+32.5
	Num. exclusive transcripts	0	210	
	Protein coding	0	124	
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 (46, 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	30.6	84.9	

Comparison to GRCh38-Y

T2T-Y reveals the previously uncharacterized ~30 Mb of sequence within the heterochromatic region on the long arm (Yq). In comparison, 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 gaps representing the heterochromatic blocks and sub-telomeric or satellite repeat sequences. The shorter GRCh38-Y sequence (NT_187395.1) is 37.2 kb in length, assigned to ChrY but left unlocalized (i.e. not placed in the primary Y chromosome assembly). 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 gap-excluded sequence identity of ~99.8% in the aligned regions, but with multiple structural differences including an incorrectly oriented centromere model for GRCh38-Y (**Fig. 1** and **Extended Data Fig. 3**).

To understand the genomic differences between T2T-Y and GRCh38-Y, we first identified each respective Y-chromosome haplogroup, determined by mutations that accumulate in the non-recombining portion of the male-specific region (MSY)[2]. Using yhaplo[34], which utilizes phylogenetically significant SNPs to build a tree and compares that to the ISOGG database, 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[35] and Europeans[36], respectively, consistent with the established ancestry of these genomes.

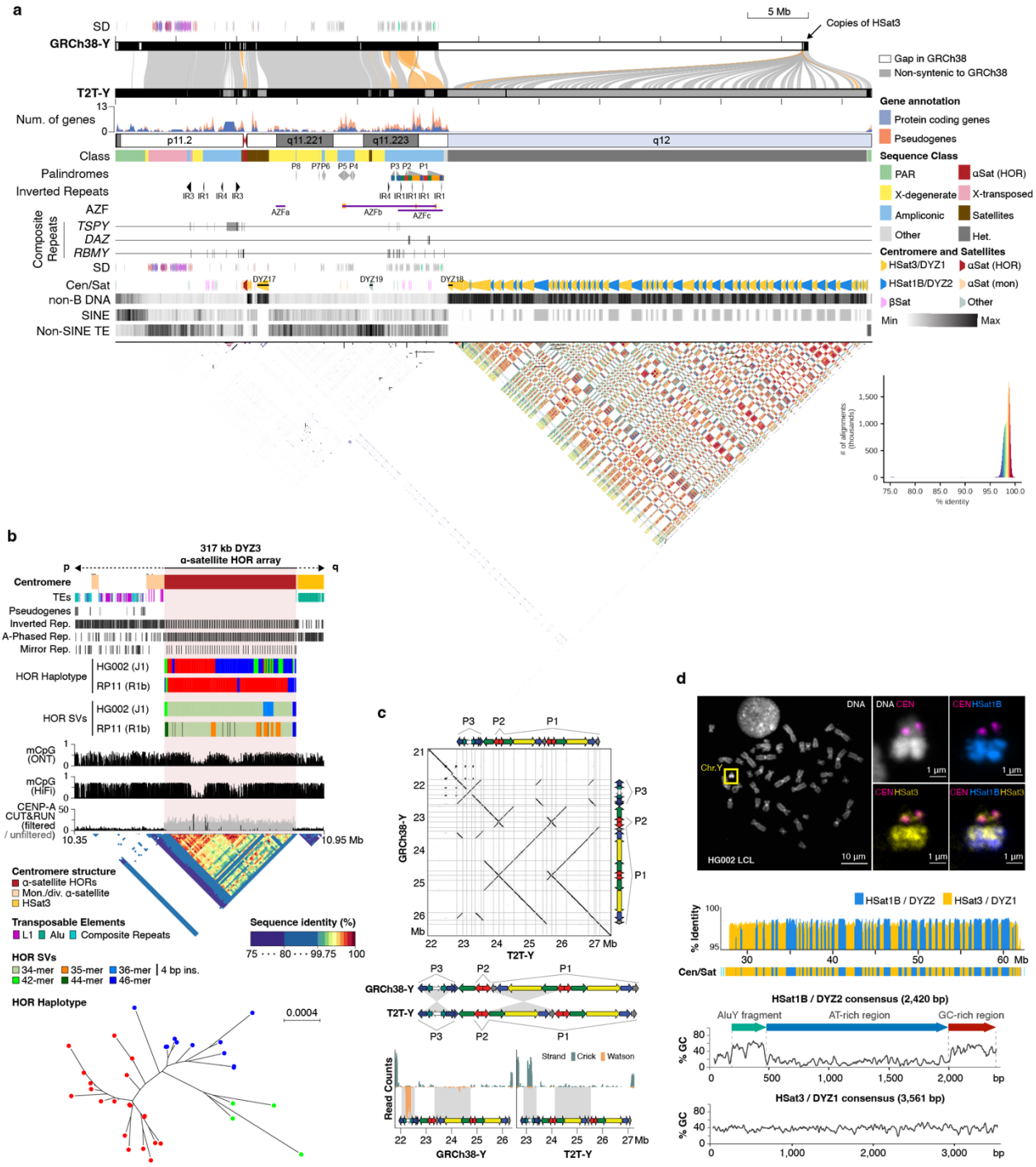


Fig. 1 | The structure of a complete Y chromosome. a. Direct comparison between GRCh38-Y and T2T-Y visualized with SafFire[37]. Regions with sequence identity over 95% are connected and colored by alignment direction (gray, forward; orange, reverse). Segmental duplications (SD) are colored by duplication types defined in DupMasker[38]. Centromere and satellite annotations (Cen/Sat) highlight the alternating HSat1 and HSat3 pattern comprising Yq12. Below, tandem repeats within T2T-Y are visualized by StainedGlass[39] with similar repeats colored by %identity in the style of an alignment dotplot. **b.** The structure of the T2T-Y centromere. No TEs were found within the DYZ3 HOR array, while L1s (proximal) and *Alu*s (distal) were found within the diverged alpha satellites (drawn taller than the other TEs). Density tracks show the prevalence of non-B DNA motifs within the 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[40]. Histograms show the fraction of methylated CpG sites called by both ONT and HiFi, and CENP-A binding signal from CUT&RUN[41]. StainedGlass plot illustrates that the repeats within the array are highly similar (99.5–100%). Tree of HOR units uses the same color coding as above. **c.** Comparison of the palindromic structure of the P1–P3 region in GRCh38-Y and T2T-Y by alignment dotplot and schematics. At bottom, Strand-seq signal from HG002 mapped against GRCh38-Y and T2T-Y confirms the two inversions in P3 and P1. **d.** Top-left panel shows 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). Maximum intensity projections are shown in all panels. Middle, % identity of each DYZ2/DYZ1 repeat unit to its consensus sequence. Bottom plots show the %GC sequence composition of the DYZ2 and DYZ1 repeat units and the position of an ancient *Alu*Y fragment in DYZ2.

Repeat annotation

Next, we generated comprehensive repeat annotations, incorporating repeat models previously updated with CHM13[42], as well as 29 previously unknown repeats identified in T2T-Y (**Extended Data Fig. 4a, Supplementary Table 4**). In GRC38-Y, only 30.58% (17.5 Mb) of the chromosome was annotated as repetitive compared to 84.86% (53 Mb) in T2T-Y (**Table 1, Supplementary Table 5**). While short interspersed nuclear elements (SINEs), specifically *Alu*s, are found embedded as part of the 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. 1a**). Moreover, TE distribution biases typify different subregions of ChrY, as *Alu*s are enriched in the PAR1 region, while other TEs (particularly L1s) are much more abundant in the X-chromosome-transposed region (XTR)[1](**Extended Data Fig. 4b** and **Supplementary Table 6**). The DYZ19 region is annotated by RepeatMasker as entirely LTRs (**Extended Data Fig. 4c**), but further sequence analyses indicate this is a satellite array spanning 265 kb and whose 125 base monomeric consensus is derived from an expanded portion of a LTR12B sequence[43]. 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 7**).

A total of 31,166 transposable element (TE) annotations were lifted from T2T-Y to GRCh38-Y, while 7,764 were unlifted, representing 2,653,492 bases of TE sequence (**Supplementary Table 8**). Of those unlifted, 98% reside in the large gaps in GRCh38 while 2% represent either polymorphic loci between the two Y chromosomes or a small collapse in the GRCh38 reference (**Supplementary Fig. 5**).

Gene annotation

We annotated T2T-Y by mapping RefSeq (v110) and GENCODE (v35) annotations from GRCh38 using Liftoff[44] and performed hand-curation of the ampliconic gene arrays (**Supplementary Table 9**). Additional Iso-Seq transcriptomes from HG002 were collected from B-Lymphocyte cells, induced pluripotent stem cell (iPS) cell-lines derived from these lymphocyte cells, and iPS cell-lines derived from blood cells for annotation (**Supplementary Table 1, Supplementary Fig. 6-7**), and used for alternative *de novo* annotations by NCBI RefSeq and EBI Ensembl along with tissue-specific expression data from other publicly available sources.

Our annotation of T2T-Y totals 693 genes and 888 transcripts, of which 107 genes (493 transcripts) are predicted to be protein-coding (**Table 1** and **Supplementary Table 10**). Only six genes differed in their annotation between GRCh38-Y and T2T-Y. Four of these genes were properly annotated in both genomes, but assigned a different paralogous gene name in T2T-Y due to presumed sequence-level differences. The other two differentially annotated genes are pseudogenes of *TSPY* that are absent from T2T-Y. These pseudogenes are adjacent to the *TSPY* assembly gap in GRCh38 and likely to be false annotations caused by assembly errors in this region (**Supplementary Table 11**). In addition to containing all genes annotated in GRCh38-Y, T2T-Y contains an additional 110 genes, among which 42 are predicted to be protein coding. The majority of these protein-coding genes (39 of 42) are additional copies of *TSPY*, filling the corresponding gap in GRCh38-Y (**Table 1**).

A typical human Y chromosome harbors 16 single-copy protein-coding X-degenerate genes, with housekeeping functions and homologs on the X chromosome; and 9 protein-coding ampliconic gene families, which have expanded specifically on the Y, are expressed in testis, function in spermatogenesis, and are associated with fertility[45] (**Table 1**). Along with the ampliconic gene annotations, we have estimated copy numbers of each ampliconic gene family in both the GRCh38-Y and T2T-Y assemblies using an adapted application of AmpliCoNE[46]. Copy number of these gene families was previously estimated for HG002 using Illumina reads and droplet digital PCR (ddPCR)[46]. These results are largely concordant with copy number in our T2T-Y assembly measured by simulated Illumina reads from the assembly and *in silico* PCR primer search (**Supplementary Table 12-15**). The only notable difference was in the *TSPY* copy number, in which we identified 46 intact protein-coding copies. The copy number was slightly higher in the assembly than the estimates derived from Illumina reads and ddPCR (46 vs. 40 and 42, respectively). The *in silico* PCR primer search matched all 45 protein coding copies in the *TSPY* gene array and *TSPY2*, and one pseudogene at the 3' end of the *TSPY* array which we were unable to avoid in the ddPCR primer design. We conclude that our AmpliCoNE and ddPCR/*in silico* PCR estimates are in agreement with the ampliconic gene annotations in the T2T-Y assembly (**Table 1**). RNA-Seq data used in the RefSeq gene annotation confirmed expression of the ampliconic genes in testis[47].

Centromere

Normal human centromeres are enriched for an AT-rich satellite family (~171 base unit, monomer), known as alpha satellite, typically arranged into higher-order repeat (HOR) structures and surrounded by more diverged alpha and other satellite classes[48]. The prior T2T-CHM13 assembly includes fully assembled centromeres for all human chromosomes, with the exception of ChrY. However, owing to its relatively short length, a prior assembly of the RP11 ChrY centromere was previously completed with ONT sequencing of the RPCI-11 BAC library[40]. Here we characterize the sequence and methylation of the T2T-Y centromeric region (J1 haplogroup) and compare it with RP11 (R1b haplogroup).

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%), a detailed analysis of the full-length repeat unit identified three HOR subtypes (red, blue, and green) and a different

organization compared to RP11 (**Fig. 1b, Supplementary Figs. 8-12** and **Supplementary Tables 16-17**). The majority of the T2T-Y centromeric array is composed of 34-mers 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). In RP11, no 36-mer variants are present, but a number of 35-mers containing internal duplication are observed (**Fig. 1b**).

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. 1b**), which has been reported to coincide with the CENP-A binding and is the putative site of kinetochore assembly[48]. 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. 13**). A similar pattern of multiple methylation dips within a single centromere was observed in other T2T-CHM13 chromosomes such as Chr11 and Chr20[49].

Rearrangements in ampliconic palindromes

We confirmed that the structure of the newly assembled T2T-Y is consistent with expectations based on previous studies[2,50,51]. We annotated regions on the T2T-Y as ampliconic, X-degenerate, X-transposed, pseudoautosomal, heterochromatic, and other, in accordance with Skaletsky *et al.*[2], but adding a more precise annotation for the satellites (including DYZ17 and DYZ19), and the centromere (**Fig. 1a** and **Supplementary Table 18**). The X-degenerate and ampliconic regions were estimated to be 8.67 Mb and 10.08 Mb in length, respectively. As expected, the ampliconic region contained eight palindromes, with palindromes P4–P8 highly concordant between T2T-Y and GRCh38-Y (i.e. in terms of arm, spacer lengths, and sequence identity). Arm-to-arm identity of these five T2T-Y palindromes, which are nested within X-degenerate regions, ranged from 99.84 to 99.96% (**Supplementary Table 19-20**). Palindromes P1–P3 are located within the AZFc region, which harbors genes critical for sperm production[52]. We discovered a large polymorphic inversion (>1.9 Mb) that likely arose from a single non-allelic homologous recombination event. Using Strand-seq, we were able to pin-point the breakpoints at two *red* amplicons (naming according to Kuroda-Kawaguchi *et al.*[53]): one forming the P2 palindrome and the other inside the P1 palindrome (**Fig. 1c**). This inversion 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[54]. Another inversion was detected in P3, which was recently reported as a recurrent variation in humans[55] (**Extended Data Fig. 5a**). Inversions between amplicons are believed to serve as substrates for subsequent AZFc deletions and duplications that might affect sperm production[50,51,54,56].

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[57–60]. The single largest gap in GRCh38-Y is at Yq12, with minimal representation of DYZ1 and DYZ2, mostly in unplaced scaffolds. Here, we have uncovered the detailed structure of the Yq12 region at single-base resolution, adding over 20 Mb of DYZ1 and

14 Mb of DYZ2 repeats to the reference sequence. The complete Yq12 assembly also enabled mapping of sister chromatid exchanges within this heterochromatic region, which we observed in approximately 15% of Strand-seq libraries (**Extended Data Fig. 5b**).

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[61]. DYZ2 is composed of an unrelated satellite family, HSat1B, which is also present in smaller amounts on the acrocentric short arms[62] and comprises a ~2.5 kb tandem repeat made up of three parts: an ancient *Alu*Y fragment (20% diverged from the *Alu*Y consensus), an extremely AT-rich region (>85% A/T), and a more GC-rich region[60,61,63]. We derived new consensus sequences for each repeat type and found the vast majority of repeat instances to be over 98% identical to their consensus, with slightly higher divergence at the more peripheral satellite blocks (**Fig. 1d**).

While HSat1B carries an *Alu*Y-derived subunit as part of its composite repeat unit (**Fig. 1d**), 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 (**Extended Data Fig. 6**). 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 expansion and homogenization locally.

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, it appears that the initial seeding events that created these arrays were independent of one another, yet were derived from *Alu*Y elements from PAR1.

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. 1d**). Yq12 is highly variable in size and sequence structure between individuals[64–66], and the number and size of these satellite blocks is expected to vary considerably. 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. 7**). These structural rearrangement patterns are consistent with evolution by unequal exchange mechanisms.

Structure of the *TSPY* ampliconic gene family

In GRCh38-Y, *TSPY* protein-coding genes are placed in an array between 9.3 to 9.6 Mb with a ~50 kb gap in the middle and most copies left unresolved[2] (**Fig. 2**). An additional protein-coding copy, *TSPY2,* is placed at 6.2 Mb in GRCh38-Y, in the proximal inverted repeat (IR3), upstream

of the array. In contrast, our T2T-Y assembly resolved 46 protein-coding *TSPY* copies, including *TSPY2*, which was found in the distal part of the IR3, downstream of the *TSPY* array (at ~10 Mb). 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[33]. All 45 protein-coding copies in the *TSPY* array were embedded in an array of composite repeat units, with one composite unit (~20.2 kb in size) per gene, such that an array of composite units includes multiple *TSPY* gene copies in tandem (**Fig. 2a** and **Supplementary Table 21**). Each composite unit also 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 human genome 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[42]).

The copy number of protein-coding *TSPY* genes in the array was polymorphic across different male samples, ranging from approximately 10–40 copies as estimated from the SGDP data (**Fig. 2b**). Phylogenetic analysis confirmed all *TSPY* protein coding copies (including *TSPY2*) originated from the same branch, distinguished from the rest of the *TSPY* pseudogenes (**Fig. 2c**). This result contradicts earlier findings[67], which concluded that *TSPY2* originated from a different lineage. However, given the presence of misassemblies at this locus in GRCh38 and earlier ChrY references, additional complete non-human primate assemblies are needed to definitively reconstruct the evolutionary history of *TSPY*. Non-B DNA motifs were also assessed within the *TSPY* composite (**Fig 2d**), as described in a latter section.

Other composite repeat arrays

In addition to the *TSPY* composite repeat array, we characterized all other composite repeats in T2T-Y, including defining the repeats and array units in and around the two ampliconic gene families, *RBMY* and *DAZ* (**Supplementary Table 21**). 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[42]. 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. 2e**). 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. 2f**). Out of the three composite arrays described herein on the Y, *DAZ* is the only one also found on an autosome (Chr3, *DAZL*), although it is only found as a single unit rather than an array, and lacks the young LINE1 (L1PA2) insertion that ChrY *DAZ* copies carry.

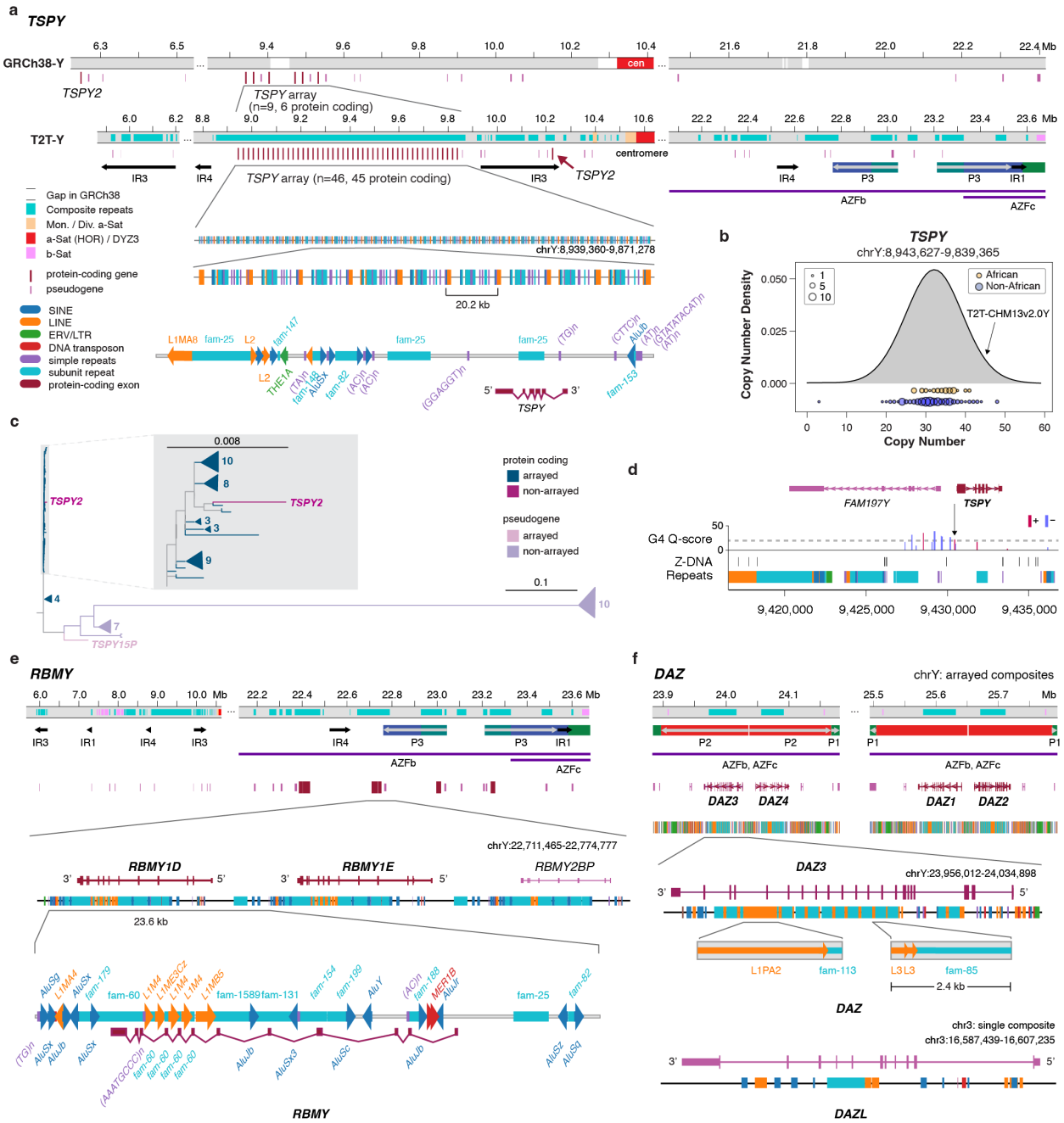


Fig. 2 | Ampliconic genes forming composite repeats. a. T2T-Y has 45 *TSPY* protein-coding genes organized in a single continuous array and a single *TSPY2* copy, compared to GRCh38-Y which has a gap in the *TSPY* array surrounded by misassembled copies annotated as pseudogenes. T2T-Y shows a more regularized 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**. Phylogenetic tree analysis of the *TSPY* gene family. Numbers next to the triangles indicate the number of *TSPY* genes in the same branch. **d.** G4 and Z-DNA structures predicted for a *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[68] (Q-score) indicates a more stable G4 structure, with scores over 19 considered stable (dotted line). **e.** Repeat composition of the *RBMY* gene family. **f**. 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.

Transduced genomic segments

Transcriptionally active transposons, especially long interspersed element 1 (L1) and short variable number tandem repeat interspersed elements (SINE-VNTR-*Alu*s, SVA), occasionally co-mobilize downstream DNA into a new locus by bypassing a canonical poly(A) termination signal in a process termed 3' DNA transduction[42,69–71]. In contrast, SVAs can also produce 5' DNA transductions by hijacking alternative promoters[42,72]. Transduction activity has been recognized as a driver of genome shuffling that includes protein-coding sequences, regulatory elements, and even whole genes[70,71,73]. Additionally, transductions can occur in soma, contributing to tumorigenesis, and frequent 3' L1-mediated DNA transductions have been observed in some cancer types[74]. To identify potential DNA transduction activities in T2T-Y, we searched for transductions mediated by L1s and SVAs (**Supplementary Methods**).

We detected six potential 3' L1 transductions within the Y, yet no SVA-driven DNA transductions (**Supplementary Table 22**). Our results show that four L1s carrying transduced segments are full-length elements (>6 kb), two of which possess a canonical poly(A) termination signal within 30 bases upstream of a poly(A) tail. The other two L1s are truncated elements with 3'-transduction signatures, consistent with prior predictions[70] since many retroposed L1s are truncated at the 5' end. We found that five L1 transductions were shared with GRCh38-Y, while one is specific to T2T-Y. Despite a genome-wide investigation of both T2T-CHM13+Y and GRCh38, we were not able to locate the potential donor elements of the transduced segments according to our criteria (e.g., they were not within 20 bases downstream of a full-length L1 in a different locus).

To investigate whether any source elements within T2T-Y gave rise to DNA transductions onto other chromosomes, we probed previously reported L1 and SVA transductions[42]. However, we were not able to detect any sign of the source elements. Given the ancient form of L1s annotated on the Y, we confirm a recent analysis[75] of 1KGP data that found no evidence for DNA transduction between the Y and the rest of the chromosomes. Our analysis revealed that the transduction rate in T2T-Y was 0.096 per 1 Mb, which is much lower than the transduction rate observed in the CHM13 autosomes (avg. 6.9 per 1 Mb) and ChrX (10.19 per 1 Mb)[42]. In conclusion, our results indicate that transposable element driven transductions are not abundant in the Y, and traffic of these events is low between this chromosome and the rest of the genome.

Non-B DNA motifs

We next located Y chromosome motifs capable of forming alternative DNA structures (non-B DNA) such as bent helix, slipped strand, G-quadruplex (G4s), cruciform, triple helix, hairpin, and Z-DNA structures. Non-B DNA structures are known to affect a variety of important cellular processes, such as replication, gene expression, and genome stability[76–79]. We identified a total of 825,526 sequence motifs with non-B DNA forming potential on T2T-Y, compared to only 138,640 on GRCh38-Y. This nearly 6-fold increase is largely attributed to the newly sequenced heterochromatic and highly repetitive region on the Yq arm (**Fig. 1a**, individual non-B DNA types in **Extended Data Fig. 8**). We found inverted, A-phased, and mirror repeats to be abundant in the centromeric alpha satellite HOR, forming a periodic pattern occurring every 5.7 kb (**Fig. 1b** and **Supplementary Table 23**). An additional pair of A-phased and inverted repeats was present in the extended light blue HOR variant, suggesting possible non-B-form variations per HOR SVs.

The presence of non-B DNA motifs, inverted repeats in particular at the human Y centromere is consistent with the proposed role of such DNA in defining centromeres[80]. Moreover, the per-base-pair density of A-phased, direct, inverted, mirror, and short tandem repeats (STRs) is higher in the newly completed regions of the Y chromosome, and the density of Z-DNA and G4 motifs was particularly high in the newly completed *TSPY* gene array (**Fig. 2d** and **Supplementary Table 23**). Among the 762 new G4 motifs in T2T-Y, 519 are located within the *TSPY* gene array. Specifically, each *TSPY* composite repeat unit contains 14 G4 motifs, seven of which are stable according to their Quadron score (≥19)[68]. Among these, one is located ~500 bases upstream of the transcription start site on the same strand, which might indicate a role in transcriptional regulation[79] (**Fig. 2d**). Along the entire T2T-Y chromosome, 242 G4 motifs overlapped CpG islands (by at least 1 base), again suggesting a role in transcription.

T2T-Y improves variant calling for XY samples

Before using T2T-CHM13+Y as an alternate reference, we investigated whether masking PARs or XTRs on ChrY would improve mapping quality (MQ) and variant calling accuracy on the sex chromosomes. Genetic diversity is higher within PAR1 compared to the non-recombining regions on ChrY[81,82]; however, the ChrX and ChrY PAR sequences in GRCh38 are represented as identical copies[83]. The perfect identity between GRCh38 PARs reduces MQ, hindering accurate variant detection in this region. Previously, the impact of hard-masking the entire ChrY for samples with an XX karyotype was shown to improve mapping quality and increase the number of variants called; however, this was not tested extensively on samples with an XY karyotype[84]. Thus, we simulated 20x coverage of 150 base Illumina reads for ten XY samples (10x coverage for each ChrX and ChrY) seeded with variants called from the high coverage samples in 1KGP, and tested if the PAR masking improves MQ and variant calling accuracy (**Supplementary Table 24-25**).

In the XY samples, the simulated read alignments showed a near-zero MQ on the PARs when no masking was applied, with almost no variants called. In comparison, when masking the PARs on ChrY, reads aligned with improved mapping quality across all samples (example of one sample shown in **Extended Data Fig. 9**), calling an average of 4,615 true positive variants on PAR1 and 365 on PAR2, respectively, with almost no false positives (**Supplementary Table 26**). Additionally, we tested the impact of masking the XTR on ChrY. Unlike the PARs, the false positives substantially increased from an average of 1 to 33,345 across the ChrX XTR (**Supplementary Table 27**), indicating that mapping and variant calling was improved by masking PARs but that XTR masking was detrimental.

After precisely identifying PARs on ChrX and ChrY (**Supplementary Table 18**), we performed short-read alignment and variant calling for 3,202 samples (1,603 XX; 1,599 XY) from the 1KGP, including 1,233 unrelated XY samples averaging at least 30x coverage of 150 base paired-end reads[23]. This set of 1,233 XY samples captures a wide array of genetic diversity comprising individuals from the 26 populations in 1KGP phase 3 and representing 35 distinct Y-chromosome haplogroups (**Supplementary Table 28**). Given our analysis of simulated data, ChrY was completely hard-masked in XX samples, while only the ChrY PARs were masked in XY samples,

forcing any reads sequenced from the ChrY PARs of the samples to align to the ChrX PARs in the reference. Variants in both XX and XY samples could then be called as diploid within the PARs[84]. Other than the optimization for ChrY PARs, the alignment and variant calling pipeline mirrors our previous analysis based on GRCh38-Y[85]. This allowed us to directly compare the effects of using T2T-Y for short-read alignment and variant calling versus using GRCh38-Y.

Across all 1,233 unrelated XY samples, we observed improved mappability on ChrY when using the complete T2T-Y assembly. Specifically, 27.6% more reads per sample mapped to ChrY on average, due in part to the large amount of added sequence in T2T-Y (**Fig. 3a**). We also observed a greater proportion of properly paired reads per sample relative to GRCh38-Y (increase of 1.4% on average, **Fig. 3b**). In addition, the per-read mismatch rate was 62% smaller per sample on average (an absolute decrease of 0.6% on average) when using the T2T-Y assembly (**Fig. 3c**). This metric includes both sequencing errors (either in the read or in the reference), which will be independent across reference genomes, as well as true genetic differences between the read and the reference genome, which is expected to vary by sample ancestry. As such, the decrease in mismatch rate observed across populations is consistent with decreased reference errors in the T2T-Y assembly, and overall, the improvements to mappability demonstrate the utility of T2T-Y for short-read alignment across populations.

From these alignments, we generated variant calls as in the prior analysis, and first compared variant calls in the PARs of XY samples. Unlike our previous approach, all variants in the PARs were now called as diploid, rather than haploid. This resulted in an average 35% increase in the number of variants called per sample in the PARs due to the improved recovery of heterozygous variants (**Fig. 3d**). Next, we sought to determine differences in variant calling across the complete ChrY. For all unrelated samples, we identified 444,584 high-quality ("PASS") variants on T2T-Y compared to 176,150 variants relative to GRCh38-Y, with an increase in the number of variants per sample observed across all super-populations (**Fig. 3e**). However, when restricting this analysis to regions syntenic with GRCh38-Y the overall number of variants called was lower (158,373 on T2T-Y vs. 166,954 on GRCh38-Y). There were also fewer variants per-sample across all super-populations and most Y-chromosome haplogroups, with the greatest reduction observed in samples sharing the same J1 haplogroup as T2T-Y (**Fig. 3f-g**). The samples identified as the R1b haplogroup (the same haplogroup as GRCh38-Y) had more variants called on T2T-Y, and showed the greatest increase in variants called on T2T-Y, as expected. We observed similar mapping and variant calling improvements for SGDP samples (**Supplementary Figs. 14-18**).

Next, we explored if the decreased number of variants called within the comparable syntenic regions between T2T-Y and GRCh38-Y was due to the improved quality of the T2T-Y reference. A simple test is to count the number of variants fixed across all samples (biallelic, alternate allele frequency of 1), which represents private variants or base errors in the reference. In total, 219 fixed variants were observed on GRCh38-Y, compared to only 30 on T2T-Y (0.14% vs. 0.02%, respectively), demonstrating a reduction in the number of private or false variants on T2T-Y relative to GRCh38-Y. In addition, we identified 24,491 variants called on GRCh38-Y and liftable to T2T-Y, which were not called on T2T-Y directly (effectively "disappearing" on T2T-Y). To investigate whether these disappearing variants may be false variants calls, we intersected them

with putative collapsed regions of GRCh38-Y that are better resolved in T2T-Y (**Supplementary Table 29**). These regions include *TSPY*, upstream of IR4, the centromere and its adjacent satellite sequences, DYZ17, DYZ18, DYZ19, and collapsed HSat3 sequences, but excludes palindromic P1-P5 regions with large structural discrepancies (**Fig. 1c**). Notably, we observe 3.4 times as many variants per-base within these collapsed regions relative to the rest of the syntenic regions, which is a signature of falsely collapsed regions[85]. Of the 24,491 disappearing variants, nearly half of them (41.7%, 10,213) fall within these collapsed regions, representing a 6.6-fold enrichment relative to the rest of the variants within the syntenic regions, reinforcing that many of these variant may be false variants caused by the incomplete nature of GRCh38-Y.

Additionally, we noticed a large decrease in the number of indels (insertions/deletions) called on T2T-Y compared to SNPs. Although we found a similar number of SNPs called in both GRCh38-Y or T2T-Y (156,670 vs. 154,249, respectively), we found 6,424 fewer indels on T2T-Y (10,817 on GRCh38-Y vs. 4,393 on T2T-Y, respectively). Many of the variants that “disappear” on T2T-Y are enriched for indels; of the 24,491 disappearing variants, 40.7% (7,235) are indels, a 11.7-fold enrichment relative to the rest of the variants within the syntenic regions. Among these variants, only 126 overlapped the collapsed regions, indicating that the T2T-Y assembly corrects a large fraction of indel-like errors on GRCh38-Y.

To illustrate the effect of variant calling in individual samples, we chose one individual each from the J1, R1b and E1b haplogroups (HG01130, HG00116 and HG01885, respectively) and compared total read depth and the fraction of reads supporting non-reference, alternate alleles (**Fig. 3h**). In all three samples, we observed more variant calls on GRCh38-Y with a higher-than-expected read depth and wide range of alternate allele support. This is a typical signature of mis-mapped reads or collapsed duplications in the reference. We confirmed that regions with excessive coverage and variant calls were located at known collapsed regions described above. In these regions among the 1,233 unrelated XY individuals, 879.8 (SD 107.3) variants were called on average per sample on GRCh38-Y, while only 234.5 (SD 58.5) were called in the corresponding regions in T2T-Y. One example is DYZ19, which is located in the AZFb region and carries multiple variants for Y-chromosome haplogroup determination. This region on GRC38-Y includes a 5 kb gap and is rearranged in comparison to T2T-Y, precluding its genotyping (**Fig. 3i**). However, using T2T-Y, it is possible to identify copy-number changes in this satellite repeat and potentially differentiate alternative haplotypes based on short-read variant calls and depth of coverage (**Fig. 3j**). Similar signatures were found on all the *TSPY* composite repeat units, with signatures of collapsed repeats in GRCh38-Y due to missing and incomplete copies. Taken together, these analyses indicate the complete T2T-Y assembly improves short-read alignment and variant calling across populations and corrects errors in the GRCh38-Y reference.

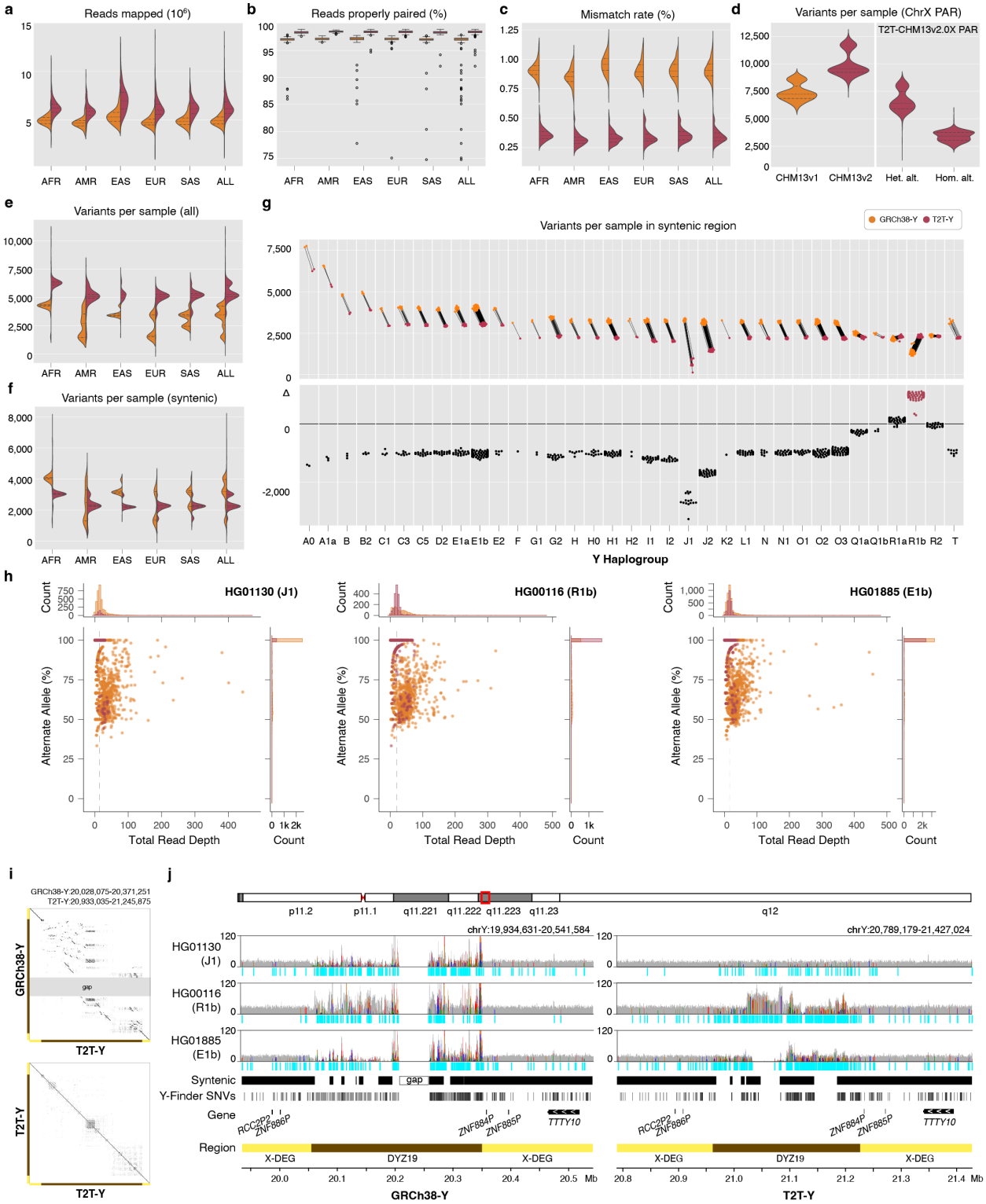


Fig. 3 | Short-read mappability and variant calling improvements on T2T-Y. In all plots, GRCh38 is orange and T2T-Y is maroon. The complete sequence of T2T-Y improves short-read alignment of the 1KGP dataset by **a.** number of reads mapped, **b**. portion properly mapped in pairs, and **c.** lower mismatch rate compared to GRCh38-Y. **d.** More variants are called on the X-PARs in diploid mode than in haploid mode. **e.** A large number of variants are called on the same sample attributed to the newly added, non-syntenic sequences on T2T-Y. Number of called variants within syntenic regions is reduced on T2T-Y, regardless

of super-population (**f**) or Y-haplogroup (**g**), except for the samples identified as R1b (GRCh38-Y haplogroup), with the increase of the variant calls reflecting possible true genetic variation. **h.** 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. **i.** Dotplot of the DYZ19 array between GRCh38-Y and T2T-Y and self-dotplot of T2T-Y drawn with Gepard[86] (word size 100). 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). Read pile-ups on DYZ19 and variants called from the collapsed alleles as shown with IGV[87]. Regardless of the haplogroup, excessive and disconnected read depth on DYZ19 hinders interpretation on GRCh38-Y (gray histogram). For T2T-Y, no large structural or copy number variation is observed on HG001130 (same J1 haplogroup as T2T-Y), while copy number changes are distinguishable for the other two samples. Colors in the coverage tracks represent alternate alleles (>60%). Each light blue bar below the coverage track indicates homozygous alternate variant calls. Syntenic regions between the two Ys are marked in black bands with a 5 kb gap in GRCh38-Y shown as a white box. 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 gray, respectively.

Nearly all known variants are liftable to the T2T reference

Acknowledging the rich amount of resources available on GRCh38, we generated a curated 1-to-1 whole-genome alignment between each GRCh reference (GRCh37 and GRCh38) and T2T-CHM13+Y to enable lifting annotations in either direction. Two different alignment methods were used to obtain homologous regions, and their boundaries were manually inspected. These alignments were split at unaligned segments over 1 kb or gaps greater than 10 kb, and removed when they aligned to non-homologous chromosomes. If there was more than one possible alignment for a locus, a dynamic programming algorithm was used to select the alignment with a higher score. To evaluate the effect of using these alignments (chain files) to lift gene annotations from GRCh38 to T2T-CHM13+Y, we used each chain file to lift GENCODEv35 gene annotations from the primary GRCh38 assembly and compared the results to gene annotations generated directly on T2T-CHM13+Y. In both cases, more than 97% of gene coordinates matched exactly. Based on manual inspection of the two different chain sets and the results of the gene-lifting comparison, we chose the chain files created using one of the two methods for use in further analyses**.** The alignments in this chain set total 747 segments, comprising 2.86 Gb of GRCh38 and T2T-CHM13+Y.

Utilizing this chain file, we sought to lift over three databases of genetic variation from GRCh38 to T2T-CHM13+Y: ClinVar (March 13, 2022 release), dbSNP build 155, and its subset intersecting the GWAS Catalog v1.0 (accessed March 8, 2022). Overall, 99% of both Clinvar and GWAS catalog variants and 98% of all dbSNP variants lifted over to the new reference (**Table 2**). For ChrY, all ClinVar and GWAS catalog variants, and 95% of dbSNP variants, were lifted successfully to T2T-CHM13+Y (**Table 2**). This includes 46.2% of GWAS catalog variants and 0.4% of dbSNP variants whose reference and alternative alleles were swapped between references. The majority of liftover failures on the T2T-Y chromosome (representing 4.7% of all Y chromosome dbSNP variants) are cases in which GRCh38 does not have an orthologous 1:1 position to T2T-CHM13+Y, due to ambiguous mapping between the two references. However, there are a small number of cases (<0.3%) in which a mapping exists between the two references

at the variant's position but T2T-CHM13+Y possesses an allele that is neither the reference nor any of the previously reported alternative alleles (**Supplementary Table 30**).

To further investigate the reasons underlying liftover failure, we intersected all Y chromosome variants in dbSNP whose position did not lift over (117,072, 4.7%) with a set of structurally variable regions between GRCh38 and T2T-CHM13+Y. Out of the 53,158 dbSNP non-PAR variants on the Y that did not lift over, 50,874 (96%) overlapped the structurally variable regions. In comparison, only 28% of the dbSNP variants that lifted over successfully overlapped structurally variable, but unambiguously mappable regions. Thus, we conclude that liftover failures are largely due to copy number and structural differences between the two references. We provide these lifted over datasets within the UCSC genome browser, as well as lists of all variants that failed liftover and the associated reasons.

For T2T-Y, we also provide a list of variants in GRCh38-Y coordinates that are expected to disappear when using T2T-Y as the reference due to the different sample sources. We produce a confident list of 2,314 SNVs and 1,291 indels smaller than 50 bases in 16.6 Mb when restricting to regions with 1:1 alignments between T2T-Y and GRCh38-Y, and not affected by SVs. We also make available the full list of variant calls between T2T-Y and GRCh38-Y, including SVs.

Table 2. Variants lifted over from GRCh38 to T2T-CHM13+Y

Database	Chrs.	GRCh38	T2T-CHM13+Y	%
ClinVar	**Y**	48	48	100.0%
	All	1,122,432	1,113,862	99.2%
GWAS Catalog	**Y**	26	26	100.0%
	All	189,051	186,904	98.9%
dbSNP155	**Y**	2,480,588	2,355,634	95.0%
	All	1,053,463,789	1,029,905,476	97.8%

Human sequence is a common contaminant 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[88,89]. For example, a recent analysis of nearly 5,000 human whole-genome data sets found an unexpected linkage between multiple bacterial species and human males, including 77,647 100-mers that were significantly enriched in the male samples[90]. 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. Using the complete T2T-Y sequence, we explored the contamination hypothesis more thoroughly. We compared male-enriched 100-mers from the Chrisman *et al.* study[90] 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. These results allowed us to estimate the extent of human ChrY contamination in bacterial databases. When counting the potentially contaminated RefSeq bacterial sequence entries matching the GRCh38-Y and T2T-Y, we found 4,179 and 5,148 sequences, respectively (**Fig. 4a,** top and middle). The sequences were relatively short in length (<1 kb), as is typical of contaminating genomic segments (**Fig. 4b**). A total of 1,009 sequences were found only on T2T-Y (**Fig. 4c**, **Supplementary Table 31**), with the vast majority of these sequences localizing to the newly added HSat1B and HSat3 repeats. Such highly repetitive sequences are common sources of contamination because their high copy-number increases the likelihood that they will be accidentally sequenced and assembled. We predict this contamination issue includes sequence from all human chromosomes and extends to all sequence databases, including non-microbial genomes.

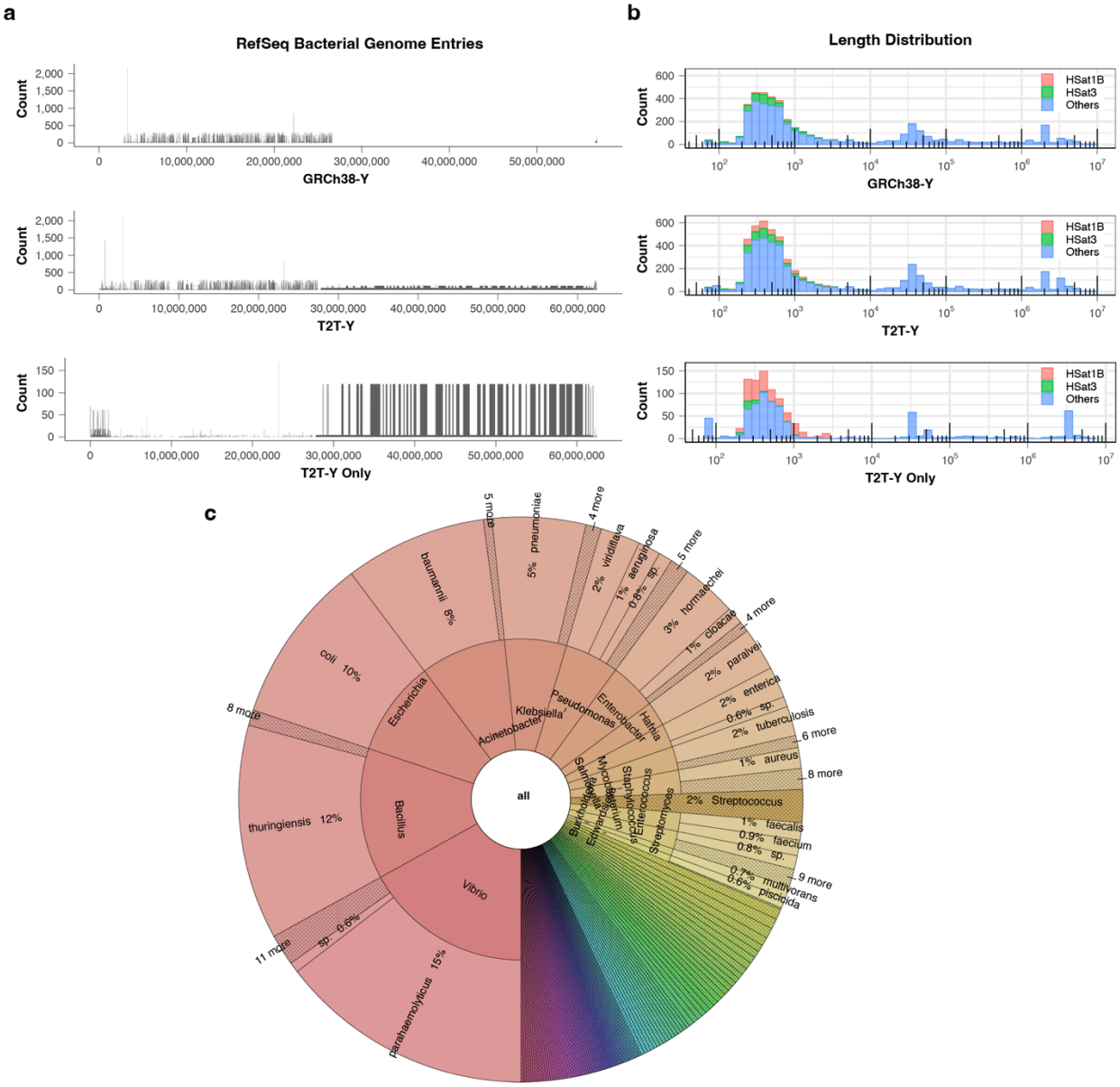


Fig. 4 | Human contaminants in bacterial reference genomes. a. Number of distinct RefSeq accessions in every 10 kb window containing 64-mers of GRCh38-Y (top), T2T-Y (middle), and in T2T-Y only (bottom). Here, RefSeq sequences with more than 20 64-mers or matching over 10% of the Y chromosome are included. **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[91].

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 GIAB 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 immediate use as an alternative reference via NCBI and the UCSC Genome Browser[92] (**Data Availability**).

Our analysis of the T2T-CHM13+Y reference assembly reveals a drastic 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 PAR-masked T2T-CHM13+Y when mapping XY samples.

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, and in some cases erroneous, 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 Yqh region, in which both HiFi and ONT are prone to sequencing biases, hindering accurate biological interpretation (**Supplementary Note** 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 essentially all common bioinformatic analyses.

Construction of the T2T-Y assembly challenged the assembly methods previously developed for the essentially 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[93], which automates the integration of HiFi and ONT data for diploid human genome assembly. The companion study of Hallast *et al.*[33] 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[94] are in the process of generating high-coverage HiFi and ONT sequencing for hundreds of 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.

Data Availability

The T2T-CHM13+Y analysis set and resources are available for download at https://github.com/marbl/CHM13. The assembly, annotation, and associated resources are also available to browse as "hs1" from the UCSC Genome Browser http://genome.ucsc.edu/cgi-bin/hgTracks?db=hub_3671779_hs1 and NCBI data-hub https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_009914755.1/ . The 1KGP and SGDP alignments and variant calls are available within AnVIL at https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T_CHRY. Potential assembly issues are listed and tracked at https://github.com/marbl/CHM13-issues.

Code Availability

Codes used for data analysis and visualization are available at https://github.com/arangrhie/T2T-HG002Y.

Supplementary Information

Supplementary Methods and Supplementary Notes are available in **Supplementary Information**.

Acknowledgements

We thank P. Hallast, 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). 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.); 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.

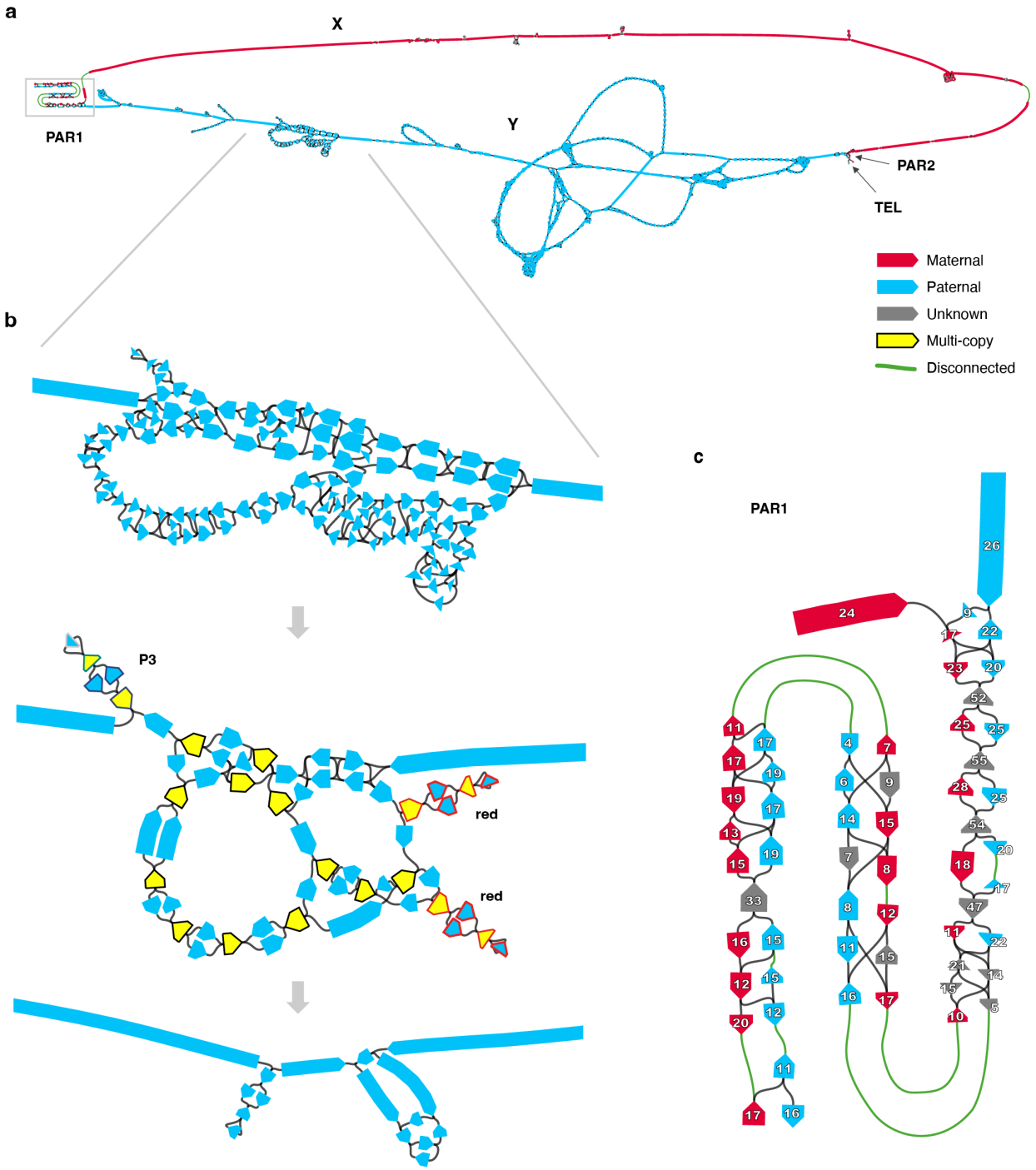
Author contributions

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. Structural annotation: A.R., M.C., H.L., P.M., 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. 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. Epigenetics: A.R., P.W.H., A.Ge., W.T., A.M.W. Retroelements: R.Hal., W.M. Non-B DNA: M.H.W., K.D.M. 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

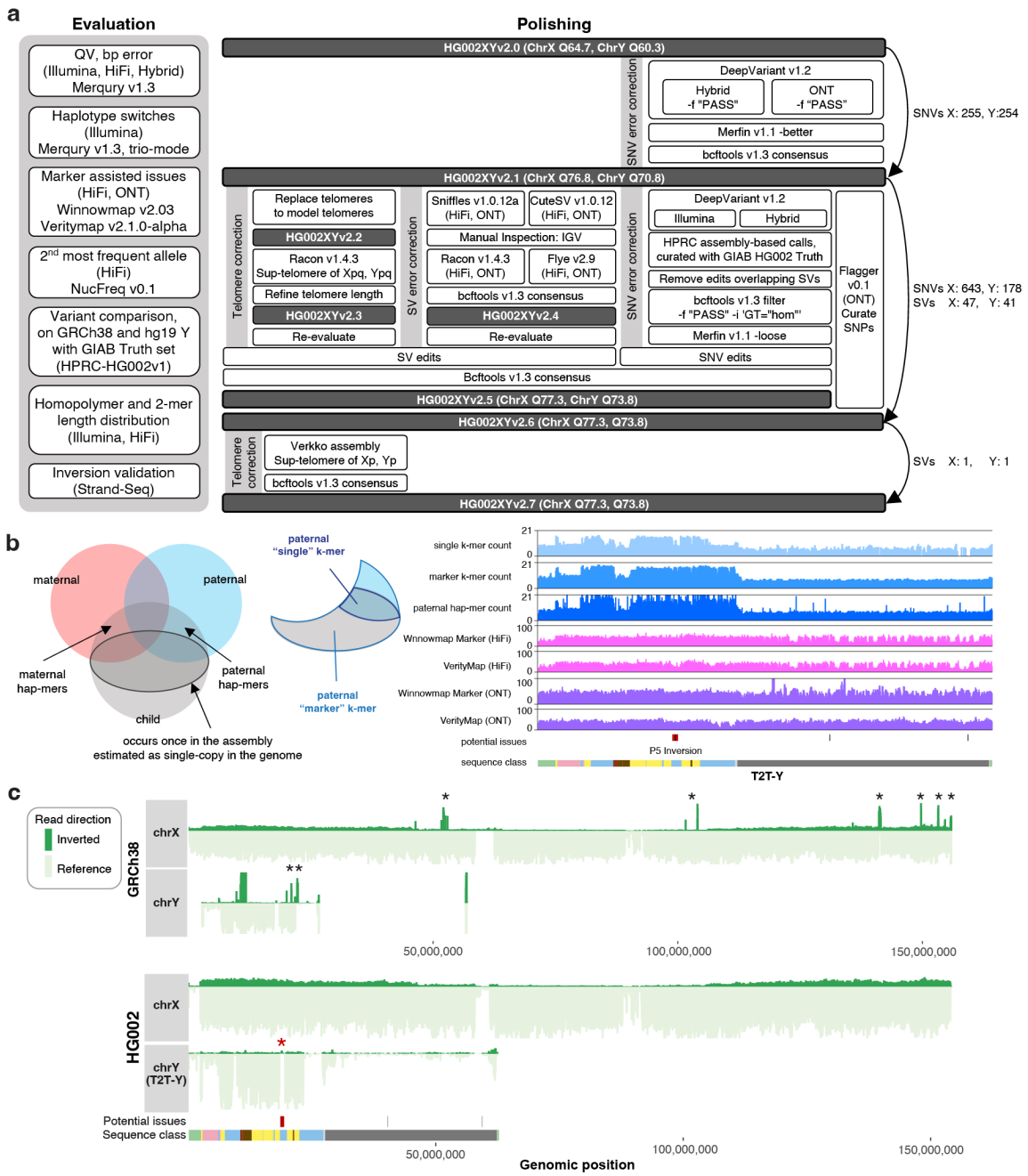
S.N. is an employee of Oxford Nanopore Technologies; A.F. is an employee of DNAnexus; C.-S.C. is an employee of Sema4 OpCo Inc.; N.-C.C. is an employee of Exai Bio; L.F.P. receives research support from Genetech; F.J.S. receives research support from Pacific Biosciences, Oxford Nanopore Technologies, 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.

Extended Data Figures



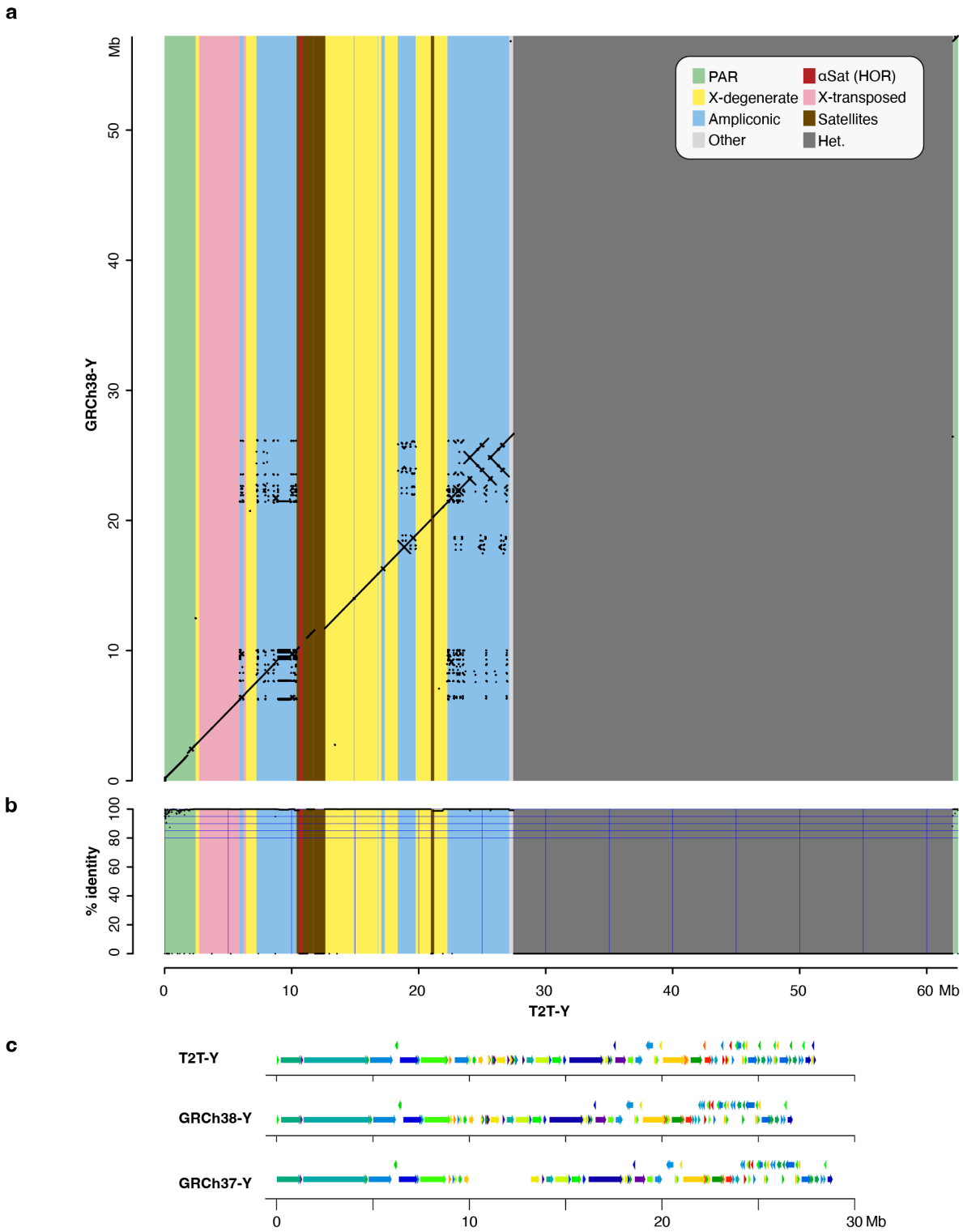
Extended Data Fig. 1 | Assembling the X and Y chromosomes of HG002. a. Chromosome X and Y components of the assembly string graph built from HiFi reads, detected based on node sequence alignments to T2T-CHM13 and GRCh38 references. Each node is colored according to the excess of paternal-specific (blue) 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 (Yqh). 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. 1c**). 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. 1c**. 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. Gaps (green edges) and uneven node coverage estimates indicate biases in HiFi sequencing across the region. **Fig. 1a** shows an enrichment of SINE repeats and non-B DNA motifs in PAR1 that may underlie the sequencing gaps in this region.

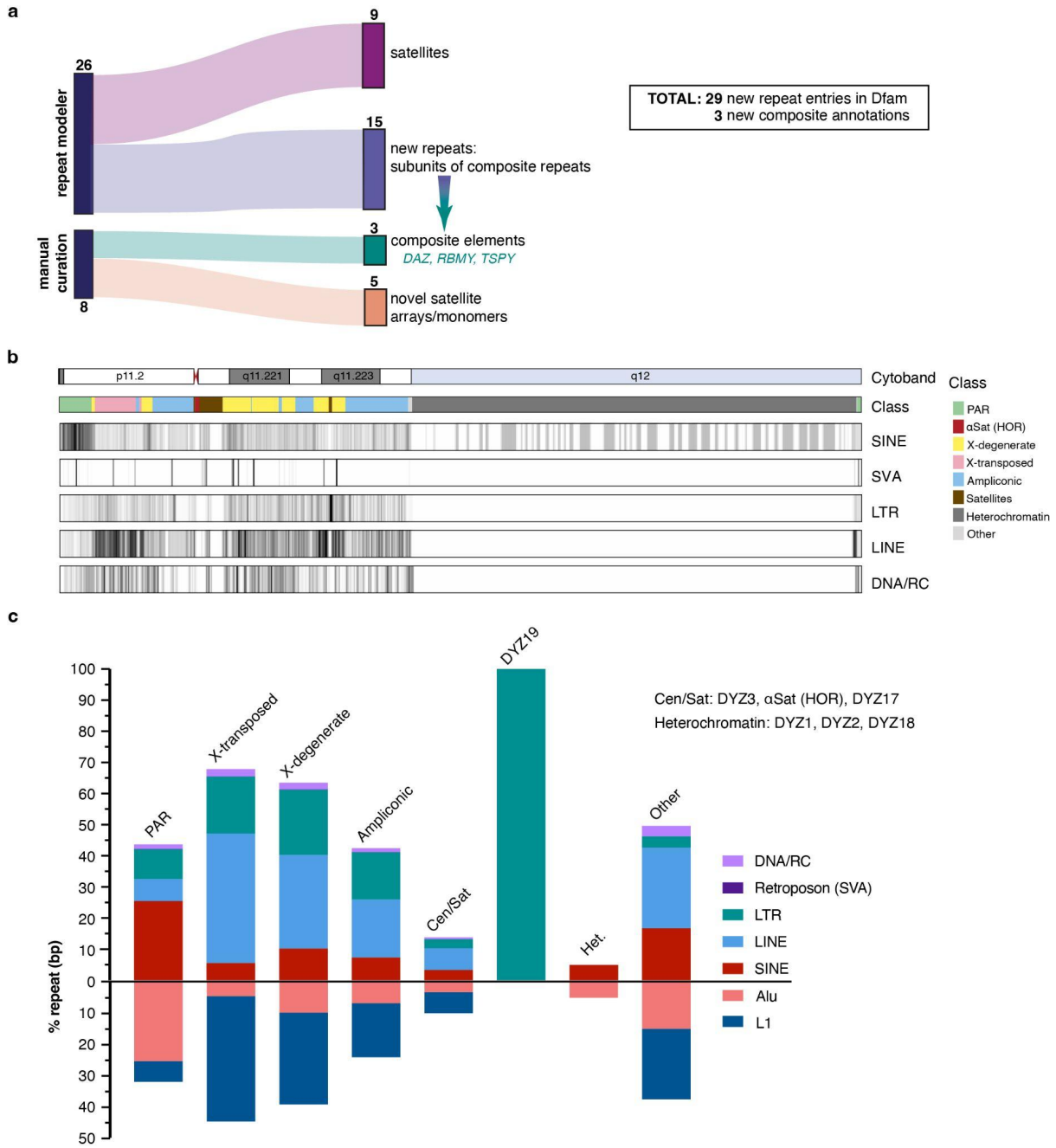


Extended Data Fig. 2 | Validation of the T2T-Y. a. Evaluation and polishing workflow performed on T2T-CHM13v1.1 autosomes + HG002 XY assemblies. **b**. Venn diagram of the k-mers from the parents and child. On the left, hap-mers[15] represent haplotype specific k-mers inherited by the child. The darker 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[14], 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 Winnowmap 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

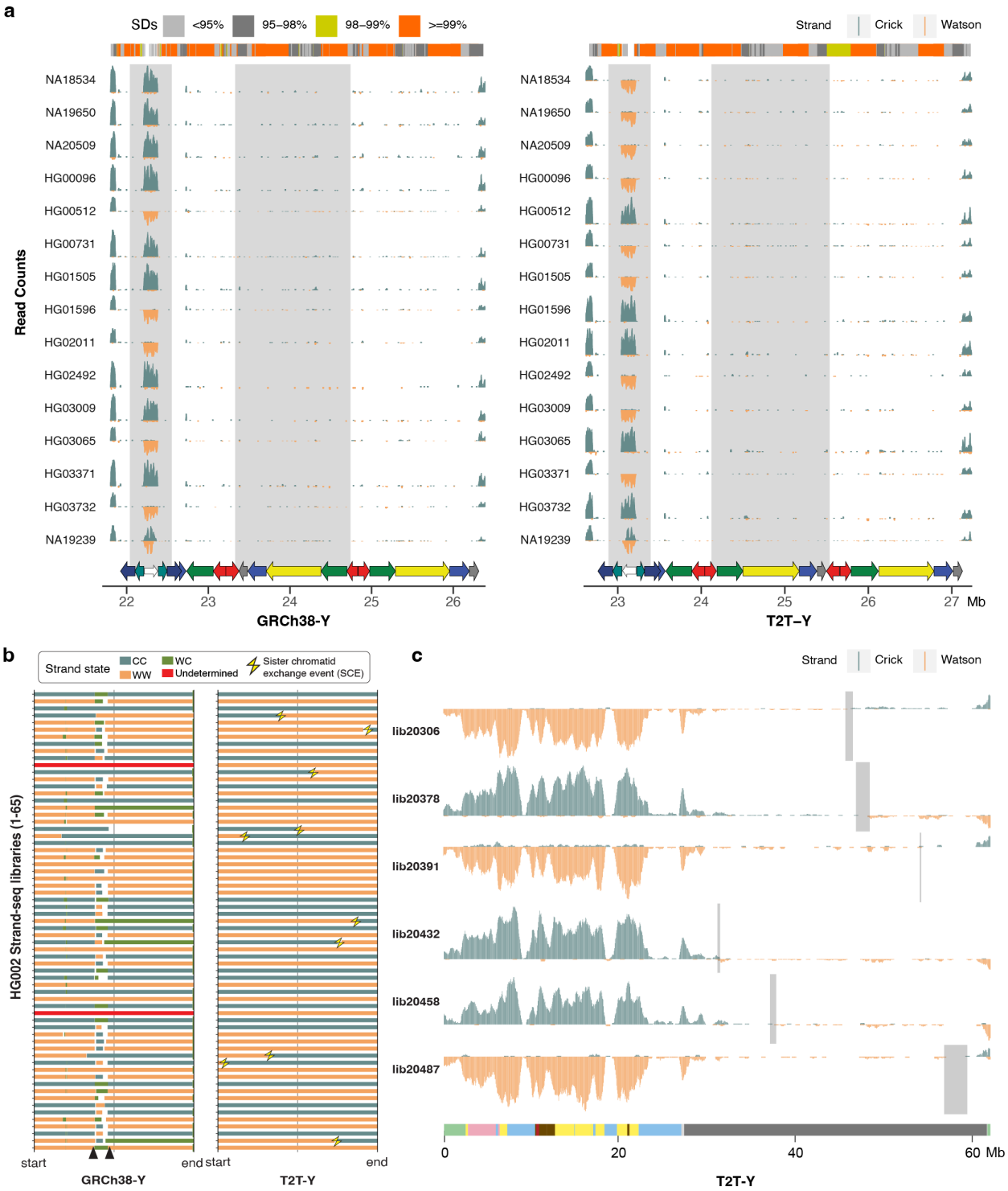
in forward direction (light green) or reverse direction (dark green). Multiple spikes in reverse direction (black asterisks) in GRCh38-Y indicate inversion polymorphisms relative to HG002, likely due to differences between the haplogroups. Such spikes in coverage are not observed on T2T-X and T2T-Y, 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[95] after softmasking with WindowMasker[96]. **b.** Identity was computed from matches and mismatches over positions with alignments, excluding gaps. **c.** Structural differences revealed using PRG-TK[97] against GRCh38-Y and GRCh37-Y in the euchromatic region of the Y chromosome.

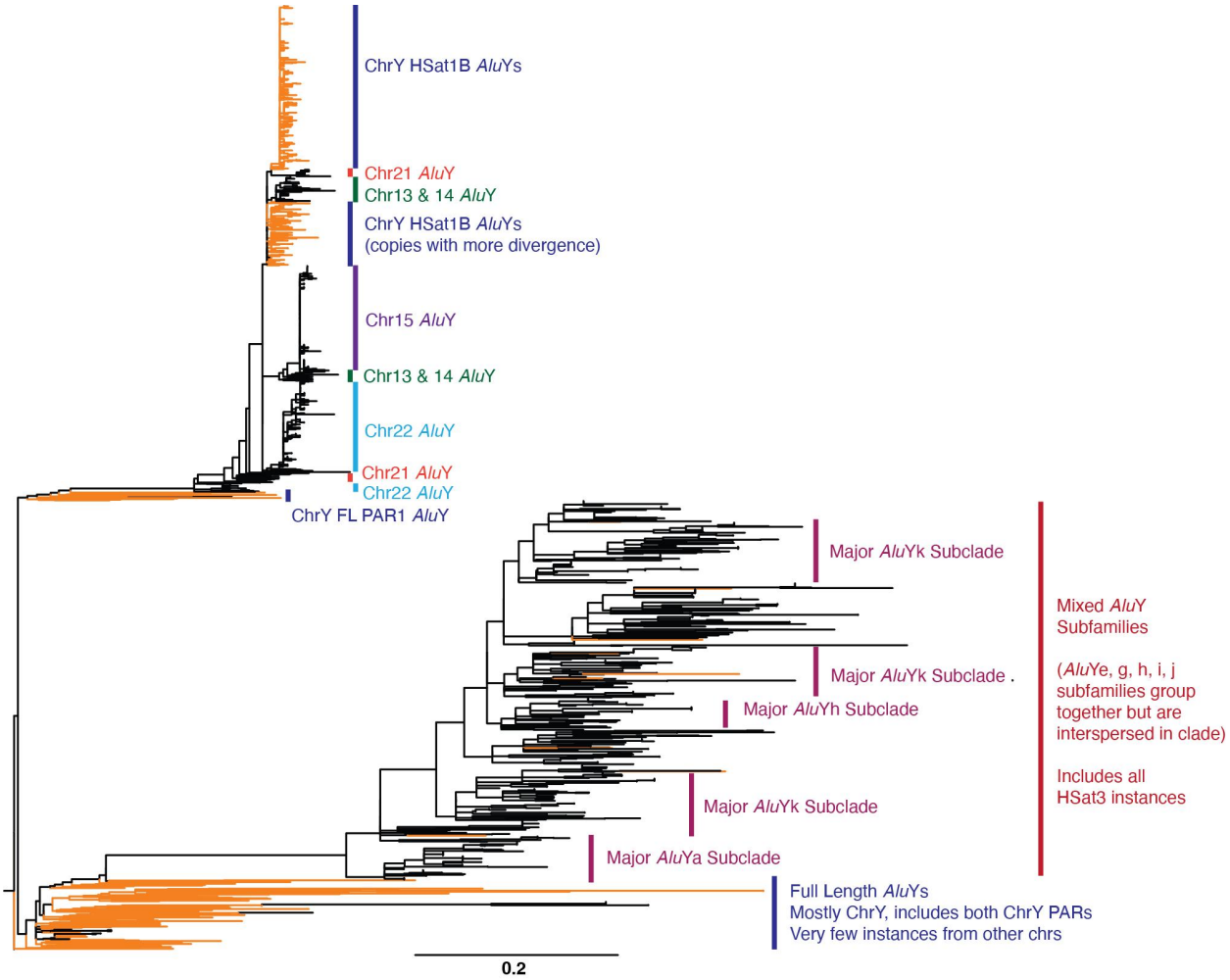


Extended Data Fig. 4 | Repeat discovery and annotation of T2T-Y. a. Assembly completion allowed for a full assessment of repeats and resulted in the identification of previously unknown satellite arrays (predominantly in the PAR1) and subunit repeats that fall within one of three composite repeat units (*TSPY*, *RBMY*, *DAZ*). **b.** Ideogram of TE density (per 100 kb bin). This is an extension of **Fig**. **1a** 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.

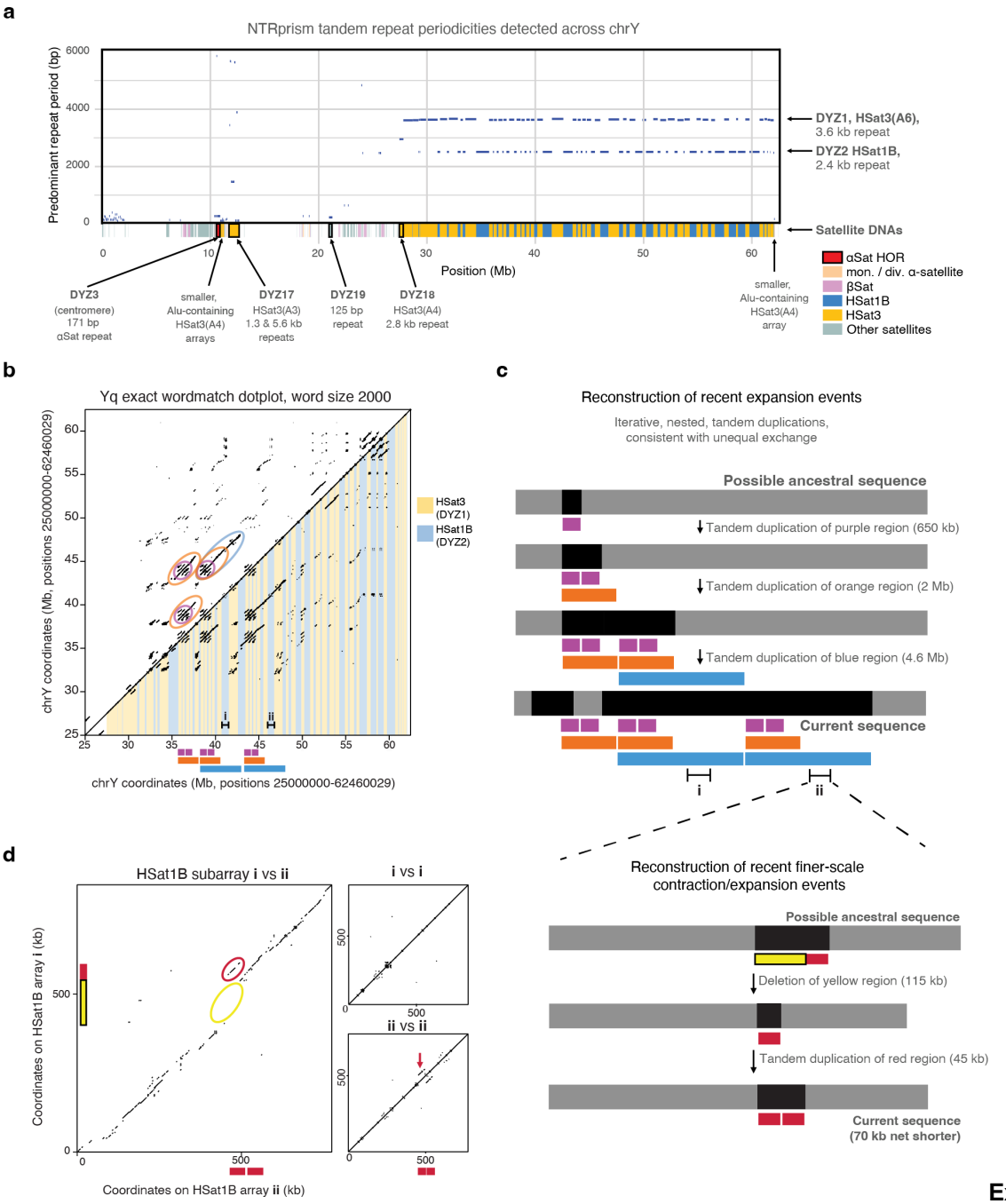


Extended Data Fig. 5 | Recurrent inversions identified with Strand-seq. a. Five out of 15 individuals have the inverted variant as present in HG002 at the P3 palindrome. Although the inversion across P1–P2 is difficult to confirm with Strand-seq because of the high sequence similarity between the palindromic arms, it is noticeable that different rearrangements are present in these individuals. **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

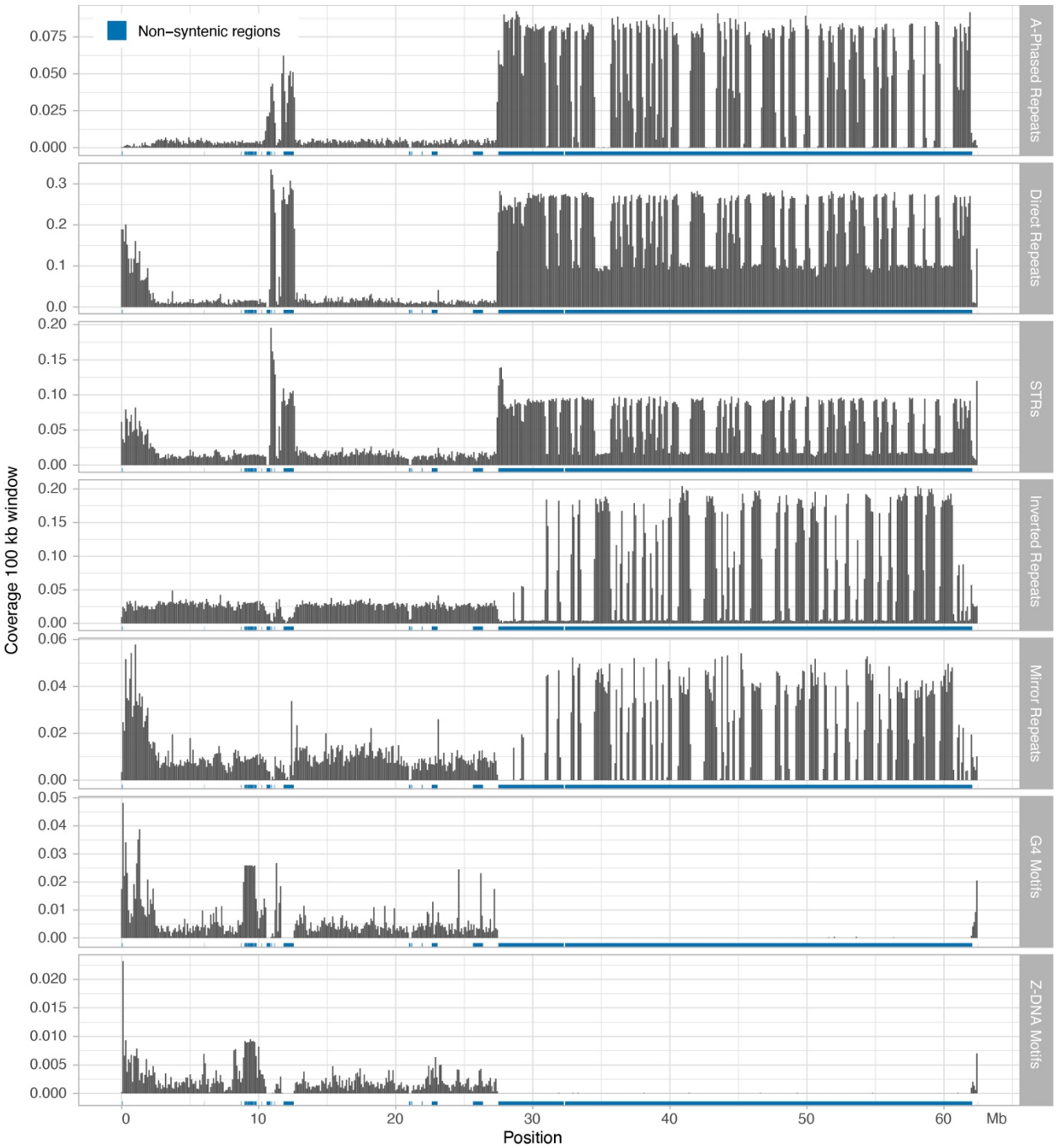
in plus orientation. Changes in strand state along a single chromosome are normally caused by a double-strand-break (DSBs) that occurred during DNA replication[98] in a random fashion and we refer to them as sister-chromatid-exchanges (SCEs, yellow thunderbolts). Recurrent change in strand state over the same region in multiple Strand-seq cells indicates misassemblies. 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 Yqh 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**.



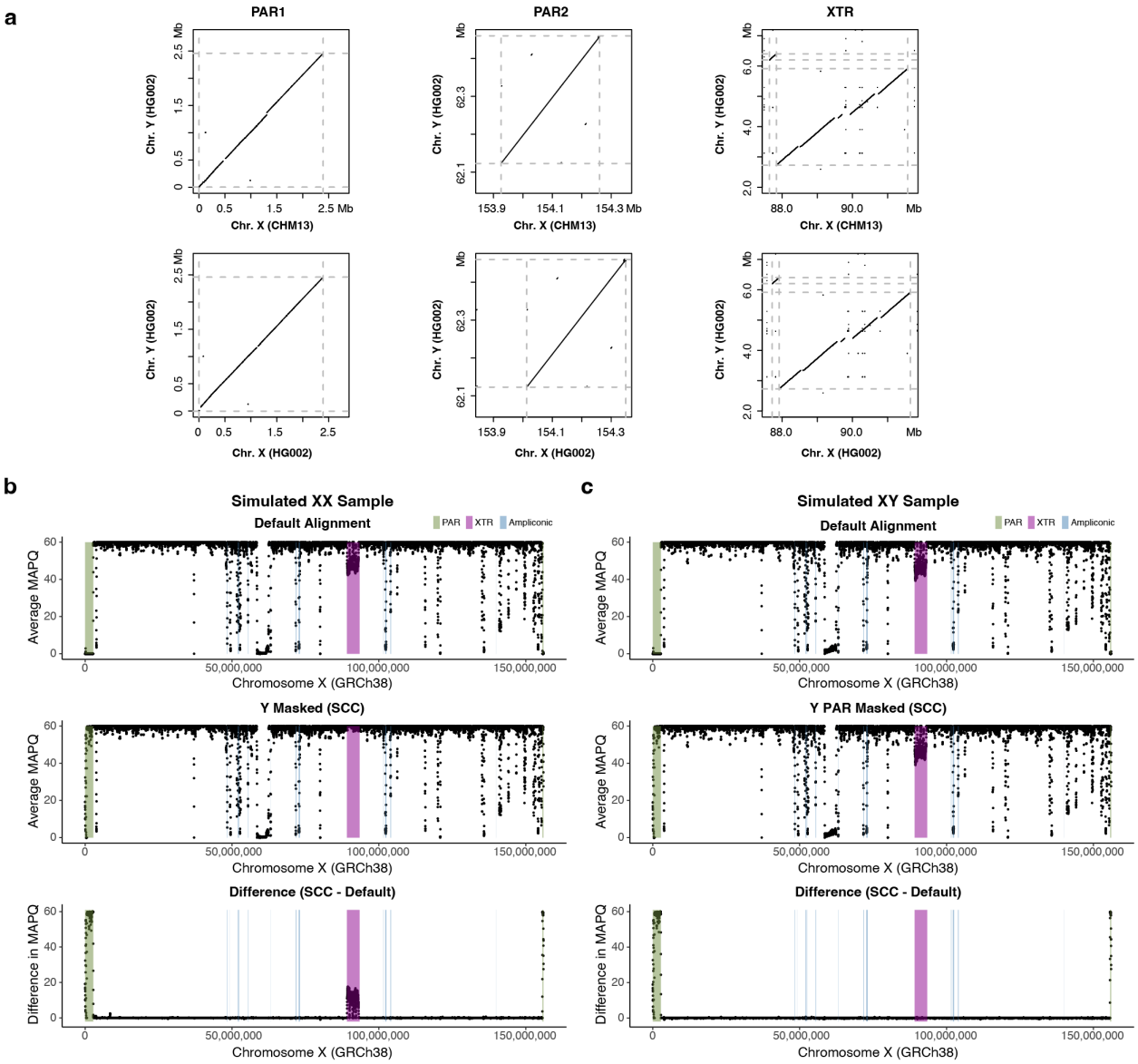
Extended Data Fig. 6 | Phylogenetic tree of *Alu*Y sequences associated with HSat1B and HSat3. Tree represents subsampling of *Alu*Y elements, both full length (FL) and truncated, including *Alu*Y sequences found within HSat1B units and associated with HSat3 arrays. The tree is rooted on the RepeatMasker/Dfam derived consensus sequence for *Alu*Sc8. Elements located on ChrY are denoted with orange branches. Analysis was run on a MAFFT[99] derived alignment using RAxML-NG[100] with 100 non-parametric bootstrap replications. Each major grouping of *Alu*Y subfamily, chromosomal location and/or HSat1B and HSat3 derivation is labeled accordingly. Note that in the *Alu*Y subfamily clade ("Mixed *Alu*Y Subfamilies") there are scattered elements across the group even though the majority are represented in the labeled subclades. The scale bar represents 0.2 substitutions per site on a branch of the same length.



Extended Data Fig. 7 | Satellite annotation and recent expansion events on the Yq heterochromatin. **a.** A plot showing the top repeat periodicities detected by NTRprism[48] 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[2], HSat subfamilies[66], 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. 8 | 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, *TSPY* gene array showed enriched G4 and Z-DNA motifs, as shown in **Fig. 2d**.



Extended Data Fig. 9 | Genomic similarity in PARs and XTR and improved MAPQ of the PARs through informed sex chromosome complement reference. a. Dot plots from LASTZ alignment 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.

References

1. 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).
2. 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).
3. Vog, P. H. *et al.* Human Y Chromosome Azoospermia Factors (AZF) Mapped to Different Subregions in Yq11. *Hum. Mol. Genet.* **5**, 933–943 (1996).
4. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
5. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
6. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
7. 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).
8. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
9. 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.
10. Rautiainen, M. & Marschall, T. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* **21**, 253 (2020).
11. Formenti, G. *et al.* Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. *Nat. Methods* **19**, 696–704 (2022).
12. Kirsche, M. *et al.* Jasmine: Population-scale structural variant comparison and analysis. *bioRxiv* (2021) doi:10.1101/2021.05.27.445886.

13. 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).

14. Mc Cartney, A. M. *et al.* Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. *Nat. Methods* **19**, 687–695 (2022).

15. 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).

16. Wang, T. *et al.* The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).

17. Jarvis, E. D. *et al.* Semi-automated assembly of high-quality diploid human reference genomes. *Nature* **611**, 519–531 (2022).

18. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).

19. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).

20. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).

21. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).

22. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

23. 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).

24. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

25. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

26. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).

27. 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).

28. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).

29. Jiang, T. *et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).

30. 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.

31. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

32. Sanders, A. D. *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).

33. Hallast et al. Assembly of 43 diverse human Y chromosomes reveals extensive complexity and variation. *bioRxiv* (2022).

34. David Poznik. yhaplo | Identifying Y-Chromosome Haplogroups. Last accessed: 2022-11-29. https://github.com/23andMe/yhaplo (2022).

35. Hammer, M. F. *et al.* Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Hum. Genet.* **126**, 707 (2009).

36. 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).

37. Vollger, M. R. SafFire. Last accessed: 2022-11-29. https://github.com/mrvollger/SafFire (2022).

38. Jiang, Z., Hubley, R., Smit, A. & Eichler, E. E. DupMasker: a tool for annotating primate segmental duplications. *Genome Res.* **18**, 1362–1368 (2008).

39. 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).

40. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.* **36**, 321–323 (2018).

41. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

42. Hoyt, S. J. *et al.* From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).

43. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics* **9**, 533 (2008).

44. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).

45. Subrini, J. & Turner, J. Y chromosome functions in mammalian spermatogenesis. *eLife* **10**, e67345 (2021).

46. 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).

47. NCBI RefSeq v110 Browser. Homo sapiens isolate NA24385 chromosome Y, alternate assembly T2T-CHM13v2.0. Last accessed: 2022-12-01. (2022).

48. Altemose, N. *et al.* Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).

49. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

50. 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).

51. Jobling, M. A. Copy number variation on the human Y chromosome. *Cytogenet. Genome Res.* **123**, 253–262 (2008).

52. Nailwal, M. & Chauhan, J. B. Azoospermia factor C subregion of the Y chromosome. *J. Hum. Reprod. Sci.* **10**, 256 (2017).

53. 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).

54. 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).

55. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986-2005.e26 (2022).

56. 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).

57. 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).

58. 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).

59. 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).

60. Altemose, N. A classical revival: Human satellite DNAs enter the genomics era. *Semin. Cell Dev. Biol.* **128**, 2–14 (2022).

61. Cooke, H. Repeated sequence specific to human males. *Nature* **262**, 182–186 (1976).

62. 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).

63. 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).

64. Gripenberg, U. Size variation and orientation of the human Y chromosome. *Chromosoma* **15**, 618–629 (1964).

65. Mathias, N., Bayés, M. & Tyler-Smith, C. Highly informative compound haplotypes for the human Y chromosome. *Hum. Mol. Genet.* **3**, 115–123 (1994).

66. 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).

67. 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).

68. Sahakyan, A. B. *et al.* Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci. Rep.* **7**, 14535 (2017).

69. Goodier, J. L., Ostertag, E. M. & Kazazian Jr, H. H. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).

70. Xing, J. *et al.* Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc. Natl. Acad. Sci.* **103**, 17608–17613 (2006).

71. Pickeral, O. K., Makałowski, W., Boguski, M. S. & Boeke, J. D. Frequent Human Genomic DNA Transduction Driven by LINE-1 Retrotransposition. *Genome Res.* **10**, 411–415 (2000).

72. Damert, A. *et al.* 5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res.* **19**, 1992–2008 (2009).

73. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Exon Shuffling by L1 Retrotransposition. *Science* **283**, 1530–1534 (1999).

74. Tubio, J. M. C. *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014).

75. Halabian, R. & Makałowski, W. A Map of 3′ DNA Transduction Variants Mediated by Non-LTR Retroelements on 3202 Human Genomes. *Biology* **11**, 1032 (2022).

76. Ghosh, A. & Bansal, M. A glossary of DNA structures from A to Z. *Acta Crystallogr. D Biol. Crystallogr.* **59**, 620–626 (2003).

77. Jain, A., Wang, G. & Vasquez, K. M. DNA triple helices: Biological consequences and therapeutic potential. *Biochimie* **90**, 1117–1130 (2008).

78. Wang, G. & Vasquez, K. M. Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. *DNA Repair* **19**, 143–151 (2014).

79. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat. Rev. Mol. Cell Biol.* **21**, 459–474 (2020).

80. Kasinathan, S. & Henikoff, S. Non-B-Form DNA Is Enriched at Centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).

81. Cotter, D. J., Brotman, S. M. & Wilson Sayres, M. A. Genetic Diversity on the Human X Chromosome Does Not Support a Strict Pseudoautosomal Boundary. *Genetics* **203**, 485–492 (2016).

82. Lien, S., Szyda, J., Schechinger, B., Rappold, G. & Arnheim, N. Evidence for Heterogeneity in Recombination in the Human Pseudoautosomal Region: High Resolution Analysis by Sperm Typing and Radiation-Hybrid Mapping. *Am. J. Hum. Genet.* **66**, 557–566 (2000).

83. Genome Reference Consortium. Assembly Terminology. Last accessed: 2022-11-29. https://www.ncbi.nlm.nih.gov/grc/help/definitions/.

84. Webster, T. H. *et al.* Identifying, understanding, and correcting technical artifacts on the sex chromosomes in next-generation sequencing data. *GigaScience* **8**, giz074 (2019).

85. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

86. Krumsiek, J., Arnold, R. & Rattei, T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026–1028 (2007).

87. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

88. 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).

89. 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).

90. Chrisman, B. *et al.* The human “contaminome”: bacterial, viral, and computational contamination in whole genome sequences from 1000 families. *Sci. Rep.* **12**, 9863 (2022).

91. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

92. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

93. Rautiainen, M. *et al.* Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv* 2022.06.24.497523 (2022) doi:10.1101/2022.06.24.497523.

94. Liao, W.-W. *et al.* A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321 (2022) doi:10.1101/2022.07.09.499321.

95. Harris, Robert S. Improved Pairwise Alignmnet of Genomic DNA. (Penn State, 2007).

96. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).

97. Chin, C.-S. *et al.* Multiscale Analysis of Pangenome Enables Improved Representation of Genomic Diversity For Repetitive And Clinically Relevant Genes. *bioRxiv* 2022.08.05.502980 (2022) doi:10.1101/2022.08.05.502980.

98. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).

99. 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.

100. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).