Article

Assembly of 43 human Y chromosomes reveals extensive complexity and variation

https://doi.org/10.1038/s41586-023-06425-6

Received: 30 November 2022

Accepted: 11 July 2023

Published online: 23 August 2023

Check for updates

Pille Hallast[1,18], Peter Ebert[2,3,4,18], Mark Loftus[5,6,18], Feyza Yilmaz[1], Peter A. Audano[1], Glennis A. Logsdon[7], Marc Jan Bonder[8,9], Weichen Zhou[10], Wolfram Höps[11], Kwondo Kim[1], Chong Li[12], Savannah J. Hoyt[13], Philip C. Dishuck[7], David Porubsky[7], Fotios Tsetsos[1], Jee Young Kwon[1], Qihui Zhu[1], Katherine M. Munson[7], Patrick Hasenfeld[11], William T. Harvey[7], Alexandra P. Lewis[7], Jennifer Kordosky[7], Kendra Hoekzema[7], Human Genome Structural Variation Consortium (HGSVC)*,**, Rachel J. O'Neill[13,14,15], Jan O. Korbel[11], Chris Tyler-Smith[16], Evan E. Eichler[7,17], Xinghua Shi[12], Christine R. Beck[1,14,15], Tobias Marschall[2,4], Miriam K. Konkel[5,6,19] & Charles Lee[1,19] ✉

The prevalence of highly repetitive sequences within the human Y chromosome has prevented its complete assembly to date[1] and led to its systematic omission from genomic analyses. Here we present de novo assemblies of 43 Y chromosomes spanning 182,900 years of human evolution and report considerable diversity in size and structure. Half of the male-specific euchromatic region is subject to large inversions with a greater than twofold higher recurrence rate compared with all other chromosomes[2]. Ampliconic sequences associated with these inversions show differing mutation rates that are sequence context dependent, and some ampliconic genes exhibit evidence for concerted evolution with the acquisition and purging of lineage-specific pseudogenes. The largest heterochromatic region in the human genome, Yq12, is composed of alternating repeat arrays that show extensive variation in the number, size and distribution, but retain a 1:1 copy-number ratio. Finally, our data suggest that the boundary between the recombining pseudoautosomal region 1 and the non-recombining portions of the X and Y chromosomes lies 500 kb away from the currently established[1] boundary. The availability of fully sequence-resolved Y chromosomes from multiple individuals provides a unique opportunity for identifying new associations of traits with specific Y-chromosomal variants and garnering insights into the evolution and function of complex regions of the human genome.

The mammalian sex chromosomes evolved from a pair of autosomes, gradually losing their ability to recombine with each other over increasing lengths of the chromosomes, leading to degradation and accumulation of large proportions of repetitive sequences on the Y chromosome[3]. The resulting sequence composition of the human Y chromosome is rich in complex repetitive regions, including highly similar segmental duplications (SDs)[1,4]. This has made the Y chromosome difficult to assemble and, paired with reduced gene content, has led to its systematic neglect in genomic analyses.

The first human Y chromosome sequence assembly was generated almost 20 years ago, providing a high-quality but incomplete sequence (53.8%, or around 30.8 Mb out of 57.2 Mb unresolved in GRCh38 Y)[1]. Less than half (around 25 Mb) of the GRCh38 Y chromosome is composed of euchromatin, which contains two pseudoautosomal regions (PARs), PAR1 and PAR2 (around 3.2 Mb in total), that actively recombine with homologous regions on the X chromosome and are therefore not considered to be part of the male-specific Y region (MSY)[1]. The remainder of the Y-chromosomal euchromatin (around 22 Mb) has been divided into three main classes according to their sequence composition and evolutionary history[1]: (1) the X-degenerate regions (XDR, around 8.6 Mb) are remnants of the ancient autosome from which the X and Y chromosomes evolved; (2) the X-transposed regions (XTR, around 3.4 Mb) resulted from a duplicative transposition event from the X chromosome followed by an inversion; and (3) the ampliconic regions (around 9.9 Mb), which contain sequences with up to 99.9% intrachromosomal identity across tens or hundreds of

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA. [2]Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany. [3]Core Unit Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany. [4]Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany. [5]Department of Genetics & Biochemistry, Clemson University, Clemson, SC, USA. [6]Center for Human Genetics, Clemson University, Greenwood, SC, USA. [7]Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA. [8]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [9]Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [10]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA. [11]Genome Biology Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany. [12]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [13]Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, USA. [14]Institute for Systems Genomics, University of Connecticut, Storrs, CT, USA. [15]The University of Connecticut Health Center, Farmington, CT, USA. [16]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [17]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. [18]These authors contributed equally: Pille Hallast, Peter Ebert, Mark Loftus. [19]These authors jointly supervised this work: Miriam K. Konkel, Charles Lee. *A list of authors and their affiliations appear at the end of the paper. **A full list of members and their affiliations appears in the Supplementary Information. ✉e-mail: charles.lee@jax.org

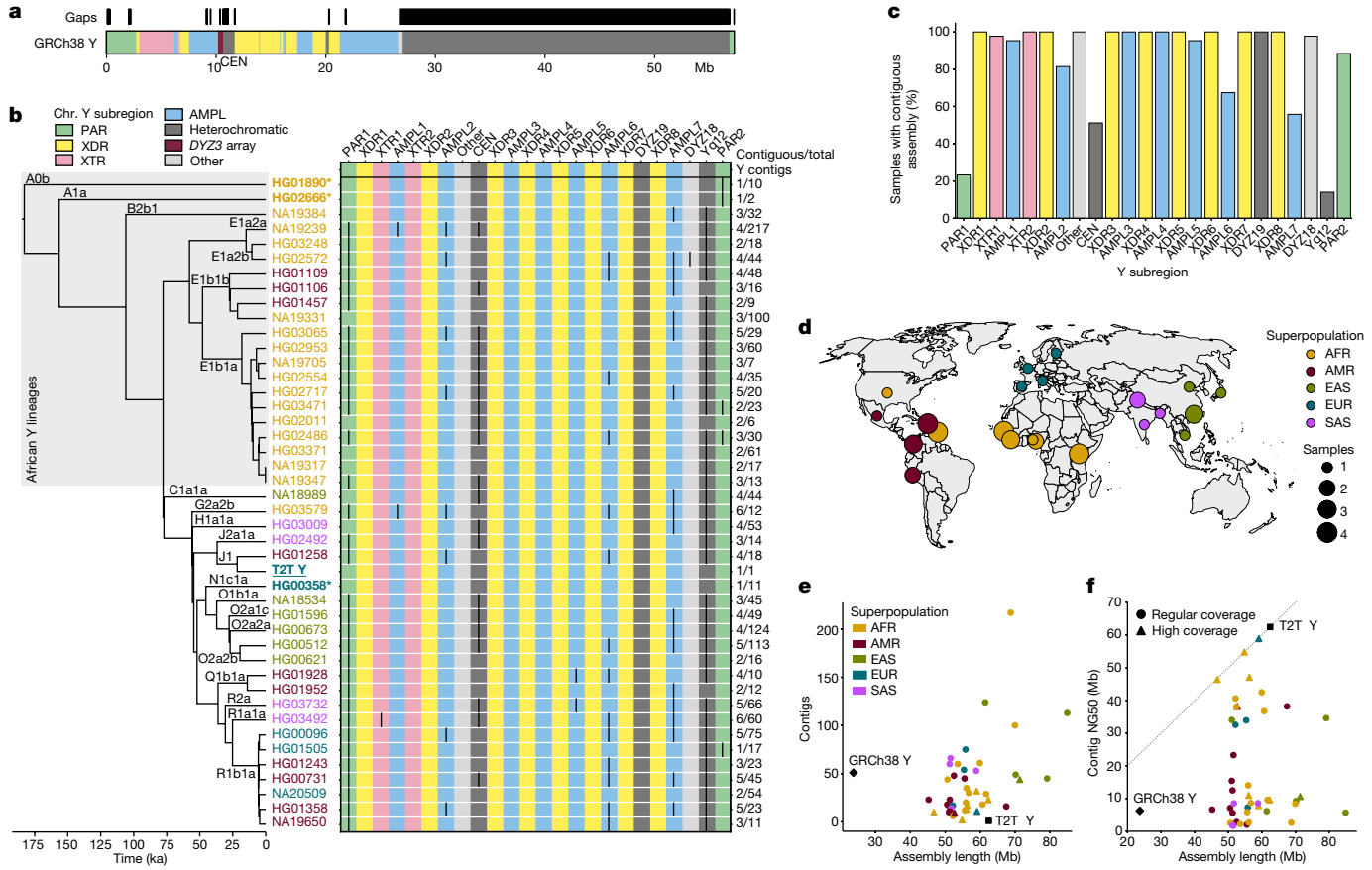


Fig. 1 | De novo assembly outcome. a, The structure of the human Y chromosome on the basis of the GRCh38 Y reference sequence. CEN, centromere. **b**, Phylogenetic relationships (left) with haplogroup labels of the analysed Y chromosomes, with branch lengths drawn proportional to the estimated times between successive splits (further details are provided in Supplementary Fig. 1 and Supplementary Table 1). Summary of Y-chromosome assembly completeness (right). The vertical black lines represent non-contiguous assembly of that region (Methods). The numbers on the right indicate the number of Y contigs needed to achieve the indicated contiguity/total number of assembled Y contigs for each sample. CEN includes the *DYZ3* α-satellite array and the pericentromeric region. Three contiguously assembled Y chromosomes are indicated by asterisks (assemblies for HG02666 and HG00358 are contiguous from telomere to telomere, whereas the HG01890 assembly has a break approximately 100 kb before the end of PAR2) and the T2T Y assembly is underlined. The colour of sample ID corresponds to the superpopulation designation (as described in **d**). Note that the GRCh38 Y sequence mostly represents Y haplogroup R1b. AMPL, ampliconic; ka, thousand years ago. **c**, The proportion of contiguously assembled Y-chromosomal subregions across 43 samples. **d**, The geographical origin and sample size of the included 1000 Genomes Project samples coloured according to the continental groups (African (AFR), American (AMR), European (EUR), South Asian (SAS) and East Asian (EAS)). **e**, Y-chromosomal assembly length versus the number of Y contigs. Gap sequences (IUPAC code N) were excluded from GRCh38 Y. **f**, Y-chromosomal assembly length versus Y contig NG50. High coverage was defined as greater than 50× genome-wide PacBio HiFi read depth. Gap sequences (IUPAC code N) were excluded from GRCh38 Y.

kilobases (Fig. 1a). Besides the euchromatin, the Y chromosome contains a large proportion of repetitive and heterochromatic sequences, including the (peri)centromeric *DYZ3* α-satellite and *DYZ17* arrays, *DYZ18* and *DYZ19* arrays, and the largest contiguous heterochromatic block in the human genome, Yq12, which is known to be highly variable in size[1,5,6]. All of these heterochromatic regions are thought to be composed predominantly of satellites, simple repeats and SDs[1,7].

Recent attempts have been made to assemble the human Y chromosome using Illumina short-read[8] and Oxford Nanopore Technologies (ONT) long-read data[9], but a contiguous assembly of the ampliconic and heterochromatic regions was not achieved. In April 2022, the de novo assembly of a human Y chromosome was reported by the Telomere-to-Telomere (T2T) Consortium[10] (from individual HG002/NA24385, carrying a rare J1a-L816 Y lineage found among Ashkenazi Jews and Europeans[11], termed T2T Y). However, understanding the composition and appreciating the complexity of the Y chromosomes in the human population requires availability of assemblies from many diverse individuals. Here, we combined PacBio HiFi and ONT long-read sequencing data to assemble the Y chromosomes from 43 male individuals, representing the five continental groups from the 1000 Genomes Project. Whereas both the GRCh38 (mostly R1b-L20 haplogroup) and the T2T Y assemblies represent European Y lineages, half of our Y chromosomes constitute African lineages and include most of the deepest-rooted human Y lineages. This newly assembled dataset of 43 Y chromosomes therefore provides a more comprehensive view of genetic variation, at the nucleotide level, across over 180,000 years of human Y chromosome evolution.

Sample selection

We selected 43 genetically diverse male individuals from the 1000 Genomes Project, representing 21 largely African haplogroups (A, B and E, including deep-rooted lineages A0b-L1038, A1a-M31 and B2b-M112)[12,13] (Fig. 1b,d, Methods, Supplementary Fig. 1 and Supplementary Table 1). The time to the most recent common ancestor (TMRCA) among our 43 Y chromosomes and the T2T Y was estimated to be approximately 182,900 years ago (95% highest posterior density (HPD) interval = 159,800–209,200 years ago) (Supplementary Fig. 1), consistent with previous reports[14,15]. A pair of closely related African Y chromosomes (NA19317 and NA19347, lineage E1b1a1a1a-CTS8030), was included for assembly

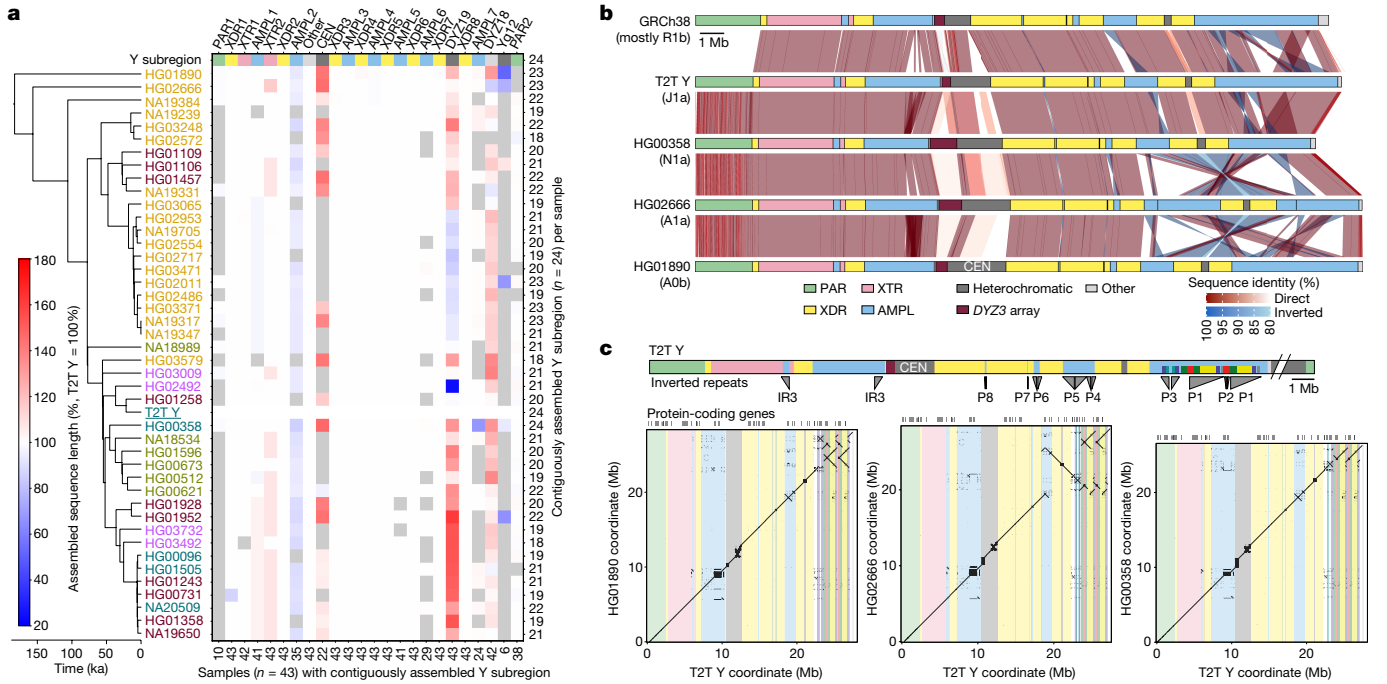


Fig. 2 | Size and structural variation in Y chromosomes. a, The size variation in contiguously assembled Y-chromosomal subregions shown as a heat map relative to the T2T Y size (as 100%). The grey boxes indicate regions that are not contiguously assembled (Methods). The numbers on the bottom indicate contiguously assembled samples for each subregion out of a total of 43 samples, and the numbers on the right indicate the contiguously assembled Y subregions out of 24 regions for each sample. Samples are coloured as in Fig. 1b. AMPL, ampliconic; CEN, centromere; ka, thousand years ago. **b**, Comparison of the three contiguously assembled Y chromosomes to GRCh38 and the T2T Y (excluding the Yq12 and PAR2 subregions). **c**, Dot plots of three contiguously assembled Y chromosomes versus the T2T Y (excluding the Yq12 and PAR2 subregions), annotated with Y subregions and SDs in ampliconic subregion 7 (Supplementary Fig. 34).

validation, as these Y chromosomes are expected to be highly similar (TMRCA 200 years ago (95% HPD interval = 0–500 years ago)).

Constructing de novo assemblies

We used the hybrid assembler Verkko[16] to generate Y-chromosome assemblies, including the ampliconic and heterochromatic regions (Methods). Verkko leverages the high accuracy of PacBio HiFi reads (>99.8% base pair calling accuracy[17,18]) with the length of ONT long/ultralong reads (median read length N50 = 134 kb) to produce highly accurate and contiguous assemblies (Supplementary Table 2). Using this approach, we generated high-quality (median QV = 48; Supplementary Table 3) whole-genome (median length = 5.9 Gb; Supplementary Table S4) assemblies for the 43 males studied. The chromosome Y sequences exhibit a high degree of completeness (median length = 55.6 Mb, 79% to 148% assembly length relative to GRCh38 Y; Fig. 1, Supplementary Fig. 2 and Supplementary Tables 5 and 6), contiguity (median NG50 = 9.6 Mb, median LG50 = 2; Supplementary Table 4), base-pair quality (median QV = 46; Supplementary Table 3) and read-depth profile consistency with the autosomal sequences in the assemblies (Supplementary Fig. 3 and Supplementary Table 7). The Verkko assembly process was robust (sequence identity for NA19317/NA19347 pair of 99.9959%; Supplementary Fig. 4, Supplementary Table 8 and Supplementary Results (De novo assembly evaluation)). We generated a gapless Y-chromosome assembly, spanning from PAR1 to PAR2, for three individuals, two of whom represent deep-rooted African haplogroups (Figs. 1b and 2 and Supplementary Table 9). These three samples are among nine samples with an increased HiFi coverage of at least 50× (high-coverage samples; Supplementary Tables 1, 2 and 7).

Following established procedures[19–21], we flagged potentially erroneous regions, comprising 0.103% (median; mean = 0.31%) up to 0.186% (median; mean = 0.467%) of the assembled Y sequence (Methods, Supplementary Fig. 5 and Supplementary Tables 10–12). Although the error rate is increased for the lower-coverage assemblies, increasing the HiFi coverage beyond 50× has limited effect on the error rate (Supplementary Fig. 6).

We further annotated each of the Y-chromosomal assemblies with respect to the 24 Y-chromosomal subregions originally proposed[1] (Fig. 1a–c, Methods, Supplementary Fig. 2 and Supplementary Table 13). In addition to the three gapless Y chromosomes, we contiguously assembled the MSY—excluding Yq12 and the (peri)centromeric region—for 17 out of 43 samples (Supplementary Tables 9 and 14–16). Overall, 17 out of 24 subregions were contiguously assembled across 41 out of 43 samples (Fig. 1b,c and Supplementary Fig. 2).

Diversity of assembled Y chromosomes

Size variation of the assembled Y chromosomes

The assembled Y chromosomes showed extensive variation both in size and structure (Figs. 2a–c and 3–5, Methods, Extended Data Fig. 1a and Supplementary Figs. 7–18) with chromosome sizes ranging from 45.2 to 84.9 Mb (mean = 57.6 Mb, median = 55.7 Mb; Methods, Supplementary Fig. 16 and Supplementary Tables 14 and 16). However, this is a slight underestimate of the true Y-chromosomal size due to assembly gaps. An analysis of the underlying assembly graphs suggest that the paths of complete assemblies would be, on average, 1.15% longer (Supplementary Table 6 and Supplementary Results (De novo assembly evaluation)). Among the gaplessly assembled Y-chromosomal subregions (including for the T2T Y), the largest variation in subregion size was seen for the heterochromatic Yq12 (17.6 to 37.2 Mb, mean = 27.6 Mb), the (peri)centromeric region (2.0 to 3.1 Mb, mean = 2.6 Mb) and the *DYZ19*-repeat array (63.5 kb to 428 kb; mean = 307 kb) (Figs. 2a and 5f, Extended Data Fig. 1b, Supplementary Figs. 7 and 16–21 and Supplementary Tables 14–16).

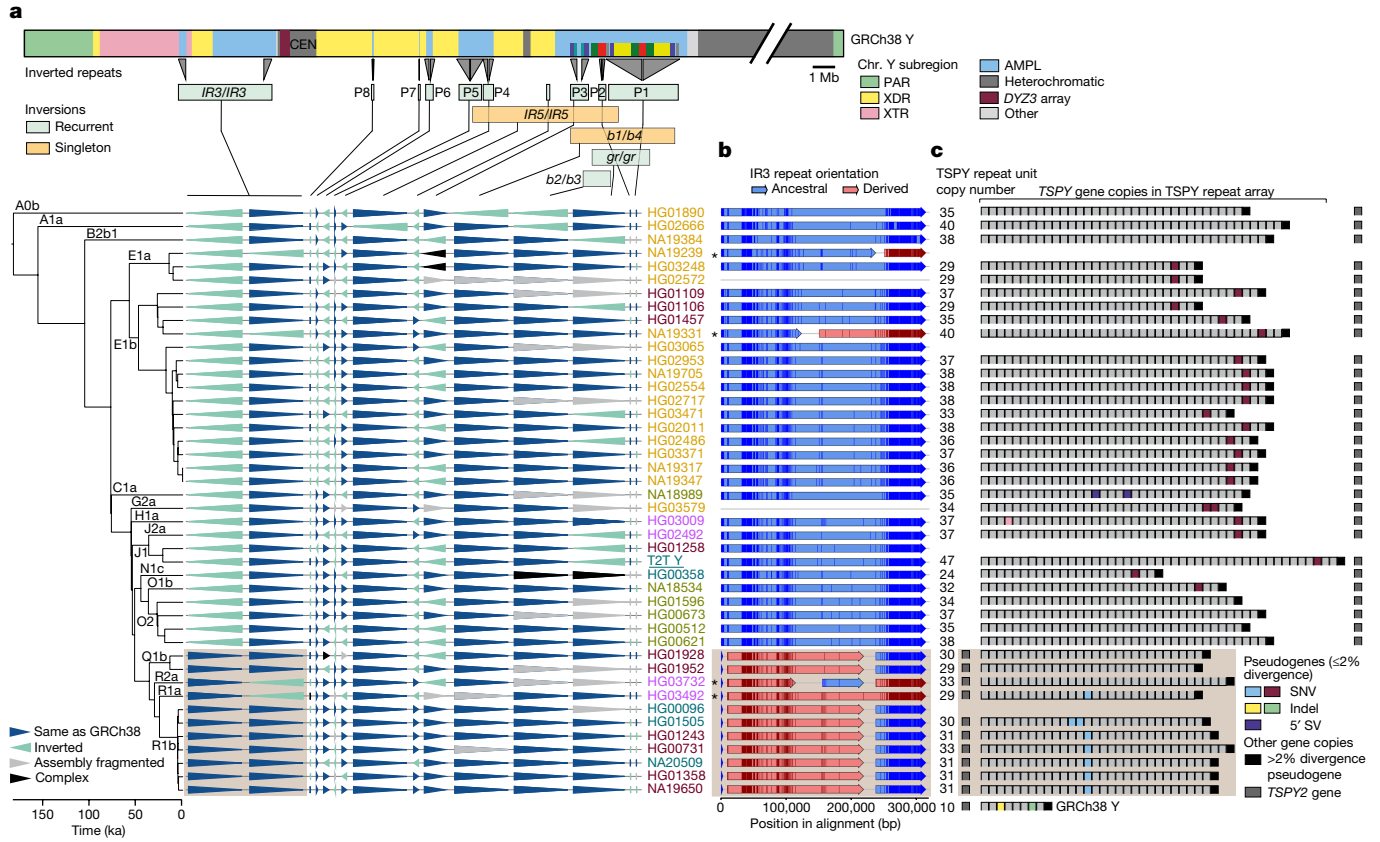


Fig. 3 | Characterization of large SVs. a, The distribution of 14 euchromatic inversions in the phylogenetic context, with the schematic of the GRCh38 Y-chromosome structure shown above, annotated with Y subregions, inverted repeat locations, palindromes (P1–P8) and SDs in ampliconic subregion 7 (Supplementary Fig. 34), with inverted segments indicated below. Samples are coloured as in Fig. 1b. AMPL, ampliconic; CEN, centromere; ka, thousand years ago. **b**, Inversion breakpoint identification in the IR3 repeats. The light brown shaded box (also in **a** and **c**) indicates samples that have probably undergone two inversions: one changing the location of the single *TSPY2* gene-containing repeat unit from the proximal to distal IR3 repeat, and the second reversing the region between the IR3 repeats, shared by haplogroup QR samples (Supplementary Fig. 34 and Supplementary Results (Y-chromosomal Inversions)). The asterisks indicate samples that have undergone an additional IR3 inversion. Informative paralogous sequence variants are shown as vertical darker lines in each of the arrows. Samples with non-contiguous IR3 assembly are indicated by grey lines. **c**, The distribution of pseudogenes within the TSPY repeat array. The total number of *TSPY* genes located within the approximately 20.3-kb TSPY repeat units is shown on the left. The samples marked with asterisks in **b** carry the TSPY array in reverse orientation and were reoriented for visualization. The low-divergence (≤2%) pseudogenes (coloured boxes) originate from five events: two nonsense mutations (light blue, maroon), two single-nucleotide indel deletions (yellow, green) and one 5′ structural variation that deletes around 370 nucleotides of the proximal half of exon 1 (purple). An additional sixth event was identified (that is, a premature stop codon within the fourth *TSPY* copy in the array of HG03009, in pink), but was deemed to be unlikely to result in nonsense-mediated decay as it was located only three codons before the canonical stop codon. Sample IDs and phylogenetic relationships are as described in **a**.

The euchromatic regions showed comparatively little variation in size (Fig. 2a, Supplementary Fig. 7 and Supplementary Tables 14 and 16) with the exception of the ampliconic subregion 2 that contains a copy-number variable TSPY repeat array, composed of 20.3-kb repeat units. The TSPY array size varies by up to 467 kb between individuals (Methods, Extended Data Fig. 1c,d, Supplementary Figs. 18 and 22, Supplementary Tables 15–18 and Supplementary Results (Gene family architecture and evolution)) and was consistently shorter among male individuals within haplogroup QR (from 567 to 648 kb, mean 603 kb) compared with male individuals in the other haplogroups (from 465 to 932 kb, mean = 701 kb) (Supplementary Figs. 18 and 22–25). The concordance of observed size variation with the phylogeny is well supported by relatively constant, phylogenetically independent contrasts across the phylogeny (Methods, Supplementary Figs. 23–24 and Supplementary Table 19). Such phylogenetic consistency reinforces the high quality of our assemblies even across homogeneous tandem arrays, as more closely related Y chromosomes are expected to be more similar, and this consequently enables the investigation of mutational dynamics across well-defined timeframes.

Distribution and frequency of genetic variants

We used our assemblies to produce a set of variant calls for each Y chromosome, including structural variants (SVs), insertions or deletions (indels) and single-nucleotide variants (SNVs). In the MSY, we report on average 88 insertion and deletion SVs (≥50 bp), three large inversions (>1 kb), 2,168 indels (<50 bp) and 3,228 SNVs per Y assembly (Methods, Extended Data Fig. 2a and Supplementary Table 20) compared with the GRCh38 Y reference. Variants were merged across all 43 samples to produce a non-redundant callset of 876 SVs (488 insertions, 378 deletions, 10 inversions), 23,459 indels (10,283 insertions, 13,176 deletions) and 53,744 SNVs (Supplementary Tables 21–25 and Supplementary Results (Orthogonal support to Y-chromosomal SVs and copy number variation)). On the basis of SV insertions, we identified an average of 81 kb (range of 46 to 155 kb) of novel, non-reference sequences per Y chromosome. After excluding simple repeats and mobile element sequences, an average of 18 kb (range of 0.6 to 47 kb) of unique non-reference sequence per Y remained (Supplementary Table 26).

Across the unique regions of the autosomes, we found 1.91 SVs, 165.66 indels and 994.42 SNVs per Mb per haplotype (Methods and

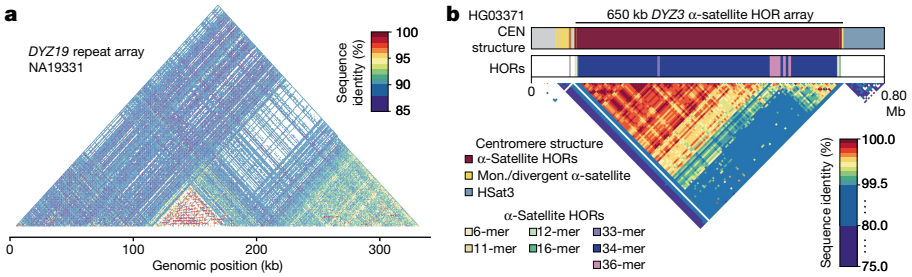


Fig. 4 | *DYZ19* and centromeric repeat arrays. a, Sequence identity heat map of the *DYZ19* repeat array from NA19331 (E1b1b1b2b2a1-M293) (using 1-kb window size), highlighting the higher sequence similarity within central and distal regions. **b**, The genetic landscape of the chromosome Y centromeric region from HG03371 (E1b1a1a1a1c1a-CTS1313). This centromere (CEN) contains the ancestral 36-monomer HORs, from which the canonical 34-monomer HOR is derived (Supplementary Fig. 46). Mon., monomer.

Supplementary Table 27). In the PAR1 region, on both the X and the Y chromosome, SV rates increased 1.98-fold to 3.79 SVs per Mb ($P = 2.37 \times 10^{-5}$, Welch's *t*-test) per haplotype, indels increased 1.56-fold to 259.14 per Mb ($P = 1.38 \times 10^{-3}$, Welch's *t*-test) and SNVs decreased slightly to 936.19 ($P = 1.00$, Welch's *t*-test) across unique loci. While PAR1 has the same ploidy as the autosomes, it is much shorter (2.8 Mb) and has a 10× increased recombination rate in male compared with female individuals[22], which may lead to the observed higher density of SVs and indels. A reduced level of variation observed in the proximal 500 kb before the currently established PAR1 boundary could be the result of a lower recombination rate closer to the sex-specific chromosomal regions, indicating a more distal location for the actual PAR1 boundary (Supplementary Figs. 26–29). As expected, the human chromosome X (excluding both PAR regions) exhibits lower genetic variation with 1.16 SVs ($P = 1.08 \times 10^{-25}$ Student's *t*-test), 106.64 indels ($P = 9.29 \times 10^{-46}$, Student's *t*-test) and 584.93 SNVs ($P = 8.23 \times 10^{-83}$, Student's *t*-test) per Mb of unique loci with most differences probably attributed to a lower effective population size for the X chromosome. The MSY has even less variation than seen for the X chromosome, with an average of 0.01 SVs, 2.11 indels and 5.72 SNVs per Mb ($P < 1 \times 10^{-100}$ for all, Welch's *t*-test) of unique loci (Supplementary Table 27). Bonferroni correction was applied to all tests.

We also identified 21 mobile element insertions across the 43 Y-chromosomal assemblies that are not present in the GRCh38 Y reference, including 15 *Alu* elements (4 out of 15 within Yq12) and six long interspersed element-1s (LINE-1s; no significant difference compared with the whole-genome distribution reported previously[20]) (Fig. 5f, Methods, Supplementary Tables 28 and 29 and Supplementary Results (Yq12 heterochromatic subregion)). Closer inspection across the three gaplessly assembled Y chromosomes, as well as the T2T Y chromosome, showed substantial differences in repeat composition between Y-chromosomal subregions (Supplementary Fig. 30 and Supplementary Tables 30 and 31). For example, the PARs showed a clear increase in short interspersed element content and a reduction in LINE and long terminal repeat content compared with the male-specific XTR, XDR and ampliconic regions (Extended Data Fig. 2b, Supplementary Fig. 30 and Supplementary Table 32).

Y-chromosomal inversions

Large inversions were identified using Strand-seq[23] and manual inspection of assembly alignments, yielding as many as 14 inversions in the euchromatic regions and two inversions within the Yq12 across the studied males (Figs. 3a and 5c, Methods, Extended Data Fig. 3, Supplementary Tables 33–35 and Supplementary Results (Y-chromosomal Inversions)). Six of these matched the ten inversions identified above by variant calling (Supplementary Table 23). The breakpoint intervals for 8 out of 14 of the euchromatic inversions were refined to DNA regions as small as 500 bp (Fig. 3b, Methods, Supplementary Figs. 31–33 and Supplementary Table 36). All of these inversions are flanked by highly similar (up to 99.97%) and large (from 8.7 kb to 1.45 Mb) inverted SDs and, although determination of the molecular mechanism generating Y-chromosomal inversions remains challenging, most are probably a result of non-allelic homologous recombination (NAHR). Moreover, we found that most (12 out of 14, 85%) euchromatic inversions are recurrent, with 2 to 13 toggling events in the Y phylogeny, which translates to an inversion rate estimate ranging from 3.68×10^{-5} (95% confidence interval (CI) = $3.25–4.17 \times 10^{-5}$) to 2.39×10^{-4} (95% CI = $2.11–2.71 \times 10^{-4}$) per locus per father-to-son Y transmission. The highest inversion recurrence is seen among the eight Y-chromosomal palindromes (P1–P8; Fig. 3a, Methods, Supplementary Fig. 34 and Supplementary Table 33). Taken together, we calculated a rate of 1 recurrent inversion per 603 (95% CI = 533–684) father-to-son Y transmissions. The per site per generation rate estimates for 12 Y-chromosomal recurrent inversions are significantly higher (greater than twofold difference between median estimates, two-tailed Mann–Whitney–Wilcoxon test, $n = 44$, $P < 0.0001$) than the rates previously estimated for 32 autosomal and X-chromosomal recurrent inversions[2].

There are two fixed inversions flanking the Yq12 subregion (Fig. 5c, Supplementary Fig. 35, Supplementary Table 35 and Supplementary Results (Y-chromosomal inversions)). The proximal inversion, observed in 10 out of 11 individuals analysed, ranged from 358.9 to 820.7 kb in size (mean = 649.0 kb) (Supplementary Table 35). By contrast, the distal inversion was observed in all 11 individuals and ranged from 259.5 to 641.4 kb in size (mean = 472.5 kb). We found that the breakpoints for these two inversions were identical among all of the individuals. This suggests that the consistent presence of these two inversions at both ends of the Yq12 subregion may prevent unequal sister chromatid exchange from occurring, restricting expansion and contraction of the repeat units to the region flanked by these two inversions.

Evolution of palindromes and multicopy gene families

To further reconstruct the evolution of Y-chromosomal palindromes, we investigated both the gene conversion patterns and evolutionary rates across the Y assemblies (Supplementary Figs. 36–38 and Supplementary Tables 15, 37 and 38). The intra-arm gene conversion patterns across seven palindromes (P1, P3–P8, but excluding the P2 palindrome and DNA sequences that are shared between different palindromes; Methods) showed a significant bias towards G or C nucleotides (942 events to G or C versus 701 to A or T nucleotides, $P = 2.75 \times 10^{-9}$, χ^2 test), but no bias towards the ancestral state (357 events to derived versus 374 to ancestral state; $P = 0.5295$, χ^2 test; Supplementary Table 37). Comparison of base substitution patterns for all eight palindromes and the eight XDR regions, across 13 Y chromosomes, indicated that different palindromes are evolving at different rates (Methods). The level of sequence variation (both in base substitutions and SVs) and estimated base substitution mutation rates were higher for palindromes P1, P2 and

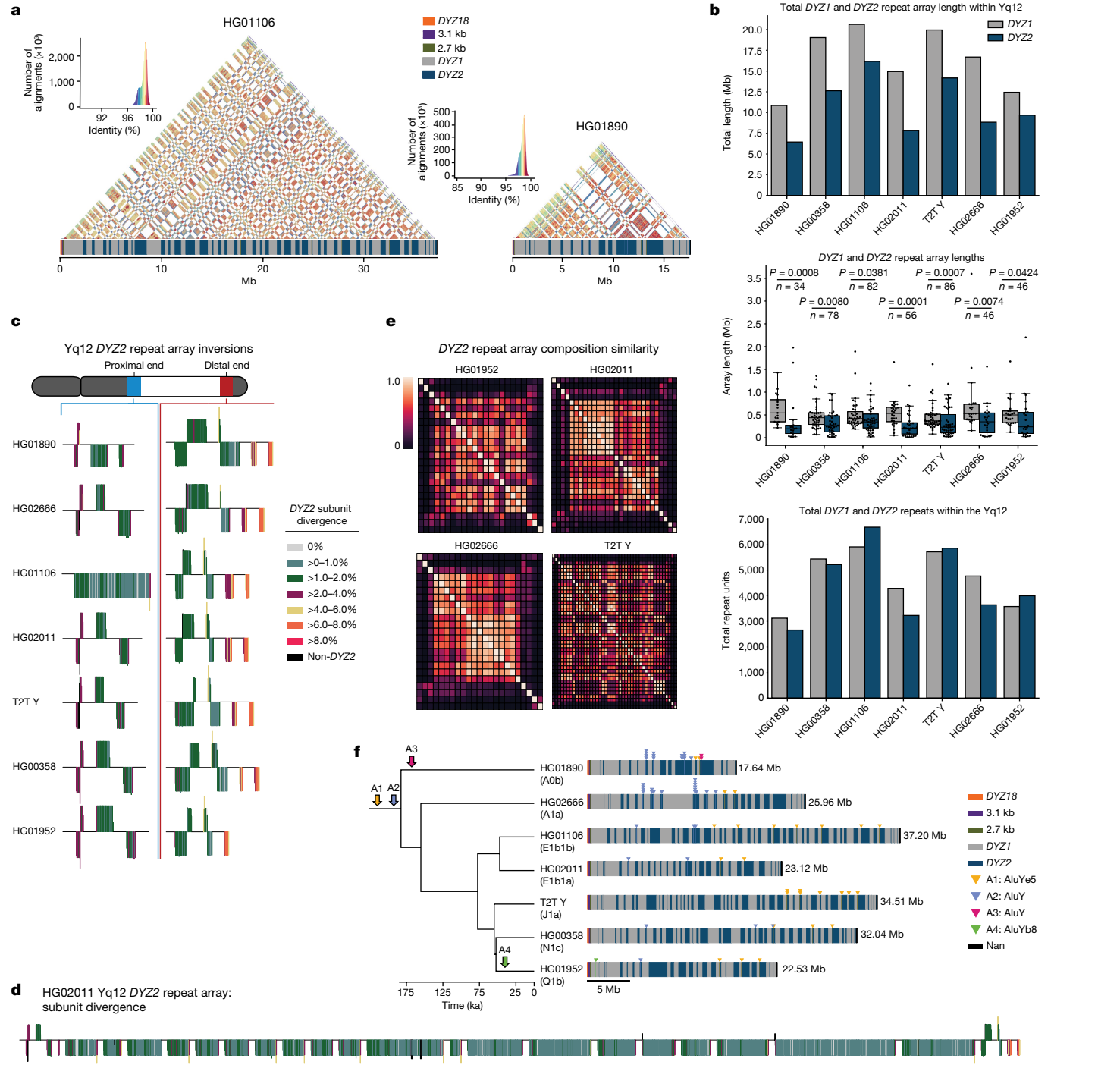


Fig. 5 | Yq12 heterochromatic region. a, Yq12 heterochromatic subregion sequence identity heat map in 5-kb windows for two samples with repeat array annotations. **b**, *DYZ1* and *DYZ2* total repeat array lengths (top), individual array lengths (middle) and the total number of *DYZ1* and *DYZ2* repeat units (bottom) within contiguously assembled genomes. The black dots represent individual arrays. Statistically significant *P* values comparing *DYZ1* and *DYZ2* array lengths within each assembly and *n* values (number of repeat arrays) are shown ($\alpha = 0.05$, two-sided Mann–Whitney *U*-test; Methods). For the box plots, the box limits indicate quartiles, the whiskers encompass the full range of the data (except for outliers) and the centre line shows the median. **c**, *DYZ2* repeat array inversions in the proximal and distal ends of the Yq12 subregion. *DYZ2* repeats are coloured on the basis of their divergence estimate and visualized based on their orientation (sense, up; antisense, down). **d**, Detailed representation of *DYZ2* subunit divergence estimates for HG02011 (the colour key in **c** applies also to **d**). **e**, The inter-*DYZ2* repeat array subunit composition similarity within a sample. Similarity was calculated using the Bray–Curtis index (1 − Bray–Curtis distance, where 1.0 = the same composition). *DYZ2* repeat arrays are shown in physical order from proximal to distal (from top down, and from left to right). **f**, Mobile element insertions identified in the Yq12 subregion. We identified four putative *Alu* insertions across the seven gapless Yq12 assemblies. Their approximate location, as well as the expansion and contraction dynamics of *Alu*-insertion-containing *DYZ* repeat units, are shown (right). After the insertion into the *DYZ* repeat units, lineage-specific contractions and expansions occurred. Two *Alu* insertions (A1 and A2) occurred before the radiation of Y haplogroups (at least 182,900 years ago), while two additional *Alu* elements represent lineage-specific insertions. The total length of the Yq12 region is indicated on the right. ka, thousand years ago.

P3, which contain higher proportions of multicopy (that is, more than two copies) segments compared with other palindromes (3.03×10^{-8} (95% CI = $2.80–3.27 \times 10^{-8}$) versus 2.12×10^{-8} (95% CI = $1.96–2.29 \times 10^{-8}$) mutations per position per generation, respectively) (Fig. 3a, Methods, Supplementary Figs. 34 and 36–38 and Supplementary Table 38). The increased variation of P1, P2 and P3 probably results from sequence exchange between multicopy regions.

The gene annotation of the Y-chromosomal assemblies showed no evidence of the loss of any MSY protein-coding genes in the 43 male individuals analysed (Supplementary Tables 39–43 and Supplementary Results (Gene annotation)). However, the investigation of three copy-number variable ampliconic gene families (*DAZ* (deleted in azoospermia), *TSPY* (testis-specific protein Y-linked 1), *RBMY1* (RNA-binding motif (RRM) gene on Y chromosome); Supplementary Results (Gene family architecture and evolution)) revealed substantial differences in their genetic diversity and evolution. Although only 2 out of 43 samples (41 assemblies, T2T Y and GRCh38 Y) showed a difference in the *DAZ* copy number (two and six *DAZ* copies versus four in all others), extensive variation was detected in the copy number of the 28 canonical exons (from 0 to 14 copies of a single exon) between samples (Methods, Supplementary Fig. 39 and Supplementary Tables 15, 44 and 45). Consistent with previous reports, *RBMY1* genes were primarily located in four separate regions, while three samples had undergone larger rearrangements[24] (Extended Data Fig. 4, Supplementary Fig. 40 and Supplementary Results (Gene family architecture and evolution))[24]. On average, 8 *RBMY1* gene copies (from 5 to 11) were identified, with most of the variation caused by expansions or contractions in regions 1 and 2 (Extended Data Fig. 4 and Supplementary Table 39). A phylogenetic analysis of *RBMY1* genes revealed that the gene copies from regions 3 and 4 have probably given rise to *RBMY1* genes located in regions 1 and 2 (Extended Data Fig. 4 and Supplementary Figs. 40 and 41), additionally supported by the analysis of the chimpanzee (PanTro6) *RBMY1* sequences (Supplementary Results (Gene family architecture and evolution)).

The majority of the *TSPY* genes are located in a tandemly organized and highly copy-number variable TSPY array. While a single repeat unit containing the *TSPY2* gene is located upstream of the *TSPY* array in GRCh38, we inferred that the ancestral position of *TSPY2* lies between the TSPY repeat array and the Y centromere in reverse orientation (Fig. 3c, Extended Data Fig. 5, Supplementary Fig. 42, Supplementary Table 36 and Supplementary Results (Y-chromosomal inversions and Gene family evolution)). Probably the result of two inversions or a complex rearrangement, the localization of *TSPY2* upstream of the *TSPY* array is shared by all QR haplogroup (including the GRCh38 Y) individuals. On average, 33 TSPY gene copies (from 23 to 39, 46 in T2T Y, counts include only low-divergence (≤2%) *TSPY* gene copies from the *TSPY* repeat array and *TSPY2*) were identified per assembly (Fig. 3c, Extended Data Fig. 5, Supplementary Fig. 43 and Supplementary Tables 15 and 40). Both network and phylogenetic analysis of *TSPY* gene sequences support identification of the ancestral gene copy (Methods and Extended Data Fig. 5b,c (medium blue)). Notably, five independently arisen pseudogenes were identified among the *TSPY* genes located within the 41 *TSPY* arrays analysed (39 contiguous assemblies, the T2T Y and the GRCh38 Y), with 31 out of 41 samples carrying at least one pseudogene (Fig. 3c). The phylogenetic distribution suggests periodic purging of pseudogenes from the array, possibly through the removal of deleterious mutations by gene conversion and NAHR. Evidence of gene conversions and NAHR was found both between the tandemly repeated *TSPY* gene copies and the *RBMY1* genes (Extended Data Figs. 4 and 5).

Epigenetic variation

The ONT sequencing data also provide a means to examine the base-level epigenetic landscape of the Y chromosomes (Extended Data Fig. 6). Here, we focused on DNA methylation at CpG sites, hereafter referred to as DNAme. In 41 samples (EBV-transformed lymphoblastoid cell lines) that passed quality control (Methods), we first tested the association between chromosome Y assembly length and global DNAme levels, as has previously been shown in *Drosophila*[25]. We detected a significant relationship between the chromosome Y assembly length and global DNAme levels, both genome wide and for the Y chromosome (linear model, $P = 0.0477$ and $P = 0.0469$ ($n = 41$); Supplementary Fig. 44 and Supplementary Results (Functional analysis)). We found 2,861 DNAme segments that vary across these Y chromosomes (Extended Data Fig. 7 and Supplementary Table 46). Notably, 21% of the variation in DNAme levels is associated with haplogroups (permutational analysis of variance (PERMANOVA), $P = 0.003$, $n = 41$), while the same is true for only 4.8% of the expression levels (PERMANOVA, $P = 0.005$, $n = 210$, leveraging the Geuvadis RNA-sequencing (RNA-seq) expression data[26]; Methods). This association is particularly strong for five genes (*BCORP1* (Supplementary Fig. 45), *LINC00280*, *LOC100996911*, *PRKY* and *UTY*), where both DNAme and gene-expression effects are observed (Supplementary Tables 46 and 47). Finally, we found 194 Y-chromosomal genetic variants, including a 171-bp insertion and one inversion, that impact DNAme levels on chromosome Y (Supplementary Table 48 and Supplementary Results (Functional analysis)). This suggests that some of the genetic background, either on the Y chromosome or elsewhere in the genome, may impact the functional outcome (the epigenetic and transcriptional profiles) of specific genes on the Y chromosome.

Variation of the heterochromatic regions

Variation in the size and structure of centromeric/pericentromeric repeat arrays

In general, the chromosome Y centromeres are composed of 171-bp *DYZ3* α-satellite repeat units[1], organized into a higher-order repeat (HOR) array, flanked on either side by short stretches of monomeric α-satellite. The α-satellite HOR arrays across gaplessly assembled Y centromeres ranged in size from 264 kb to 1.165 Mb (mean = 671 kb), with smaller arrays found in haplogroup R1b samples compared with other lineages[27,28] (mean = 341 kb versus 787 kb, respectively; Methods, Extended Data Fig. 8a, Supplementary Figs. 18, 23 and 24 and Supplementary Tables 15 and 16). We determined that the *DYZ3* α-satellite HOR array is mostly composed of a 34-monomer repeating unit that is the most prevalent HOR type found in the 21 analysed samples (Fig. 4b, Methods and Supplementary Fig. 46). However, we identified two other HORs that were present at a high frequency among the analysed Y chromosomes: a 35-monomer HOR found in 14 out of 21 samples and a 36-monomer HOR found in 11 out of 21 samples (Methods). While the 35-monomer HOR is present across different Y lineages in the Y phylogeny, the 36-monomer HOR has been lost in phylogenetically closely related Y chromosomes representing the QR haplogroups (Extended Data Fig. 8a). Analysis of the sequence composition of these HORs revealed that the 36-monomer HOR probably represents the ancestral state of the canonical 35-mer and 34-mer HOR after deletion of the 22nd α-satellite monomer in the resulting HORs, respectively (Methods and Supplementary Fig. 46).

The overall organization of the *DYZ3* α-satellite HOR array is similar to that found on other human chromosomes, with near-identical α-satellite HORs in the core of the centromere that become increasingly divergent towards the periphery[29–32]. There is a directionality of the divergent monomers at the periphery of the Y centromeres such that a larger block of diverged monomers is consistently found at the p-arm side of the centromere compared with the block of diverged monomers juxtaposed to the q-arm (Fig. 4b, Extended Data Fig. 8b and Supplementary Figs. 47 and 48).

Adjacent to the *DYZ3* α-satellite HOR array on the q-arm is a human satellite III (*HSat3*) repeat array, which ranges in size from 372 to 488 kb (mean = 378 kb), followed by a *DYZ17* repeat array, which ranges in size from 858 kb to 1.740 Mb (mean = 1.085 Mb). A comparison of the

sizes of these three repeat arrays revealed no significant correlation among their sizes (Supplementary Figs. 47–49 and Supplementary Tables 15 and 16).

The *DYZ19* repeat array is located on the long arm, flanked by XDRs (Fig. 1a) and composed of 125-bp repeat units (fragment of a long terminal repeat) in a head-to-tail manner. This subregion was completely assembled across all 43 Y chromosomes and, among subregions, exhibits the highest variation with a 6.7-fold difference in size (from 63.5 to 428 kb). The HG02492 individual (haplogroup J2a) with the smallest-sized *DYZ19* repeat array has a deletion of approximately 200 kb in this subregion (Supplementary Table 16). In 43 out of 44 Y chromosomes (including T2T Y), we found evidence of at least two rounds of mutation/expansion (Fig. 4a (green and red coloured blocks) and Supplementary Figs. 19–21), leading to directional homogenization of the central and distal parts of the region in all Y chromosomes. Finally, we observed a recent approximately 80-kb duplication event shared by the 11 phylogenetically related haplogroup QR samples (Supplementary Figs. 19–21) that must have occurred approximately 36,000 years ago (Fig. 1b and Supplementary Fig. 1), resulting in a substantially larger overall *DYZ19* subregion in these Y chromosomes.

Between the Yq11 and the Yq12 subregions lies the *DYZ18* subregion, which comprises three distinct repeat arrays: a *DYZ18* repeat array and novel 3.1-kb and 2.7-kb repeat arrays (Extended Data Fig. 9 and Supplementary Figs. 50–56). The 3.1-kb repeat array is composed of degenerate copies of the *DYZ18* repeat unit, exhibiting 95.8% sequence identity (using SNVs only) across the length of the repeat unit. The 2.7-kb repeat array seems to have originated from both the *DYZ18* (23% of the 2.7-kb repeat unit shows 86.3% sequence identity to *DYZ18*) and *DYZ1* (77% of the 2.7-kb repeat unit shows 97% sequence identity to *DYZ1*) repeat units (Supplementary Fig. 50). All three repeat arrays (*DYZ18*, 3.1 kb and 2.7 kb) show a similar pattern and level of methylation compared to the *DYZ1* repeat arrays (Supplementary Fig. 57), in that we observe constitutive hypermethylation.

Composition of the Yq12 heterochromatic subregion

The Yq12 subregion is the most challenging portion of the Y chromosome to assemble contiguously owing to its highly repetitive nature and size. Here we completely assembled the Yq12 subregion for six individuals and compared it to the Yq12 subregion of the T2T Y chromosome (Figs. 1a and 5a,f, Supplementary Tables 14–16 and Supplementary Results (Yq12 heterochromatic subregion)). This subregion is composed of alternating arrays of the repeat units *DYZ1* and *DYZ2*[1,6,33–36]. The *DYZ1* repeat unit is approximately 3.5 kb and consists mainly of simple repeats and pentameric satellite sequences, and has been recently referred to as HSat3A6[5]. The *DYZ2* repeat (which has been recently referred to as HSat1B[31]) is approximately 2.4 kb and consists mainly of a tandemly repeated AT-rich simple repeat fused to a 5′ truncated *Alu* element followed by an HSATI satellite sequence (Supplementary Fig. 50).

The *DYZ1* repeat units are tandemly arranged into larger *DYZ1* repeat arrays, as are the *DYZ2* repeat units, and the *DYZ1* and *DYZ2* repeat arrays alternate with one another (Fig. 5). The total number of *DYZ1* and *DYZ2* arrays (range, 34 to 86; mean = 61) were significantly positively correlated (Spearman correlation = 0.90, $P = 0.0056$, $n = 7$, $\alpha = 0.05$) with the total length of the analysed Yq12 region (Supplementary Fig. 58), whereas the length of the individual *DYZ1* and *DYZ2* repeat arrays were found to be widely variable (Fig. 5b and Supplementary Fig. 59). The *DYZ1* arrays were significantly longer (range, 50,420 to 3,599,754 bp; mean = 535,314 bp) than the *DYZ2* arrays (range, 11,215 to 2,202,896 bp; mean = 354,027 bp, two-tailed Mann–Whitney *U*-test, $P < 0.05$ for all seven assemblies with a complete Yq12 region). However, the total number of each repeat unit was nearly equal within each Y chromosome (*DYZ1* to *DYZ2* ratio ranges from 0.88 to 1.33; mean = 1.09) (Fig. 5b and Supplementary Table 49). From ONT data, we observed a consistent hypermethylation of the *DYZ2* repeat arrays compared with the *DYZ1* repeat arrays, the sequence composition of the two repeats is markedly different in terms of CG content (24% *DYZ2* versus 38% *DYZ1*) and the number of CpG dinucleotides (1 CpG per 150-bp *DYZ2* versus 1 CpG per 35-bp *DYZ1*) potentially explains the marked differences in DNA methylation (Extended Data Fig. 6).

Sequence analysis of the repeat units in the Yq12 suggests that the *DYZ1* and *DYZ2* repeat arrays and the entire Yq12 subregion may have evolved in a similar manner, and similarly to the centromeric region (see above). Specifically, repeat units near the middle of a given array showed a higher level of sequence similarity to each other than to the repeat units at the distal regions of the repeat array (Fig. 5d and Extended Data Figs. 9 and 10). This suggests that expansion and contraction tend to occur in the middle of the repeat arrays, homogenizing these units yet allowing divergent repeat units to accumulate towards the periphery. Similarly, when looking at the entire Yq12 subregion, we observed that repeat arrays located in the middle of the Yq12 subregion tend to be more similar in sequence to each other than to repeat arrays at the periphery (Fig. 5e, Extended Data Figs. 9 and 10 and Supplementary Figs. 60). This observation is supported by results from the *DYZ2* repeat divergence analysis, the inter-*DYZ2* array profile comparison and the construction of a *DYZ2* phylogeny (Methods and Supplementary Fig. 61).

Discussion

The mammalian Y chromosome has been notoriously difficult to assemble owing to its extraordinarily high-repeat content. Here we present the Y-chromosomal assemblies of 43 male individuals from the 1000 Genomes Project dataset and a comprehensive analysis of their genetic and epigenetic variation and composition. Although both the GRCh38 Y and the T2T Y assemblies represent relatively recently emerged (TMRCA 54,500 years ago (95% HPD interval: 47,600–62,400 years ago); Supplementary Fig. 1) European Y lineages, half of our Y chromosomes carry African Y lineages, including two of the deepest-rooted human Y lineages (A0b and A1a, TMRCA 182,900 years ago (95% HPD interval: 159,800–209,200 years ago)), which we gaplessly assembled, enabling us to investigate how the Y chromosome has changed over 180,000 years of human evolution.

We were able to comprehensively and precisely examine the extent of genetic variation down to the nucleotide level across multiple human Y chromosomes. The MSY can be approximately divided into two portions: the euchromatic and the heterochromatic regions. The single-copy protein-coding MSY genes, present in the GRCh38 Y reference sequence, are conserved in all 43 Y assemblies with few SNVs. The low SNV diversity in Y is concordant with previous studies and consistent with models of natural demographic processes such as extreme male-specific bottlenecks in recent human history and purifying selection removing deleterious mutations and linked variation[12,14,37]. The multi-copy protein-coding MSY genes are often copy-number variable. We found that 5 out of 8 multicopy gene families showed variation in terms of copy number, with the highest variation observed in the *TSPY* gene family (23 to 39 copies, 46 in the T2T Y; Fig. 3c and Supplementary Table 40). An investigation of three copy-number variable gene families (*TSPY*, *RBMY1* and *DAZ*) revealed different modes of evolution, probably resulting from differences in structural composition of the genomic regions. For example, the majority of the *TSPY* genes are located within a tandemly repeated array, undergoing frequent expansions and contractions, where we also find evidence of lineage-specific acquisition and purging of pseudogenes.

The euchromatic region contains additional structural variation across the 43 individuals. Notably, we identified 14 inversions that together affect half of the Y-chromosomal euchromatin, with only the most closely related pair of African Y chromosomes (from NA19317 and NA19347) showing the exact same inversion composition. Of these 14 inversions, 12 showed recurrent toggling in recent human

history, including five previously undescribed recurrent inversions[2]. We narrowed down the breakpoints for all of the inversions and have refined the breakpoints down to a 500-bp region for 8 out of 14 inversions. The determination of the molecular mechanism causing the inversions remains challenging; however, the increased recurrent inversion rate on the Y chromosome compared with the rest of the human genome may be in part due to DNA double-stranded breaks being repaired by intrachromatid recombination[2,38]. The enrichment in highly similar (inverted) SDs[1,4] prone to NAHR, coupled with reduced selection to maintain gene order, may explain the high prevalence of recurrent inversions on the Y. The majority of the recurrent inversions (8 out of 14) occurs between highly similar SDs termed palindromes P1–P8 (Fig. 3a). Three of the palindromes appear to be evolving at faster rates compared with the other five palindromes and the unique XDR regions of the Y chromosome, probably due to sequence exchange between multicopy (that is, more than two copies) SDs.

In the PAR1 region, we found evidence of enrichment of indels and SVs compared with in the autosomes, and the rest of the X and the Y chromosomes, potentially resulting from a higher recombination rate in this region during male meiosis[22]. Notably, there is a reduction of genetic variation in the proximal 500 kb of PAR1, indicating a reduced recombination rate here and suggesting that the actual PAR1 boundary probably lies distal to the currently established boundary[1].

There are four heterochromatic subregions in the human Y chromosome: the (peri)centromeric region, *DYZ18*, *DYZ19* and Yq12. Heterochromatin is usually defined by the preponderance of highly repetitive sequences and the constitutive dense packaging of the chromatin within[39]. When we examined the DNA sequence and the methylation patterns for these four heterochromatic subregions, the highly repetitive sequence content and the high level of methylation (Extended Data Fig. 6 and Supplementary Fig. 57) observed is consistent with the definition of heterochromatin. Furthermore, resolving the complete structural variation in the heterochromatic regions of the human Y chromosome provides molecular archaeological evidence for evolutionary mechanisms. For example, we show how the higher-order structure at the centromeric region of the Y chromosome evolved from an ancestral 36-mer HOR to a 34-mer HOR, which predominates in the centromeres of male humans[40]. Moreover, the degeneration of these repeat units of the (peri)centromeric region of the Y chromosome has a directional bias towards the p-arm side. The presence of an *Alu* element right at the q-arm boundary, but not on the p-arm side, raises the possibility that two *Alu* insertions, over 180,000 years ago, led to a subsequent *Alu*–*Alu* recombination that deleted the region in between and removed the diverged centromeric sequence block[41]. In the Yq12 subregion, we found evidence for localized expansions and contractions of the *DYZ1* and *DYZ2* repeat units, although the preservation of nearly a 1:1 ratio among all of the male individuals studied indicates functional or evolutionary constraints.

Here, we fully sequenced and analysed 43 diverse Y chromosomes and identified the full extent of variation of this chromosome across more than 180,000 years of human evolution, offering a major advance to our understanding of how non-recombining regions of the genome evolve and persist. Sequence-level resolution across multiple human Y chromosomes reveals new DNA sequences and new elements of conservation, and provides molecular data that give us important insights into genomic stability and chromosomal integrity. It also offers the possibility to investigate the molecular mechanisms and evolution of repetitive sequences across a well-defined timeframe without the encumbrances of meiotic recombination. Ultimately, the ability to effectively assemble the complete human Y chromosome has been a long-awaited yet crucial milestone towards understanding the full extent of human genetic variation and also provides the starting point to associate Y-chromosomal sequences to specific human traits and more thoroughly study human evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06425-6.

1. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
2. Porubsky, D. et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005 (2022).
3. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B* **355**, 1563–1572 (2000).
4. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
5. 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).
6. Nakahori, Y., Mitani, K., Yamada, M. & Nakagome, Y. A human Y-chromosome specific repeated DNA family (DYZ1) consists of a tandem array of pentanucleotides. *Nucleic Acids Res.* **14**, 7569–7580 (1986).
7. Cooke, H. Repeated sequence specific to human males. *Nature* **262**, 182–186 (1976).
8. Skov, L., The Danish Pan Genome Consortium & Schierup, M. H. Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion. *PLoS Genet.* **13**, e1006834 (2017).
9. Kuderna, L. F. K. et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. *Nat. Commun.* **10**, 4 (2019).
10. Rhie, A. et al. The complete sequence of a human Y chromosome. *Nature* https://doi.org/10.1038/s41586-023-06457-y (2023).
11. Sahakyan, H. et al. Origin and diffusion of human Y chromosome haplogroup J1-M267. *Sci. Rep.* **11**, 6659 (2021).
12. 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).
13. The Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339–348 (2002).
14. Karmin, M. et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* **25**, 459–466 (2015).
15. Hallast, P., Agdzhoyan, A., Balanovsky, O., Xue, Y. & Tyler-Smith, C. A Southeast Asian origin for present-day non-African human Y chromosomes. *Hum. Genet.* **140**, 299–307 (2021).
16. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01662-6 (2023).
17. 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).
18. Lang, D. et al. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* **9**, giaa123 (2020).
19. Mikheenko, A., Bzikadze, A. V., Gurevich, A., Miga, K. H. & Pevzner, P. A. TandemTools: mapping long reads and assessing/improving assembly quality in extra-long tandem repeats. *Bioinformatics* **36**, i75–i83 (2020).
20. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
21. Vollger, M. R. et al. Long-read sequence and assembly of segmental duplications. *Nat. Methods* **16**, 88–94 (2019).
22. Bergman, J. & Schierup, M. H. Evolutionary dynamics of pseudoautosomal region 1 in humans and great apes. *Genome Biol.* **23**, 215 (2022).
23. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
24. Shi, W. et al. Evolutionary and functional analysis of *RBMY1* gene copy number variation on the human Y chromosome. *Hum. Mol. Genet.* **28**, 2785–2798 (2019).
25. Brown, E. J., Nguyen, A. H. & Bachtrog, D. The *Drosophila* Y chromosome affects heterochromatin integrity genome-wide. *Mol. Biol. Evol.* **37**, 2808–2824 (2020).
26. Lappalainen, T. et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
27. Miga, K. H. et al. Centromere reference models for human chromosomes X and Y satellite arrays. *Genome Res.* **24**, 697–707 (2014).
28. Oakey, R. & Tyler-Smith, C. Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics* **7**, 325–330 (1990).
29. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**, 79–84 (2020).
30. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
31. Altemose, N. et al. Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
32. Gershman, A. et al. Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
33. Cooke, H. J. & McKay, R. D. Evolution of a human Y chromosome-specific repeated sequence. *Cell* **13**, 453–460 (1978).

34. Rahman, M. M., Bashamboo, A., Prasad, A., Pathak, D. & Ali, S. Organizational variation of DYZ1 repeat sequences on the human Y chromosome and its diagnostic potentials. *DNA Cell Biol.* **23**, 561–571 (2004).
35. Pathak, D., Premi, S., Srivastava, J., Chandy, S. P. & Ali, S. Genomic instability of the DYZ1 repeat in patients with Y chromosome anomalies and males exposed to natural background radiation. *DNA Res.* **13**, 103–109 (2006).
36. 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).
37. Wilson Sayres, M. A., Lohmueller, K. E. & Nielsen, R. Natural selection reduced diversity on human Y chromosomes. *PLoS Genet.* **10**, e1004064 (2014).
38. Lange, J. et al. Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**, 855–869 (2009).
39. Verma, R. S. *Heterochromatin: Molecular and Structural Aspects* (Cambridge Univ. Press, 1988).
40. Tyler-Smith, C. & Brown, W. R. Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* **195**, 457–470 (1987).
41. Cooper, K. F., Fisher, R. B. & Tyler-Smith, C. Structure of the sequences adjacent to the centromeric alphoid satellite DNA array on the human Y chromosome. *J. Mol. Biol.* **230**, 787–799 (1993).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature Limited 2023

Human Genome Structural Variation Consortium (HGSVC)

Pille Hallast[1,18], Peter Ebert[2,3,4,18], Mark Loftus[5,6,18], Feyza Yilmaz[1], Peter A. Audano[1], Glennis A. Logsdon[7], Marc Jan Bonder[8,9], Weichen Zhou[10], Wolfram Höps[11], Kwondo Kim[1], Chong Li[12], Philip C. Dishuck[7], David Porubsky[7], Fotios Tsetsos[1], Jee Young Kwon[1], Qihui Zhu[1], Katherine M. Munson[7], Patrick Hasenfeld[11], William T. Harvey[7], Alexandra P. Lewis[7], Jennifer Kordosky[7], Kendra Hoekzema[7], Jan O. Korbel[11], Evan E. Eichler[7,17], Xinghua Shi[12], Christine R. Beck[1,14,15], Tobias Marschall[2,4], Miriam K. Konkel[5,6,19] & Charles Lee[1,19]

Methods

Sample selection

Samples were selected from the 1000 Genomes Project Diversity Panel[42] and one representative was selected from 21 populations (Supplementary Table 1). A total of 13 out of 28 samples were included from the Human Genome Structural Variation Consortium (HGSVC) Phase 2 dataset, which was published previously[20]. Furthermore, for 15 out of 28 samples, data were newly generated as part of the HGSVC efforts (see the 'Data production' section for details). We also included 15 samples from the Human Pangenome Reference Consortium (HPRC) (Supplementary Table 1). Notably, there is an African Y lineage (A00) older than the lineages in our dataset (TMRCA 253,884 years ago; 95% CI = 191,875–307,116 years ago[14,43]) that we could not include due to sample availability issues.

Data production

Data generated as part of this project were derived from lymphoblast lines available from the Coriell Institute for Medical Research for research purposes (https://www.coriell.org/), and authenticated using Illumina high-coverage data from a previous study[44]. Regular checks for mycoplasma contamination are performed at the Coriell Institute, which maintains the cell lines.

PacBio HiFi sequence production. University of Washington. Sample HG00731 data have been previously described[20]. Additional samples HG02554 and HG02953 were prepared for sequencing in the same way but with the following modifications: isolated DNA was sheared using the Megaruptor 3 instrument (Diagenode) twice using settings 31 and 32 to achieve a peak size of ~15–20 kb. The sheared material was processed for SMRTbell library preparation using the Express Template Prep Kit v2 and SMRTbell Cleanup Kit v2 (PacBio). After checking for size and quantity, the libraries were size-selected on the Pippin HT instrument (Sage Science) using the protocol '0.75% agarose, 15–20 kb high pass' and a cut-off of 14–15 kb. Size-selected libraries were checked by fluorometric quantitation (Qubit) and pulse-field sizing (FEMTO Pulse). All cells were sequenced on the Sequel II instrument (PacBio) using 30 h video times using version 2.0 sequencing chemistry and 2 h pre-extension. HiFi/CCS analysis was performed using SMRT Link (v.10.1) using an estimated read-quality value of 0.99.

The Jackson Laboratory. High-molecular-mass DNA was extracted from 30 million frozen pelleted cells using the Gentra Puregene extraction kit (Qiagen). Purified gDNA was assessed using fluorometric (Qubit, Thermo Fisher Scientific) assays for quantity and FEMTO Pulse (Agilent) for quality. For HiFi sequencing, samples exhibiting a mode size above 50 kb were considered to be good candidates. Libraries were prepared using the SMRTBell Express Template Prep Kit 2.0 (Pacbio). In brief, 12 µl of DNA was first sheared using gTUBEs (Covaris) to target 15–18 kb fragments. Two 5 µg of sheared DNA were used for each prep. DNA was treated to remove single-stranded overhangs, followed by DNA damage repair and end repair/A-tailing. The DNA was then ligated with a V3 adapter and purified using Ampure beads. The adapter ligated library was treated with Enzyme mix 2.0 for nuclease treatment to remove damaged or non-intact SMRTbell templates, followed by size selection using Pippin HT generating a library with a size >10 kb. The size-selected and purified >10 kb fraction of libraries was used for sequencing on the Sequel II (Pacbio) system.

ONT-UL sequence production. University of Washington. High-molecular-mass DNA was extracted from 2 aliquots of 30 million frozen pelleted cells using the phenol–chloroform approach as described previously[45]. Libraries were prepared using the Ultra-Long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendations. In brief, DNA from around 10 million cells was incubated with 6 µl of fragmentation mix at room temperature for 5 min and 75 °C for 5 min. This was followed by an addition of 5 µl of adapter (RAP-F) to the reaction mix and incubated for 30 min at room temperature. The libraries were cleaned up using Nanobind disks (Circulomics) and long fragment buffer (SQK-ULK001, ONT) and eluted in elution buffer. Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT) for 96 h. A library was split into 3 loads, with each load going 24 h followed by a nuclease wash (EXP-WSH004, ONT) and subsequent reload.

The Jackson Laboratory. High-molecular-mass DNA was extracted from 60 million frozen pelleted cells using the phenol–chloroform approach as previously described[46]. Libraries were prepared using the Ultra-Long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendations. In brief, 50 µg of DNA was incubated with 6 µl of fragmentation mix at room temperature for 5 min and 75 °C for 5 min. This was followed by an addition of 5 µl of adapter (RAP-F) to the reaction mix and incubated for 30 min at room temperature. The libraries were cleaned up using Nanodisks (Circulomics) and eluted in elution buffer. Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT) system for 96 h. A library was generally split into 3 loads with each loaded at an interval of about 24 h or when pore activity dropped to 20%. A nuclease wash was performed using the Flow Cell Wash Kit (EXP-WSH004) between each subsequent load.

Bionano Genomics optical genome maps production. Optical mapping data were generated at Bionano Genomics. Lymphoblastoid cell lines were obtained from Coriell Cell Repositories and grown in RPMI 1640 medium with 15% FBS, supplemented with L-glutamine and penicillin–streptomycin, at 37 °C and 5% CO_2. Ultra-high-molecular-mass DNA was extracted according to the Bionano Prep Cell Culture DNA Isolation Protocol (document number 30026, revision F) using the Bionano SP Blood & Cell DNA Isolation Kit (80030). In brief, 1.5 million cells were centrifuged and resuspended in a solution containing detergents, proteinase K and RNase A. DNA was bound to a silica disk, washed, eluted and homogenized by 1 h end-over-end rotation at 15 rpm, followed by an overnight rest at room temperature. Isolated DNA was fluorescently tagged at the motif CTTAAG by the enzyme DLE-1 and counter-stained using the Bionano Prep DNA Labeling Kit—DLS (8005) according to the Bionano Prep Direct Label and Stain (DLS) protocol (document number 30206, revision G). Data collection was performed using Saphyr 2nd generation instruments (60325) and Instrument Control Software (ICS; v.4.9.19316.1).

Strand-seq data generation and data processing. Strand-seq data were generated at EMBL and the protocol was as follows. EBV-transformed lymphoblastoid cell lines from the 1000 Genomes Project (Coriell Institute; Supplementary Table 1) were cultured in BrdU (100 µM final concentration; Sigma-Aldrich, B9285) for 18 or 24 h, and single isolated nuclei (0.1% NP-40 substitute lysis buffer[47]) were sorted into 96-well plates using the BD FACSMelody and BD Fusion cell sorter. In each sorted plate, 94 single cells plus one 100-cell positive control and one zero-cell negative control were deposited. Strand-specific single-cell DNA sequencing libraries were generated using the previously described Strand-seq protocol[23,47] and automated on the Beckman Coulter Biomek FX P liquid handling robotic system[48]. After 15 rounds of PCR amplification, 288 individually barcoded libraries (amounting to three 96-well plates) were pooled for sequencing on the Illumina NextSeq 500 platform (MID-mode, 75 bp paired-end protocol). The demultiplexed FASTQ files were aligned to the GRCh38 reference assembly (GCA_000001405.15) using BWA aligner (v.0.7.15-0.7.17)[49] for standard library selection. Aligned reads were sorted by genomic position using SAMtools (v.1.10)[50,51] and duplicate reads were marked using sambamba (v.1.0). Low-quality libraries were excluded from future analyses if

they showed low read counts (<50 reads per Mb), uneven coverage or an excess of 'background reads' (reads mapped in opposing orientation for chromosomes expected to inherit only Crick or Watson strands) yielding noisy single-cell data, as previously described[47]. Aligned BAM files were used for inversion discovery as described previously[2].

Hi-C data production. Lymphoblastoid cell lines were obtained from Coriell Cell Repositories and cultured in RPMI 1640 medium supplemented with 15% FBS. Cells were maintained at 37 °C in an atmosphere containing 5% CO_2. Hi-C libraries using 1.5 million human cells as input were generated with Proximo Hi-C kits v4.0 (Phase Genomics) according to the manufacturer's protocol with the following modification: in brief, cells were cross-linked, quenched, lysed sequentially with lysis buffers 1 and 2, and liberated chromatin was immobilized on magnetic recovery beads. A four-enzyme cocktail composed of DpnII (GATC), DdeI (CTNAG), HinfI (GANTC) and MseI (TTAA) was used during the fragmentation step to improve coverage and aid haplotype phasing. After fragmentation and fill-in with biotinylated nucleotides, fragmented chromatin was proximity-ligated for 4 h at 25 °C. Cross-links were then reversed, DNA was purified and biotinylated junctions were recovered using magnetic Streptavidin beads. Bead-bound proximity ligated fragments were then used to generate a dual-unique indexed library compatible with Illumina sequencing chemistry. The Hi-C libraries were evaluated using fluorescent-based assays, including qPCR with the Universal KAPA Library Quantification Kit and Tapestation (Agilent). Sequencing of the libraries was performed at New York Genome Center (NYGC) on the Illumina NovaSeq 6000 instrument using 2 × 150 bp cycles.

RNA-seq data production. Total RNA of cell pellets was isolated using the QIAGEN RNeasy Mini Kit according to the manufacturer's instructions. In brief, each cell pellet (10 million cells) was homogenized and lysed in Buffer RLT Plus, supplemented with 1% β-mercaptoethanol. The lysate containing RNA was purified using the RNeasy spin column, followed by an in-column DNase I treatment by incubating for 10 min at room temperature, and then washed. Finally, total RNA was eluted in 50 µl RNase-free water. RNA-seq libraries were prepared with 300 ng total RNA using KAPA RNA Hyperprep with RiboErase (Roche) according to the manufacturer's instructions. First, ribosomal RNA was depleted using RiboErase. Purified RNA was then fragmented at 85 °C for 6 min, targeting fragments ranging 250–300 bp. Fragmented RNA was reverse-transcribed with an incubation of 25 °C for 10 min, 42 °C for 15 min and an inactivation step at 70 °C for 15 min. This was followed by second-strand synthesis and A-tailing at 16 °C for 30 min and 62 °C for 10 min. The double-stranded cDNA A-tailed fragments were ligated with Illumina unique dual-index adapters. Adapter-ligated cDNA fragments were then purified by washing with AMPure XP beads (Beckman). This was followed by 10 cycles of PCR amplification. The final library was cleaned up using AMPure XP beads. Quantification of libraries was performed using quantitative PCR (qPCR) (Thermo Fisher Scientific). Sequencing was performed on the Illumina NovaSeq platform generating paired-end reads of 100 bp at The Jackson Laboratory for Genomic Medicine.

Iso-seq data production. Iso-seq data were generated at The Jackson Laboratory. Total RNA was extracted from pellets of 10 million human cells. In total, 300 ng total RNA was used to prepare Iso-seq libraries according to Iso-seq Express Template Preparation (Pacbio). First, full-length cDNA was generated using the NEBNext Single Cell/Low Input cDNA synthesis and Amplification Module in combination with the Iso-seq Express Oligo Kit. Amplified cDNA was purified using ProNex beads. The cDNA yield of 160–320 ng then underwent SMRTbell library preparation including DNA damage repair, end repair and A-tailing and was finally ligated with overhang barcoded adapters. Libraries were sequenced on the Pacbio Sequel II system. Iso-seq reads were processed with the default parameters using the PacBio Iso-seq3 pipeline.

Construction and dating of Y phylogeny

The genotypes were jointly called from the 1000 Genomes Project Illumina high-coverage data from a previous study[44] using the approximately 10.4 Mb of chromosome Y sequence previously defined as accessible to short-read sequencing[52]. BCFtools (v.1.9)[50,51] was used with minimum base quality and mapping quality 20, defining ploidy as 1, followed by filtering out SNVs within 5 bp of an indel call (SnpGap) and removal of indels. Furthermore, we filtered for a minimum read depth of 3. If multiple alleles were supported by reads, then the fraction of reads supporting the called allele should be ≥0.85; otherwise, the genotype was converted to missing data. Sites with ≥6% of missing calls, that is, missing in more than 3 out of 44 samples, were removed using VCFtools (v.0.1.16)[53]. After filtering, a total of 10,406,108 sites remained, including 12,880 variant sites. As Illumina short-read data were not available from two samples, HG02486 and HG03471, data from their fathers (HG02484 and HG03469, respectively) was used for Y phylogeny construction and dating.

The Y haplogroups of each sample were predicted as previously described[15] and correspond to the International Society of Genetic Genealogy nomenclature (ISOGG, https://isogg.org, v.15.73, accessed August 2021). We used the coalescence-based method implemented in BEAST (v.1.10.4)[54] to estimate the ages of internal nodes in the Y phylogeny. A starting maximum-likelihood phylogenetic tree for BEAST was constructed with RAxML (v.8.2.10)[55] with the GTRGAMMA substitution model. Markov chain Monte Carlo samples were based on 200 million iterations, logging every 1,000 iterations. The first 10% of iterations were discarded as burn-in. A constant-sized coalescent tree prior, the GTR substitution model, accounting for site heterogeneity (gamma) and a strict clock with a substitution rate of 0.76×10^{-9} (95% CI = 0.67×10^{-9}–0.86×10^{-9}) single-nucleotide mutations per bp per year was used[56]. A prior with a normal distribution based on the 95% CI of the substitution rate was applied. A summary tree was produced using TreeAnnotator (v.1.10.4) and visualized using the FigTree software (v.1.4.4).

The closely related pair of African E1b1a1a1a-CTS8030 lineage Y chromosomes carried by NA19317 and NA19347 differ by 3 SNVs across the 10,406,108 bp region, with the TMRCA estimated to 200 years ago (95% HPD interval = 0–500 years ago).

A separate phylogeny (Fig. 5f) was reconstructed using seven samples (HG01890, HG02666, HG01106, HG02011, T2T Y from NA24385/HG002, HG00358 and HG01952) with contiguously assembled Yq12 region following an identical approach to that described above, with a single difference that sites with any missing genotypes were filtered out. The final callset used for phylogeny construction and split time estimates using Beast contained a total of 10,382,177 sites, including 5,918 variant sites.

De novo assembly generation

Reference assemblies. We used the GRCh38 (NCBI: GCA_000001405.15) and the CHM13 (GCA_009914755.3) plus the T2T Y assembly from GenBank (CP086569.2) released in April 2022. We note that we did not use the unlocalized GRCh38 contig chrY_KI270740v1_random (37,240 bp, composed of 289 *DYZ19* primary repeat units) in any of the analyses presented in this study.

Constructing de novo assemblies. All 28 HGSVC and 15 HPRC samples were processed using the same Snakemake[57] v.6.13.1 workflow (see "Code Availability" statement in main text) to first produce a de novo whole-genome assembly from which selected sequences were extracted in downstream steps of the workflow. The de novo whole-genome assembly was produced using Verkko (v.1.0)[16] with the default parameters, combining all available PacBio HiFi and ONT data per sample to

create a whole-genome assembly: verkko -d work_dir/ --hifi {hifi_reads} --nano {ont_reads}.

The Verkko assembly process includes several steps that lower the base error rate in the resulting assembly. Sequence overlaps among the HiFi reads are leveraged to correct errors, which further increases the accuracy of the HiFi reads before they are used for the initial genome graph construction. The final output sequence is generated by combining the available sequence information to form a consensus. The Verkko assembly process therefore generates highly accurate assemblies[16]. However, there may be a very small number of SNV errors that escape this correction process (which could benefit from additional corrections using Illumina polishing), that may minimally impact downstream SNV-based analyses.

Note that we had to manually modify the assembly FASTA file produced by Verkko for the sample NA19705 for the following reason: at the time of assembly production, the Verkko assembly for the sample NA19705 was affected by a minor bug in Verkko v.1.0 resulting in an empty output sequence for contig 0000598. The Verkko development team suggested removing the affected record, that is, the FASTA header plus the subsequent blank line, because the underlying bug is unlikely to affect the overall quality of the assembly. We followed that advice and continued the analysis with the modified assembly FASTA file. Our discussion with the Verkko development team is publicly documented in the Verkko GitHub issue #66. The assembly FASTA file was adapted as follows: egrep -v "(^$|unassigned\-0000598)" assembly.original.fasta > assembly.fasta.

For the samples with at least 50× HiFi input coverage (termed high-coverage samples; Supplementary Fig. 62 and Supplementary Tables 1 and 2), we generated alternative assemblies using hifiasm (v.0.16.1-r375)[58] for quality-control purposes. Hifiasm was executed using the default parameters using only HiFi reads as input, therefore producing partially phased output assemblies hap1 and hap2 (compare with the hifiasm documentation): hifiasm -o {out_prefix} -t {threads} {hifi_reads}.

The two hifiasm haplotype assemblies per sample are comparable to the Verkko assemblies in that they represent a diploid human genome without further identification of specific chromosomes, that is, the assembled Y sequence contigs have to be identified in a subsequent process that we implemented as follows.

We used a simple rule-based strategy to identify and extract assembled sequences for the two quasi-haploid chromosomes X and Y. The following rules were applied in the order stated here:

Rule 1: the assembled sequence has primary alignments only to the target sequence of interest, that is, to either chromosome Y or chromosome X. The sequence alignments were produced using minimap2 (v.2.24)[59]: minimap2 -t {threads} -x asm20 -Y --secondary=yes -N 1 --cs -c --paf-no-hit.

Rule 2: the assembled sequence has mixed primary alignments, that is, not only to the target sequence of interest, but exhibits Y-chromosome-specific sequence motif hits for any of the following motifs: *DYZ1*, *DYZ18* and the secondary repeat unit of *DYZ3* from ref. 1. The motif hits were identified using HMMER v.3.3.2dev (commit hash #016cba0)[60]: nhmmer --cpu {threads} --dna -o {output_txt} --tblout {output_table} -E 1.60E-150 {query_motif} {assembly}.

Rule 3: the assembled sequence has mixed primary alignments, that is, not only to the target sequence of interest, but exhibits more than 300 hits for the Y-unspecific repeat unit *DYZ2* (see the 'Yq12 *DYZ2* consensus and divergence' section for details on *DYZ2* repeat unit consensus generation). The threshold was determined by expert judgement after evaluating the number of motif hits on other reference chromosomes. The same HMMER call as for rule 2 was used with an *E*-value cut-off of 1.6×10^{-15} and a score threshold of 1,700.

Rule 4: the assembled sequence has no alignment to the chromosome Y reference sequence, but exhibits Y-chromosome-specific motif hits as for rule 2.

Rule 5: the assembled sequence has mixed primary alignments, but more than 90% of the assembled sequence (in bp) has a primary alignment to a single target sequence of interest; this rule was introduced to resolve ambiguous cases of primary alignments to both chromosome X and chromosome Y.

After identification of all assembled chromosome Y and chromosome X sequences, the respective records were extracted from the whole-genome assembly FASTA file and, if necessary, reverse-complemented to be in the same orientation as the T2T reference using custom code.

Assembly evaluation and validation. Error detection in de novo assemblies. According to established procedures[16,20], we implemented three independent approaches to identify regions of putative misassemblies for all 43 samples. First, we used VerityMap (v.2.1.1-alpha-dev #8d241f4)[19], which generates and processes read-to-assembly alignments, to flag regions in the assemblies that exhibit spurious signal, that is, regions of putative assembly errors, but that may also indicate difficulties in the read alignment. Given the higher accuracy of HiFi reads, we executed VerityMap only with HiFi reads as input: python repos/VerityMap/veritymap/main.py --no-reuse --reads {hifi_reads} -t {threads} -d hifi -l SAMPLE-ID -o {out_dir} {assembly_FASTA}.

Second, we used DeepVariant (v.1.3.0)[61] and the PEPPER-Margin-DeepVariant pipeline (v.0.8, DeepVariant v.1.3.0 (ref. 62)) to identify heterozygous SNVs using both HiFi and ONT reads aligned to the de novo assemblies. Given the quasi-haploid nature of the chromosome Y assemblies, we counted all heterozygous SNVs remaining after quality filtering (bcftools v.1.15 "filter" QUAL>=10) as putative assembly errors:

```
/opt/deepvariant/bin/run_deepvariant --model_type="PACBIO" --ref={assembly_FASTA} --num_shards={threads} --reads={HiFi-to-assembly_BAM} --sample_name=SAMPLE-ID --output_vcf={out_vcf} --output_gvcf={out_gvcf} --intermediate_results_dir=$TMPDIR
```

```
run_pepper_margin_deepvariant call_variant --bam {ONT-to-assembly_BAM} --fasta {assembly_FASTA} --output_dir {out_dir} --threads {threads} --ont_r9_guppy5_sup --sample_name SAMPLE-ID --output_prefix {out_prefix} --skip_final_phased_bam --gvcf
```

Third, we used the tool NucFreq (v.0.1)[21] to identify positions in the HiFi read-to-assembly alignments in which the second most common base is supported by at least 10% of the alignments. BAM files were filtered with SAMtools using the flag 2308 (drop secondary and supplementary alignments) following the information in the NucFreq readme. Moreover, we processed only assembled contigs larger than 500 kb to limit the effect of spurious alignments in short contigs. NucFreq was then executed with the default parameters: NucPlot.py --obed OUTPUT.bed --threads {threads} --bed ASSM-CONTIGS.bed HIFI.INPUT.bam OUTPUT.png.

Note that NucFreq could not successfully process the alignments for sample HG00512 due to an error in the graphics output. We therefore omitted this sample from the following processing steps. Again, following the information in the NucFreq readme, we then created flagged regions if more than five positions were flagged in a 500 bp window, and subsequently merged overlapping windows (Supplementary Table 12).

As a final processing step, we merged the VerityMap- and NucFreq-flagged regions (subsuming heterozygous SNVs called by either DeepVariant or PEPPER) by tripling each region's size (flanking region upstream and downstream) and then merging all overlapping regions with bedtools: bedtools merge -c 4 -o collapse -i CONCAT-ALL-REGIONS.bed > OUT.MERGED.bed.

The resulting clusters and all regions separately were post-processed with custom code to derive error estimates for the assemblies

(see 'Code Availability' and 'Data availability' for access to BED files listing all flagged regions/positions and merged clusters; Supplementary Table 10).

As the PAR1 subregion was contiguously assembled from only ten samples (Supplementary Tables 14 and 16), all regions highlighted as putative assembly errors by VerityMap were visually evaluated in the HiFi and ONT read alignments to the assembly using the Integrative Genomics Viewer (IGV, v.2.14.1)[63] (Supplementary Table 11).

Assembly QV estimates were produced with yak v.0.1 (https://github.com/lh3/yak) following the examples in its documentation (see readme in the referenced repository). The QV estimation process requires an independent sequencing data source to derive a (sample-specific) reference *k*-mer set to compare the *k*-mer content of the assembly (Supplementary Fig. 63). In our case, we used available short-read data to create said reference *k*-mer set, which necessitated excluding the samples HG02486 and HG03471 because no short reads were available. For the chromosome-Y-only QV estimation, we restricted the short reads to those with primary alignments to our chromosome Y assemblies or to the T2T Y assembly, which we added during the alignment step to capture reads that would align to chromosome Y sequences missing from our assemblies.

Assembly gap detection. We used the recently introduced tool Rukki (packaged with Verkko v.1.2)[16] to derive estimates of potential gaps in our assemblies. After having identified chromosome Y and chromosome X contigs as described above, we used this information to prepare annotation tables for Rukki to identify chromosome X/chromosome Y paths in the assembly graph: rukki trio -g {assembly_graph} -m {XY_contig_table} -p {out_paths} --final-assign {out_node_assignment} --try-fill-bubbles --min-gap-size 1 --default-gap-size 1000.

The resulting set of paths including gap estimates was summarized using custom code (Supplementary Table 6). See the 'Data availability' section for access to NucFreq plots generated for all samples.

Assembly evaluation using Bionano Genomics optical mapping data. To evaluate the accuracy of Verkko assemblies, all samples ($n = 43$) were first de novo assembled using the raw optical mapping molecule files (bnx), followed by alignment of assembled contigs to the T2T whole-genome reference genome assembly (CHM13 + T2T Y) using Bionano Solve (v.3.5.1) pipelineCL.py.

```
python2.7 Solve3.5.1_01142020/Pipeline/1.0/pipelineCL.py -T 64 -U -j
64 -jp 64 -N 6 -f 0.25 -i 5 -w -c 3 -y -b ${ bnx} -l ${output_dir} -t
Solve3.5.1_01142020/RefAligner/1.0/ -a
Solve3.5.1_01142020/RefAligner/1.0/optArguments_haplotype_DLE1_
saphyr_human.xml -r ${ref}
```

To improve the accuracy of optical mapping Y chromosomal assemblies, unaligned molecules, molecules that align to T2T chromosome Y and molecules that were used for assembling contigs but did not align to any chromosomes were extracted from the optical mapping de novo assembly results. These molecules were used for the following three approaches: (1) local de novo assembly using Verkko assemblies as the reference using pipelineCL.py, as described above; (2) alignment of the molecules to Verkko assemblies using refAligner (Bionano Solve (v.3.5.1)); and (3) hybrid scaffolding using optical mapping de novo assembly consensus maps (cmaps) and Verkko assemblies by hybridScaffold.pl.

```
perl Solve3.5.1_01142020/HybridScaffold/12162019/hybridScaffold.
pl -n ${fastafile} -b ${bionano_cmap} -c
Solve3.5.1_01142020/HybridScaffold/12162019/hybridScaffold_
DLE1_config.xml -r Solve3.5.1_01142020/RefAligner/1.0/RefAligner -o
${output_dir} -f -B 2 -N 2 -x -y -m ${bionano_bnx} -p
Solve3.5.1_01142020/Pipeline/12162019/ -q
Solve3.5.1_01142020/RefAligner/1.0/optArguments_nonhaplotype_
DLE1_saphyr_human.xml
```

Inconsistencies between optical mapping data and Verkko assemblies were identified based on variant calls from approach 1 using the 'exp_refineFinal1_merged_filter_inversions.smap' output file. Variants were filtered out on the basis of the following criteria: (1) variant size smaller than 500 bp; (2) variants labelled as heterozygous; (3) translocations with a confidence score of ≤0.05 and inversions with a confidence score of ≤0.7 (as recommended on Bionano Solve Theory of Operation: Structural Variant Calling - Document Number: 30110); (4) variants with a confidence score of <0.5. Variant reference start and end positions were then used to evaluate the presence of single molecules, which span the entire variant using alignment results from approach 2. Alignments with a confidence score of <30.0 were filtered out. Hybrid scaffolding results, conflict sites provided in the 'conflicts_cut_status.txt' output file from approach 3 were used to evaluate whether inconsistencies identified above based on optical mapping variant calls overlapped with conflict sites (that is, sites identified by hybrid scaffolding pipeline representing inconsistencies between sequencing and optical mapping data) (Supplementary Table 50). Furthermore, we used molecule alignment results to identify coordinate ranges on each Verkko assembly, which had no single DNA molecule coverage using the same alignment confidence score threshold of 30.0, as described above, dividing assemblies into 10 kb bins and counting the number single molecules covering each 10 kb window (Supplementary Table 51).

De novo assembly annotation. Annotation of Y-chromosomal subregions. The 24 Y-chromosomal subregion coordinates (Supplementary Table 13) relative to the GRCh38 reference sequence were obtained from a previous study[8]. As Skov et al.[8] produced their annotation on the basis of a coordinate liftover from GRCh37, we updated some coordinates to be compatible with the following publicly available resources: for the PARs, we used the coordinates from the UCSC Genome Browser for GRCh38.p13 as they slightly differed. Moreover, Y-chromosomal amplicon start and end coordinates were edited according to more recent annotations from a previous study[64], and the locations of *DYZ19* and *DYZ18* repeat arrays were adjusted on the basis of the identification of their locations using HMMER3 (v.3.3.2)[65] with the respective repeat unit consensus sequences from a previous study[1].

The locations and orientations of Y-chromosomal subregions in the T2T Y were determined by mapping the subregion sequences from the GRCh38 Y to the T2T Y using minimap2 (v.2.24, see above). The same approach was used to determine the subregion locations in each de novo assembly with subregion sequences from both GRCh38 and the T2T Y (Supplementary Table 13). The locations of the *DYZ18* and *DYZ19* repeat arrays in each de novo assembly were further confirmed (and coordinates adjusted if necessary) by running HMMER3 (see above) with the respective repeat unit consensus sequences from[1]. Only tandemly organized matches with HMMER3 score thresholds higher than 1,700 for *DYZ18* and 70 for *DYZ19*, respectively, were included and used to report the locations and sizes of these repeat arrays.

A Y-chromosomal subregion was considered to be contiguous if it was assembled contiguously from the subclass on the left to the subclass on the right (note that the *DYZ18* subregion is completely deleted in HG02572), except for PARs in which they were defined as >95% length of the T2T Y PARs and with no unplaced contigs. Note that, due to the requirement of no unplaced contigs, the assembly for HG02666 appears to have a break in PAR2 subregion, while it is contiguously assembled from the telomeric sequence of PAR1 to telomeric sequence in PAR2 without breaks (however, there is a ~14 kb unplaced PAR2 contig aligning best to a central region of PAR2). However, the assembly of HG01890 has a break approximately 100 kb before the end of PAR2. The assembly of PAR1 remains especially challenging owing to its sequence composition and sequencing biases[9,10] and, among our samples, was contiguously assembled for 10 out of 43 samples, whereas PAR2 was contiguously assembled for 39 out of 43 samples.

Annotation of centromeric and pericentromeric regions. To annotate the centromeric regions, we first ran RepeatMasker (v.4.1.0, http://www.repeatmasker.org/) on 26 Y-chromosomal assemblies (22 samples with contiguously assembled pericentromeric regions, 3 samples with a single gap and no unplaced centromeric contigs and the T2T Y) to identify the locations of α-satellite repeats using the following command: RepeatMasker -species human -dir {path_to_directory} -pa {num_of_threads} {path_to_fasta}.

Then, we subsetted each contig to the region containing α-satellite repeats and ran HumAS-HMMER (v.3.3.2; https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) to identify the location of α-satellite HORs, using the following command: Hmmer-run.sh {directory_with_fasta} AS-HORs-hmmer3.0-170921.hmm {num_of_threads}.

We combined the outputs from RepeatMasker (v.4.1.0) and HumAS-HMMER to generate a track that annotates the location of α-satellite HORs and monomeric or diverged α-satellite within each centromeric region.

To determine the size of the α-satellite HOR array (reported for 26 samples in Supplementary Table 16, while the size estimates reported in the main text include 23 gapless assemblies (Supplementary Table 15)), we used the α-satellite HOR annotations generated using HumAS-HMMER (v.3.3.2; described above) to determine the location of *DYZ3* α-satellite HORs, focusing on only those HORs annotated as 'live' (for example, S4CYH1L). Live HORs are those that have a clear higher-order pattern and are highly (>90%) homogenous[66]. This analysis was conducted on 21 centromeres (including the T2T Y; Extended Data Fig. 8a), excluding 5 out of 26 samples (NA19384, HG01457, HG01890, NA19317, NA19331), in which, despite a contiguously assembled pericentromeric subregion, the assembly contained unplaced centromeric contig(s).

To annotate the *HSat3* and *DYZ17* arrays within the pericentromere, we ran StringDecomposer (v.1.0.0) on each assembly centromeric contig using the *HSat3* and *DYZ17* consensus sequences described previously[67] and available at GitHub (https://github.com/altemose/HSatReview/blob/main/Output_Files/HSat123_consensus_sequences.fa). We ran the following command: stringdecomposer/run_decomposer.py {path_to_contig_fasta} {path_to_consensus_sequence+fasta} -t {num_of_threads} -o {output_tsv}.

The *HSat3* array was determined as the region that had a sequence identity of 60% or greater, while the *DYZ17* array was determined as the region that had a sequence identity of 65% or greater.

Downstream analysis

Effect of input read depth on assembly contiguity. We examined a putative dependence between the characteristics of the input read sets, such as read length N50 or genomic coverage, and the resulting assembly contiguity by training multivariate regression models (ElasticNet from scikit-learn v.1.1.1; Code availability). The models were trained according to standard procedures with fivefold nested cross-validation (see the scikit-learn documentation for ElasticNetCV). Note that we did not use the haplogroup information owing to the unbalanced distribution of haplogroups in our dataset. We selected basic characteristics of both the HiFi and ONT-UL input read sets (read length N50, mean read length, genomic coverage and genomic coverage for ONT reads exceeding 100 kb in length, that is, the ultralong fraction of ONT reads; Supplementary Tables 52 and 53) as model features, and assembly contig NG50, assembly length or number of assembled contigs as target variable.

Locations of assembly gaps. The assembled Y-chromosomal contigs were mapped to the GRCh38 and the CHM13 plus T2T Y reference assemblies using minimap2 with the flags -x asm20 -Y -p 0.95 --secondary=yes -N 1 -a -L --MD --eqx. The aligned Y-chromosomal sequences for each reference were partitioned to 1 kb bins to investigate assembly gaps. Gap presence was inferred in bins where the average read depth was either lower or higher than 1. To investigate the potential factors associated with gap presence, we analysed these sequences to compare the GC content, SD content and Y subregion (Supplementary Figs. 64–68). Read depth for each bin was calculated using mosdepth[68] and the flags -n -x. GC content for each bin was calculated using the BedTools nuc function[69]. SD locations for GRCh38 Y were obtained from the UCSC genome browser, and for the CHM13 plus T2T Y from[4]. Y-chromosomal subregion locations were determined as described in the 'De novo assembly annotation with Y-chromosomal subregions' section. The bin read depth and GC content statistics were merged into matrices and visualized using matplotlib and seaborn[70,71].

Comparison of assembled Y subregion sizes across samples. Sizes for each chromosome's (peri)centromeric regions were obtained as described in the 'Annotation of pericentromeric regions' section. The size variation of (peri)centromeric regions (*DYZ3* alpha-satellite array, *Hsat3*, *DYZ17* array and total (peri)centromeric region), and the *DYZ19*, *DYZ18* and *TSPY* repeat arrays were compared across samples using a heat map, incorporating phylogenetic context. The sizes of the (peri)centromeric regions (*DYZ3* alpha-satellite array, *Hsat3* and *DYZ17* array) were regressed against each other using the OLS function in statsmodels, and visualized using matplotlib and seaborn[70].

Comparison and visualization of de novo assemblies. The similarities of three contiguously assembled Y chromosomes (HG00358, HG02666, HG01890), including comparison to both GRCh38 and the T2T Y, was assessed using blastn[72] with a sequence identity threshold of 80% (95% threshold was used for PAR1 subregion) (Fig. 2b) and excluding non-specific alignments (that is, showing alignments between different Y subregions), followed by visualization with genoPlotR (v.0.8.11)[73]. Y subregions were uploaded as DNA segment files and alignment results were uploaded as comparison files following the file format recommended by the developers of the genoplotR package. Unplaced contigs were excluded, and all Y-chromosomal subregions, except for Yq12 heterochromatic region and PAR2, were included inqueries:

```
blastn -query $file1 -subject $file2 -subject_besthit -outfmt '7 qstart
qend sstart send qseqid sseqid pident length mismatch gaps evalue
bitscore sstrand qcovs qcovhsp qlen slen' -out ${outputfile}.out
```

```
plot_gene_map(dna_segs=dnaSegs, comparisons=comparisonFiles,
xlims=xlims, legend = TRUE, gene_type = "headless_arrows", dna_
seg_scale=TRUE, scale=FALSE)
```

For other samples, three-way comparisons were generated between the GRCh38 Y, Verkko de novo assembly and the T2T Y sequences, removing alignments with less than 80% sequence identity. The similarity of closely related NA19317 and NA19347 Y-chromosomal assemblies was assessed using the same approach.

Sequence identity heat maps. Sequence identity within repeat arrays was investigated by running StainedGlass[74]. For the centromeric regions, StainedGlass was run with the following configuration: window = 5785 and mm_f = 30000. We adjusted the colour scale in the resulting plot using a custom R script that redefines the breaks in the histogram and its corresponding colours. This script is publicly available online (https://eichlerlab.gs.washington.edu/help/glogsdon/Shared_with_Pille/StainedGlass_adjustedScale.R). The command used to generate the new plots is as follows: StainedGlass_adjustedScale.R -b {output_bed} -p {plot_prefix}. For the *DYZ19* repeat array, window = 1000 and mm_f = 10000 were used, 5 kb of flanking sequence was included from both sides, followed by adjustment of colour scale using the custom R script (see above).

For the Yq12 subregion (including the *DYZ18* repeat array), window = 5000 and mm_f = 10000 were used, and 10 kb of flanking sequence was included. In addition to samples with contiguously assembled

Yq12 subregion, the plots were generated for two samples (NA19705 and HG01928) with a single gap in Yq12 subregion (the two contigs containing Yqhet sequence were joined into a single contig with 100 Ns added to the joint location). HG01928 contains a single unplaced Yqhet contig (approximately 34 kb in size), which was not included. For the Yq11/Yq12 transition region, 100 kb proximal to the *DYZ18* repeat and 100 kb of the first *DYZ1* repeat array was included in the StainedGlass runs, using window = 2000 and mm_f = 10000.

Dot plot generation. Dot plot visualizations were created using the NAHRwhals package (v.0.9), which provides visualization utilities and a custom pipeline for pairwise sequence alignment based on minimap2 (v.2.24). In brief, NAHRwhals initiates pairwise alignments by splitting long sequences into chunks of 1–10 kb, which are then aligned to the target sequence separately, enhancing the ability of minimap2 to correctly capture inverted or repetitive sequence alignments. Subsequently, alignment pairs are concatenated whenever the end point of one alignment falls in close proximity to the start point of another (base pair distance cut-off: 5% of the chunk length). Pairwise alignment dot plots are created with a pipeline based on the ggplot2 package, with optional .bed files accepted for specifying colorization or gene annotation. The NAHRwhals package and further documentation are available at GitHub (https://github.com/WHops/nahrchainer); see the 'Data availability' for information on accessing the dot plots.

Inversion analyses. Inversion calling using Strand-seq data. The inversion calling from Strand-seq data, available for 30 out of 43 samples and the T2T Y, using both the GRCh38 and the T2T Y sequences as references was performed as described previously[2].

Note that, for the P5 palindrome spacer direction in the T2T Y assembly, the P5 spacer region is present in the same orientation in both GRCh38 (where the spacer orientation had been chosen randomly, see supplementary figure 11 from ref. 1 for more details) and the T2T Y sequence, while high-confidence calls from the Strand-seq data from individual HG002/NA24385 against both the GRCh38 and T2T Y report it to be in inverted orientation. It is therefore likely that the P5 spacer orientations are incorrect in both GRCh38 Y and the T2T Y and in the P5 inversion recurrence estimates we therefore considered HG002/NA24385 to carry the P5 inversion (as shown in Fig. 3a, inverted relative to GRCh38).

Inversion detection from the de novo assemblies. To determine the inversions from the de novo assemblies, we aligned the Y-chromosomal repeat units/SDs as published previously[64] to the de novo assemblies as described above (see the 'Annotation with Y-chromosomal subregions' section). Inverted alignment orientation of the unique sequences flanked by repeat units/SDs relative to the GRCh38 Y was considered as evidence of inversion. The presence of inversions was further confirmed by visual inspection of de novo assembly dot plots generated against both GRCh38 and T2T Y sequences (see the 'Dot plot generation' section), followed by merging with the Strand-seq calls (Supplementary Table 33).

Inversion rate estimation. To estimate the inversion rate, we counted the minimum number of inversion events that would explain the observed genotype patterns in the Y phylogeny (Fig. 3a). A total of 12,880 SNVs called in the set of 44 male individuals and the Y-chromosomal substitution rate from above (see the 'Construction and dating of Y phylogeny' section) was used. A total of 126.4 years per SNV mutation was then calculated $(0.76 \times 10^{-9} \times 10{,}406{,}108\ \text{bp})^{-1}$, and converted into generations assuming a 30-year generation time[75]. Thus, each SNV corresponds to 4.21 generations, translating into a total branch length of 54,287 generations for the 44 samples. For a single inversion event in the phylogeny, this yields a rate of 1.84×10^{-5} (95% CI = 1.62×10^{-5} to 2.08×10^{-5}) mutations per father-to-son Y transmission. The confidence interval of the inversion rate was obtained using the confidence interval of the SNV rate.

Determination of inversion breakpoint ranges. We focused on the following eight recurrent inversions to narrow down the inversion breakpoint locations: IR3/IR3, IR5/IR5, and palindromes P8, P7, P6, P5, P4 and P3 (Fig. 3a), and leveraged the 'phase' information (that is, proximal/distal) of paralogous sequence variants (PSVs) across the SDs mediating the inversions as follows. First, we extracted proximal and distal inverted repeat sequences flanking the identified inversions (spacer region) and aligned them using MAFFT (v.7.487)[76,77] with the default parameters. From the alignment, we selected only informative sites (that is, not identical across all repeats and samples), excluding singletons and removing sites within repetitive or poorly aligned regions as determined by Tandem Repeat Finder (v.4.09.1)[78] and Gblocks (v.0.91b)[79], respectively. We inferred the ancestral state of the inverted regions following the maximum parsimony principle as follows: we counted the number of inversion events that would explain the distribution of inversions in the Y phylogeny by assuming (1) that the reference (that is, same as GRCh38 Y) state was ancestral; and (2) that the inverted (that is, inverted compared with GRCh38 Y) state was ancestral. The definition of ancestral state for each of the regions was defined as the lesser number of events to explain the tree (IR3: reference; IR5: reference; P8: inverted; P7: reference; P5: reference; P4: reference; P3: reference). As we observed a clear bias of inversion state in both African (Y lineages A, B and E) and non-African Y lineages for the P6 palindrome (the African Y lineages have more inverted states (17 out of 21) and non-African Y lineages have more reference states (17 out of 23)), we determined the ancestral state and inversion breakpoints for African and non-African Y lineages separately in the following analyses.

We then defined an ancestral group as any samples showing an ancestral direction in the spacer region, and selected sites that have no overlapping alleles between the proximal and distal alleles in the defined ancestral group, which were defined as the final set of informative PSVs. For IR3, we used the ancestral group as samples with Y-chromosomal structure 1 (that is, with the single ~20.3 kb TSPY repeat located in the proximal IR3 repeat) and ancestral direction in the spacer region. According to the allele information from the PSVs, we determined the phase (proximal or distal) for each PSV across samples. Excluding non-phased PSVs (for example, the same alleles were found in both proximal and distal sequences), any two adjacent PSVs with the same phase were connected as a segment while masking any single PSVs with a different phase from the flanking ones to only retain reliable contiguous segments. An inversion breakpoint was determined to be a range in which phase switching occurred between two segments, and the coordinate was converted to the T2T Y coordinate on the basis of the multiple-sequence alignment (MSA) and to the GRCh38 Y coordinate using the LiftOver tool at the UCSC Genome Browser web page (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Samples with non-contiguous assembly of the repeat regions were excluded from each analysis of the corresponding repeat region.

Molecular evolution of Y-chromosomal palindromes. Alignments of palindromes and the XDRs. To construct alignments of P1, P2, P3 palindromes, and P1 yellow and P3 teal SDs (Supplementary Fig. 34), we first mapped XDR8, blue, teal, green, red, grey, yellow, and other2 sequences derived from the GRCh38 Y reference (using coordinates derived from ref. 64) to each assembly using minimap2 (v.2.24, see above). We then reconstructed the order of non-overlapping SDs by selecting the longest from any overlapping mappings for each SD. The P1, P2, P3, P1 yellow and P3 teal sequences were collected from each sample only if the directionality and origin of two palindromic arms could clearly be determined in the context of inter-palindromic inversion status. For example, sample HG02486 has the GRCh38 Y segments mappings in the order of XDR8, blue, teal, teal, blue, green, red, red, green, yellow, blue, grey, red, red, green, yellow, blue, grey and other2 on the same contig, which matches with the expected SD order given the inversion status of HG02486 (one *gr/rg* inversion).

As we do not know the exact breakpoint locations for the *gr/rg* inversion in HG02486, the green and red SDs could be mixtures of P1 and P2 origin (as both contain green and red SDs and the *gr/rg* inversions breakpoints are likely to be located within these regions) due to the inversion. Thus, to be conservative and to avoid SDs, which are potential mixes from different origins, we excluded the green and red segments from the analysis. Similar filtering was applied to all of the samples to reduce the possibility of including variation that has been caused by, for example, interpalindromic inversions. As a result, we collected 18, 24, 28, 34 and 39 sequences for P1, P2, P3, P1 yellow and P3 teal, respectively (Supplementary Table 15), and aligned them using MAFFT (v.7.487)[76,77] with the default parameters. For the other palindromes (P4, P5, P6, P7 and P8), we used the same alignments from the inversion analysis described above (see the 'Determination of inversion breakpoint ranges' section for details). The eight XDR sequences were collected by mapping the XDR sequences derived from the GRCh38 Y-chromosome reference to each assembly using minimap2 and aligning them using MAFFT (v.7.487)[76,77] with the default parameters. The raw alignments were summarized by averaging frequencies of major allele (including a gap) across 100 bp windows (Supplementary Figs. 36 and 37).

For all of the aligned sequences, we trimmed both ends of each alignment so that the GRCh38 Y coordinate system could be used. To minimize alignment errors and recurrent mutations, we masked sites residing in repetitive or poorly aligned regions as defined by Tandem Repeat Finder (v.4.09.1)[78] and Gblocks (v.0.91b)[79]. Moreover, sites with a gap in any of the samples or with more than two alleles were masked from all samples. Lastly, we manually curated alignments by masking regions with any potential structural rearrangements to consider only point mutations in the following analyses. The final curated alignments contain unmasked regions ranging from 57.63 to 97.40% of raw alignment depending on the region.

Estimation of point mutation rates in palindromes. To estimate the point mutation rates of the Y-chromosomal palindromes, we first selected a set of 13 samples for which, after the stringent filtering and manual curation described above, alignments of all palindromic regions and XDRs were available. We then counted the number of SNVs in the XDR regions, which was used to calibrate the number of generations spanned by the 13 samples according to a previously described approach[8] (and described in more detail below), using the mutation rate estimates of the XDRs (3.14×10^{-8} per position per generation, PPPG) from ref. 80. Mutation events in the palindromes were determined considering the phylogenetic relationships of the 13 samples following a previous approach[8]. Finally, the mutation rate (PPPG) for palindromes was estimated by dividing the number of mutation events by the estimated generations and two times the unmasked palindrome length for each palindrome (Supplementary Table 38). For P1 and P3 palindromes, we analysed regions with and without multicopy (that is, >2 copies) SDs separately.

Detection of gene conversion events in palindromes. The gene conversion analysis was performed for palindromes P1 yellow, P3 teal, P4, P5, P6, P7 and P8 with 34, 39, 36, 33, 43, 44 and 44 samples for the respective palindromes (including the T2T Y; Supplementary Tables 15 and 37). For each position in the alignment, we determined the genotypes for all internal nodes on the basis of their child nodes and assigned gene conversion or mutation events for each node according to the previously described approach[8]. Starting with all observed genotypes in the tree, we filled out genotypes of all ancestral nodes based on their child nodes. We then determined gene conversion or mutation events if the genotype of the parent was different from that of the child(ren). In case multiple scenarios could explain the phylogenetic tree and the observed genotypes at a particular position, the one with the lowest number of mutations was selected. Positions, for which the best scenario included more than one mutation, were excluded from this analysis. The bias towards the ancestral state or GC bases was statistically tested using the χ^2 test.

Note that gene conversion events towards the ancestral state might be underestimated compared with the actual number of events as it is not possible to detect a gene conversion event towards the ancestral state in case it had occurred on the same branch where the mutation generating the paralogous sequence variant took place. To adjust for this bias in the detection of 'to ancestral' and 'to derived' gene conversion events, we changed the derived homozygous genotypes to ancestral homozygous genotypes in all gene conversion events detected in the initial gene conversion analysis. Using the modified genotypes, we then determined the gene conversion events using the same approach of the initial analysis, and recalculated the ancestral bias by discarding the gene conversion events that were not identified in the new analysis.

Variant calling. Variant calling using de novo assemblies. Variants were called from assembly contigs using PAV (v.2.1.0)[20] with the default parameters using minimap2 (v.2.17) contig alignments to GRCh38 (primary assembly only; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/reference/20200513_hg38_NoALT/). Supporting variant calls were done against the same reference with PAV (v.2.1.0) using LRA[81] alignments (commit e20e67) with assemblies, PBSV (v.2.8.0) (https://github.com/PacificBiosciences/pbsv) with PacBio HiFi reads, SVIM-asm (v.1.0.2)[82] with assemblies, Sniffles (v.2.0.7)[83] with PacBio HiFi and ONT, DeepVariant (v.1.1.0)[17,61] with PacBio HiFi, Clair3 (v.0.1.12)[84] with ONT, CuteSV (v.2.0.1)[85] with ONT, and longshot (v.0.4.5)[86] with ONT. A validation approach based on the subseq command was used to search for raw-read support in PacBio HiFi and ONT[20].

A merged callset was created from the PAV calls with minimap2 alignments across all samples with SV-Pop (v.3.3.5)[20,87] to create a single non-redundant callset. We used merging parameters nr::exact:ro(0.5):szro(0.5,200) for SV and indel insertions and deletions (exact size and position, then 50% reciprocal overlap, then 50% overlap by size and within 200 bp), nr::exact:ro(0.2) for inversions (exact size and position, then 20% reciprocal overlap) and nrsnv::exact for SNVs (exact position and REF/ALT match). The PAV minimap2 callset was intersected with each orthogonal support source using the same merging parameters. SVs were accepted into the final callset if they had support from two orthogonal sources with at least one being another caller (that is, support from only subseq PacBio HiFi and subseq ONT was not allowed). Indels and SNVs were accepted with support from one orthogonal caller. Inversions were manually curated using dot plots and density plots generated by PAV. A chromosome X callset was constructed from chromosome X assemblies, and quality control was applied using the same callers and parameters for the chromosome Y callset, which was subsequently used for variant density estimations in PAR1.

An increase in SVs near contig ends can indicate errors, which we did not see evidence for with a minimum distance to a contig end of 6.9 kb for SV insertions and 198 kb for SV deletions. All SVs were anchored more than 1 kb into contig ends except for a single deletion in HG00673 (624 bp), and an average of 1.53 insertions and 0.77 deletions were anchored less than 10 kb from contig ends.

To search for probable duplications within insertion calls, insertion sequences were re-mapped to the reference with minimap2 (v.2.17) with the parameters -x asm20 -H --secondary=no -r 2k -Y -a --eqx -L -t 4.

To characterize variant densities in PAR1, MSY and chromosome X against autosomes, we split variants into four callsets: (1) MSY; (2) autosomes; (3) chromosome X outside PARs; and (4) PAR1 from both chromosome X and chromosome Y (post-quality-control PAV calls). The MSY and PAR1 callsets were derived from this study, and the autosomes and chromosome X was derived from ref. 20. For chromosome X, we included only female samples to avoid technical biases in the analysis. We excluded tandem repeats and SDs to avoid overwhelming our signal by higher mutability rates and potential technical biases within these regions. For variant call rates, we computed the variants per Mb, eliminating any uncallable and repetitive loci from

the denominator, which includes assembly gaps, centromeres and difficult to align regions around them (low-confidence filter published with ref. 20), tandem repeats, SDs and any loci that were not contiguously mappable for variant calling as flagged by PAV. For all statistics, the choice between Student's and Welch's *t*-test was made by an F-test with a *P*-value cut-off of 0.01.

Validation of large SVs using optical mapping data. Orthogonal support for merged PAV calls was evaluated using optical mapping data (Supplementary Table 54). Molecule support was evaluated using local de novo assembly maps, which aligned to GRCh38 reference assembly. This evaluation included all 29 SVs 5 kb or larger in size (including 15 insertions and 14 deletions; Supplementary Table 22). Although variants of less than 5 kb could be resolved using optical mapping techniques, there were loci without any fluorescent labels, which could lead to misinterpretation of the results. Variant reference (GRCh38) start and end positions were used to evaluate the presence of single molecules, which span the variant breakpoints using alignment results using Bionano Access (v.1.7). Alignments with a confidence score of <30.0 were filtered out.

TSPY repeat array copy-number analysis. To perform a detailed analysis of the TSPY repeat array, which is known to be highly variable in copy number[88], the consensus sequence of the repeat unit was first constructed as follows. The repeat units were determined from the T2T Y sequence, the individual repeat unit sequences extracted and aligned using MAFFT (v.7.487)[76,77] with the default parameters. A consensus sequence was generated using EMBOSS cons (v.6.6.0.0) command line version with the default parameters, followed by manual editing to replace sites defined as Ns with the major allele across the repeat units. The constructed TSPY repeat unit consensus sequence was 20,284 bp.

The consensus sequence was used to identify TSPY repeat units from each de novo assembly using HMMER3 (v.3.3.2)[65], excluding five samples (HG03065, NA19239, HG01258, HG00096, HG03456) with non-contiguous assembly of this region. TSPY repeat units from each assembly were aligned using MAFFT as described above, followed by running HMMER functions esl-alistat and esl-alipid to obtain sequence identity statistics (Supplementary Table 18).

Dot plots of the TSPY repeat array were generated using EMBOSS dotpath (v.6.6.0.0) command line version with the default parameters, with varying window sizes (2, 5 and 10 kb). The TSPY repeat array locations as reported in Supplementary Table 17 were used, with 5 kb of flanking sequence added to both sides.

Phylogenetic analysis of *TSPY*, *RBMY1* and *DAZ* ampliconic genes. We used Ensembl[89] to retrieve the exon sequences for the *TSPY* (*TSPY1*), *RBMY1* (*RBMY1B*) and *DAZ* (*DAZ1*) ampliconic genes. The copy numbers of *RBMY1* were also compared to previous reports[24] (Supplementary Table 55). The sequences of these exons were incorporated with the curated Dfam library[90] into a custom RepeatMasker library. A local RepeatMasker installation was then used to screen all (43 assemblies, T2T Y and GRCh38) Y chromosome assemblies for exon sequence hits. For each assembly, we identified all low-divergence exon hits below 2% divergence. Individual *TSPY* and *RBMY1* gene copies were counted as protein coding if all gene exons (6 exons for *TSPY* and 12 exons for *RBMY1*) were present and classified as low divergence. Whereas, for *DAZ*, there was a high variation in exon copy number across individual *DAZ* gene copies. Thus, an individual *DAZ* gene copy was defined as simply all low divergence exons that are present in the same orientation within a *DAZ* gene cluster. After retrieving the individual exon sequences using SAMtools[50], the exon sequences of each individual gene copy were 'fused' together using Python. Assemblies containing breaks in contigs within the *TSPY* array (excluding GRCh38) or *DAZ* gene clusters were dropped from all subsequent analyses. For each gene family, we generated an MSA (one alignment per ampliconic gene family) using MUSCLE[91] and manually curated the alignment if needed. MSA of *DAZ* sequences was deemed to be unfruitful due to the large variability in gene exon copy numbers. Subsequently, the *TSPY* and *RBMY1* MSAs were given to IQ-Tree[92] (web server, v.1.6.12) for phylogenetic analysis (see the 'Phylogenetic analysis of *DYZ2* consensus sequences' section for details).

We also performed a network analysis for *TSPY* and *RBMY1* to identify clusters of sequences. First, we constructed a gene copy sequence distance matrix by computing the hamming distance between each sequence. We next built an undirected weighted network, using NetworkX (v.2.8.4)[93], in which each gene copy sequence is represented as a node and a weighted edge/link is placed between the nodes that are the most similar to each other (that is, lowest hamming distance). After network construction, we used an asynchronous label propagation algorithm[94] (as implemented in NetworkX to identify gene sequence communities (that is, clusters/subgraphs/subnetworks) within the network. Within each community, the node with the most connections (that is, the hub node) was identified and their sequence was considered to be the best representation of the community. If a network community was comprised of less than five nodes (*TSPY*) or three nodes (*RBMY1*), the sequence of each node within said community was compared to that of all hub nodes and then reassigned to the community of the hub node they most closely resembled (that is, least hamming distance). Separate community size cut-offs were used due to the large difference in total sequences (nodes) within each ampliconic gene network (*TSPY*, 1,344 sequences; *RBMY1*, 353 sequences). Next, we created a community consensus sequence of each community. This was constructed from the MSA of the sequences of all gene copies within a community using MUSCLE and applying a majority rule approach to build the consensus sequence. The sequence of every node within a community was then compared with their community's consensus sequence. The gene copy community assignments were projected onto the gene phylogenetic tree to better understand the evolutionary relationships between network communities.

MEI analysis. MEI calling. We leveraged an enhanced version of PALMER (Pre-Masking Long reads for Mobile Element Insertion, v.2.0.0)[95] to detect mobile element insertions (MEIs). Reference-aligned (to both GRCh38 Y and T2T Y) Y contigs from Verkko assembly were used as input. PALMER identified putative non-reference insertions (that is, L1, SVA or *Alu* elements) using a pre-masking module based on a library of mobile element sequences. PALMER then identifies the hallmarks of retrotransposition events induced by target-primed reverse transcription, including target site duplication (TSD) motifs, transductions and poly(A) tract sequences. Further manual inspection was carried out on the basis of the information of large inversions, SVs, heterochromatic regions, concordance with the Y phylogeny and alignment of flanking sequences. Low-confidence calls overlapping with large SVs, discordant with the Y phylogeny or observing multiple matches to the flanking sequences were excluded, and high-confidence calls were annotated with further genomic content details.

To compare the ratios of non-reference MEIs from the Y chromosome to the rest of the genome, the following approach was used. The size of the GRCh38 Y reference of 57.2 Mb was used, while the total GRCh38 reference sequence length is 3.2 Gb. At the whole-genome level, this results in a ratio for non-reference *Alu* elements of 0.459 per Mb (1,470 per 3.2 Gb) and for non-reference LINE-1 of 0.066 per Mb (210 per 3.2 Gb)[20]. In chromosome Y, the ratio for non-reference *Alu* elements and LINE-1 is 0.262 per Mb (15 per 57.2 Mb) and 0.105 per Mb (6 per 57.2 Mb), respectively. The ratios within the MEI category were compared using the χ^2 test.

Gene annotation. Liftoff. Genome annotations of chromosome Y assemblies were obtained using T2T Y and GRCh38 Y gff annotation files using liftoff[96]: liftoff -db $dbfile -o $outputfile.gff -u $outputfile.unmapped -dir $outputdir -p 8 -m $minimap2dir -sc 0.85 -copies $fastafile -cds $refassembly.

To evaluate which of the GRCh38 Y protein-coding genes were not detected in Verkko assemblies, we selected genes that were labelled as 'protein_coding' from the GENCODEv41 annotation file (that is, a total of 63 protein-coding genes) We compared protein-coding genes' open reading frames (ORFs) with the ORFs obtained from Ensembl 109 to check whether any pseudogenes were miscategorized. First, exon sequence coordinates were collected from liftoff results. Then, transcripts with the highest sequence identity were selected and used for evaluating ORFs. Concatenated exon fasta sequences were uploaded to AliView (v.1.28)[97]. Exon sequences were aligned using the 'Realign everything' option and sequences were translated using a built-in tool.

Methylation analysis. Read-level CpG DNAme likelihood ratios were estimated using nanopolish v.0.11.1. Nanopolish (https://github.com/jts/nanopolish) was run on the alignment to GRCh38 for all samples including the two Genome in a bottle samples[98] (28 HGSVC, 15 HPRC and 2 GIAB (NA24385/HG002 and NA24149/HG003), totalling 45 samples before quality control), for the three complete assemblies (HG00358, HG01890, HG02666), we also mapped the reads back to the assembled Y chromosomes and performed a separate nanopolish run. On the basis of the GRCh38 mappings, we first performed sample quality control. We found four samples (NA19331, NA19347, HG03009 and HG03065) with genome wide methylation levels of below 50% that were removed by quality control. Using information on the multiple runs on some samples, we observed a high degree of concordance between multiple runs from the same donor, average difference between the replicates over the segments of 0.01 (0–0.15) in methylation beta space.

After quality control, we used pycoMeth to de novo identify interesting methylation segments on chromosome Y. pycoMeth (v.2.2)[99] Meth_Seg is a Bayesian changepoint-detection algorithm that determines regions with a consistent methylation rate from the read-level methylation predictions. Over the 139 quality-controlled flow cells of the 41 samples, we found 2,861 segments that behave consistently in terms of methylation variation in a sample. After segmentation, we derived methylation rates per segment per sample by binarizing methylation calls thresholded at absolute log-likelihood ratio of 2.

To test for methylation effects of haplogroups, we first used the PERMANOVA test, implemented in the R package vegan (v.2.6-4, https://github.com/vegandevs/vegan; https://www.r-project.org/), to identify the impact of the 'aggregated' haplotype group on the DNAme levels over the segments. Owing to the low sample numbers per haplotype group, we aggregated haplogroups to meta groups on the basis of the genomic distance and sample size. We aggregated A, B and C to 'ABC', G and H to 'GH', N and O to 'NO', and Q and R to 'QR'. The E haplogroup and J haplogroup were kept as individual units for our analyses. In the analysis, we corrected for sequencing centre and global DNAme levels. Next, we assessed the link between chromosome Y assembly length and global DNAme levels, either genome wide or on chromosome Y, using PERMANOVA and linear models. We extended the previous PERMANOVA model by adding chromosome Y assembly length as an extra explanatory variable. The linear model was built up like the PERMANOVA test, correcting for the sequencing centre and haplogroup. We also tested individual segments for differential meta-haplogroup DNAme using Kruskal–Wallis tests. Regions with FDR ≤ 0.2, as derived from the Benjamini–Hochberg procedure, are reported as differentially methylated regions (DMRs). For follow up tests on the regions that were found to be significantly different on the basis of the Kruskal–Wallis test, we used a one versus all strategy using a Mann–Whitney *U*-test.

Next to assess the effects of haplogroup and chromosome Y length, we also tested for local methylation quantitative trait loci (*cis*-meQTL) using limix-QTL[100,101]. Specifically, we tested the impact of the genetic variants called on GRCh38 (see the 'Variant calling using de novo assemblies' section) versus the DNAme levels in the 2,861 segments discovered by pycoMeth. For this, we used a linear mixed model implemented in limix-QTL, methylation levels were arcsin-transformed and we used population as a random effect term. Variants with a minor allele frequency of >10% and a call rate >90% were selected for meQTL testing (leaving 11,226 variants). For each DNAme segment, we tested variants within the segment or within 100,000 bases around it, yielding a total of 245,131 tests. Using 1,000 permutations we determined the number of independent tests per segment and *P* values were corrected for this estimated number of tests using the Bonferroni procedure. To account for the number of tested segments we used the Benjamini–Hochberg procedure over the top variants per segment to correct for this.

Expression analysis. Gene expression quantification for the HGSVC[20] and the Geuvadis dataset[26] was derived from a previous study[20]. In brief, RNA-seq quality control was conducted using Trim Galore! (v.0.6.5) (https://github.com/FelixKrueger/TrimGalore) and reads were mapped to the GRCh38 reference using STAR (v.2.7.5a)[102], followed by gene expression quantification using FeatureCounts (v.2)[103]. After quality control, gene expression data are available for 210 Geuvadis male and 21 HGSVC male individuals.

As with the DNAme analysis, we used the PERMANOVA test to quantify the overall impact of haplogroup on gene expression variation. Here we focused only on the Geuvadis samples initially and tested for the effect of the single character haplotype groups, specifically E, G, I, J, N, R and T. Moreover, we tested for single gene effects using the Kruskal–Wallis test and the Mann–Whitney *U*-test. For *BCORP1*, we used the HGSVC expression data to assess the link between DNAme and expression variation.

Iso-seq data analysis. Iso-seq reads were aligned independently using minimap (v.2.24; -ax splice:hq -f1000) to each chromosome Y Verkko assembly, as well as the T2T v.2.0 reference including HG002 chromosome Y and GRCh38. Read alignments were compared between the HG002-T2T chromosome Y reference and each de novo Verkko chromosome Y assembly (Supplementary Figs. 69 and 70 and Supplementary Tables 56 and 57). Existing testis Iso-seq data from seven individuals were also analysed (SRX9033926 and SRX9033927).

Hi-C data analysis. We analysed 40 out of 43 samples for which Hi-C data were available (Hi-C data are missing for HG00358, HG01890 and NA19705) (Supplementary Fig. 71). For each sample, the GRCh38 reference genome was used to map the raw reads and Juicer software tools (version 1.6)[104] with the default aligner BWA mem (v.0.7.17)[49] was used to preprocess and map the reads. Read pairs with low mapping quality (MAPQ < 30) were filtered out, and unmapped reads, such as abnormal split reads and duplicate reads, were also removed. Using these filtered read pairs, Juicer was then used to create a Hi-C contact map for each sample. To leverage the collected chromosome Y Hi-C data from these 40 samples with various resolutions (Supplementary Fig. 72a), we combined the chromosome Y Hi-C contact maps of these 40 samples using the mega.sh script[104] given by Juicer to produce a 'mega' map. Knigh–Ruiz matrix balancing was applied to normalize Hi-C contact frequency matrices[105].

We then calculated insulation score[106], which was initially developed to find TAD boundaries on Hi-C data with a relatively low resolution, to call TAD boundaries at 10 kb resolution for the merged sample and each individual sample (Supplementary Fig. 72b). For the merged sample, the FAN-C toolkit (v.0.9.23b4)[107] with the default parameters was applied to calculate the insulation score and boundary score on the basis of the Knigh–Ruiz normalized mega map at 10 kb resolution and 100 kb window size (using the same setting as in the 4DN domain calling protocol)[108]. For each individual sample, the Knigh–Ruiz-normalized contact matrix of each sample served as the input to the same procedure as in analysing the merged sample. The previous merged result was

treated as a catalogue of TAD boundaries in lymphoblastoid cell lines (LCLs) for chromosome Y to finalize the location of TAD boundaries and TADs of each individual sample. More specifically, 25 kb flanking regions were added on both sides of the merged TAD boundary locations. Any sample boundary located within the merged boundary with the added flanking region was considered to be the final TAD boundary. The final TAD regions were then derived from the two adjacent TAD boundaries excluding those regions where more than half the length of the regions have NA insulation score values (Supplementary Table 58).

The average and variance (maximum difference between any of the two samples) insulation scores of our 40 chromosome Y samples were calculated to show the differences among these samples and were plotted aligned with the methylation analysis and the chromosome Y assembly together. Owing to the limited Hi-C sequencing depth and resolution, some of the chromosome Y regions have the missing reads and those regions with 'NA' insulation scores were shown as blank regions in the plot. Kruskal–Wallis (one-way ANOVA) tests (SciPy v.1.7.3 scipy.stats.kruskal) were performed on the insulation scores (10 kb resolution) of each sample with the same six meta haplogroups classified in the methylation analysis to detect any associations between differentially insulated regions and DMR (Supplementary Table 59). Within each DMR, *P* values were adjusted and those insulated regions with FDR ≤ 0.20 were defined as the regions that are significantly differentially insulated and methylated.

Yq12 subregion analyses. Yq12 partitioning. RepeatMasker (v.4.1.0) was run using the default Dfam library to identify and classify repeat elements within the sequence of the Yq12 region[92]. The RepeatMasker output was parsed to determine the repeat organization and any recurring repeat patterns. A custom Python script that capitalized on the patterns of repetitive elements, as well as the sequence length between *Alu* elements, was used to identify individual *DYZ2* repeats, as well as the start and end boundaries for each *DYZ1* and *DYZ2* array.

Yq12 *DYZ2* consensus and divergence. The two assemblies with the longest (T2T Y from HG002) and shortest (HG01890) Yq12 subregions were selected for *DYZ2* repeat consensus sequence building. Among all *DYZ2* repeats identified within the Yq12 subregion, most (sample collective mean, 46.8%) were exactly 2,413 bp in length. Thus, 500 *DYZ2* repeats 2,413 bp in length were randomly selected from each assembly, and their sequences were retrieved using Pysam (v.0.19.1)[109] (https://github.com/pysam-developers/pysam). Next, a multiple sequence alignment of these 500 sequences was performed using MUSCLE (v.5.1)[91]. On the basis of the MSA, a *DYZ2* consensus sequence was constructed using a majority rule approach. Alignment of the two 2,413 bp consensus sequences, built from both assemblies, confirmed 100% sequence identity between the two consensus sequences. Further analysis of the *DYZ2* repeat regions revealed the absence of a seven-nucleotide segment (ACATACG) at the intersection of the *DYZ2* HSATI and the adjacent *DYZ2* AT-rich simple repeat sequence. To address this, ten nucleotides downstream of the HSATI sequence of all *DYZ2* repeat units were retrieved, an MSA was performed using MUSCLE (v.5.1)[91] and a consensus sequence was constructed using a majority rule approach. The resulting consensus was then fused to the 2,413 bp consensus sequence creating a final 2,420 bp *DYZ2* consensus sequence. *DYZ2* arrays were then rescreened using HMMER (v.3.3.2) and the 2,420 bp *DYZ2* consensus sequence (see Supplementary Fig. 73 for the identified *DYZ2* repeat copy numbers in phylogenetically closely related NA19317 and NA19347 samples).

In view of the AT-rich simple repeat portion of *DYZ2* being highly variable in length, only the *Alu* and HSATI portion of the *DYZ2* consensus sequence was used as part of a custom RepeatMasker library to determine the divergence of each *DYZ2* repeat sequence within the Yq12 subregion. Divergence was defined as the percentage of substitutions in the sequence matching region compared to the consensus. The *DYZ2* arrays were then visualized using a custom Turtle (https://docs.python.org/3.5/library/turtle.html#turtle.textinput) script written in Python. To compare the compositional similarity between *DYZ2* arrays within a genome, a *DYZ2* array (rows) by *DYZ2* repeat composition profile (columns; *DYZ2* repeat length + orientation + divergence) matrix was constructed. Next, the SciPy (v.1.8.1) library was used to calculate the Bray–Curtis distance/dissimilarity (as implemented in scipy.spatial.distance.braycurtis) between *DYZ2* array composition profiles[110]. The complement of the Bray–Curtis dissimilarity was used in the visualization as, typically, a Bray–Curtis dissimilarity closer to zero implies that the two compositions are more similar (Fig. 5e and Supplementary Fig. 60).

Yq12 *DYZ1* array analysis. Initially, RepeatMasker (v.4.1.0) was used to annotate all repeats within *DYZ1* arrays. However, consecutive RepeatMasker runs resulted in variable annotations. These variable results were also observed using a custom RepeatMasker library approach with inclusion of the existing available *DYZ1* consensus sequence[1]. In light of these findings, *DYZ1* array sequences were extracted using Pysam, and each sequence then underwent a virtual restriction digestion with HaeIII using the Sequence Manipulation Suite (v.2)[111] (Supplementary Fig. 74). HaeIII, which has a 'ggcc' restriction cut site, was chosen on the basis of previous research of the *DYZ1* repeat in monozygotic twins[112]. The resulting restriction fragment sequences were oriented based on the sequence orientation of satellite sequences within them detected by RepeatMasker (base Dfam library). A new *DYZ1* consensus sequence was constructed by retrieving the sequence of digestion fragments 3,569 bp in length (as fragments of this length were in the greatest abundance in 6 out of 7 analysed genomes), performing a MSA using MUSCLE (v.5.1) and then applying a majority-rule approach to construct the consensus sequence.

To classify the composition of all restriction fragments, a *k*-mer profile analysis was performed. First, the relative abundance of *k*-mers within fragments as well as consensus sequences (*DYZ18*, 3.1 kb, 2.7 kb, *DYZ1*) were computed. A *k*-mer of length 5 was chosen as *DYZ1* is probably ancestrally derived from a pentanucleotide[5,113]. Next, the Bray–Curtis dissimilarity between *k*-mer abundance profiles of each fragment and consensus sequence was computed, and fragments were classified on the basis of their similarity to the consensus sequence *k*-mer profile (using a 75% similarity minimum) (Supplementary Fig. 51). Afterwards, the sequence fragments with the same classification adjacent to one another were concatenated, and the fully assembled sequence was provided to HMMER (v.3.3.2) to detect repeats and partition fragment sequences into individual repeat units[65]. The HMMER output was filtered by *E*-value (only *E*-value of zero was retained). Once individual repeat units (*DYZ18*, 3.1 kb, 2.7 kb and *DYZ1*) were characterized (Supplementary Fig. 52), the Bray–Curtis dissimilarity of their sequence *k*-mer profile versus the consensus sequence was computed and then visualized with the custom Turtle script written in Python (Extended Data Fig. 9). A two-sided Mann–Whitney *U*-test (SciPy v.1.7.3 scipy.stats.mannwhitneyu[110]) was used to test for differences in length between *DYZ1* and *DYZ2* arrays for each sample with a completely assembled Yq12 region ($n = 7$) (T2T Y HG002, MWU = 541.0, $P = 0.000786$; HG02011, MWU = 169.0, $P = 0.000167$; HG01106, MWU = 617.0, $P = 0.038162$; HG01952, MWU = 172.0, $P = 0.042480$; HG01890, MWU = 51.0, $P = 0.000867$; HG02666, MWU = 144.0, $P = 0.007497$; HG00358, MWU = 497.0, $P = 0.008068$) (Fig. 5b). A two-sided Spearman rank-order correlation coefficient (SciPy v.1.7.3 scipy.stats.spearmanr[110]) was calculated using all samples with a completely assembled Yq12 ($n = 7$) to measure the relationship between the total length of the analysed Yq12 region and the total *DYZ1* plus *DYZ2* arrays within this region (correlation = 0.90093, $P = 0.005620$) (Supplementary Fig. 58). All statistical tests performed were considered significant using an $\alpha = 0.05$.

Phylogenetic analysis of *DYZ2* consensus sequences. We retrieved the *DYZ2* repeat sequence, which was previously identified on all other human acrocentric chromosomes (13, 14, 15, 21 and 22)[114], for our phylogenetic analyses from CHM13 using BLAST[72]. More specifically,

we queried the T2T-CHM13v1.1 reference genome[115] using our *DYZ2* consensus sequence and retrieved all matches with an *e*-value of zero and greater than 500 nucleotides in length using SAMtools[53]. We next performed an MSA, for each chromosome separately, using MUSCLE[91]. We then generated a chromosome-specific consensus sequence using a majority-rule approach. To reflect sequence variation within the Yq12 heterochromatin region, we also constructed two Y chromosome *DYZ2* consensus sequences. One Yq12 *DYZ2* consensus sequence was built from *DYZ2* repeat sequences originating from arrays outside of the Yq12 inversions (that is, end arrays). The second consensus sequence was constructed from *DYZ2* sequences located within arrays internal to the Yq12 inversions (that is, middle arrays). Next, we performed an MSA of all *DYZ2* consensus sequences using MUSCLE. We elected to use only the HSATI and *Alu* portions of the *DYZ2* consensus sequences as the AT-rich simple repeat region was highly variable across consensus sequences. Next, a phylogenetic tree was inferred using maximum likelihood from the MSA with IQ-Tree (web server, v.1.6.12) (a GTR + gamma model was used, unless indicated)[116]. IQ-Tree was ran using Ultrafast boostrap (UFBoot[117]) approximation (1,000 iterations) from which an unrooted maximum-likelihood tree was generated. The phylogenetic tree was then rooted at the midpoint and visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Yq12 MEI analysis. The RepeatMasker output was screened for the presence of additional transposable elements, in particular MEIs. Putative MEIs (that is, elements with a divergence <4%) plus the flanking 100 bp were retrieved from the respective assemblies. After an MSA using MUSCLE, the ancestral sequence of the MEI was determined and used for all downstream analyses (this step was necessary as some of the MEIs duplicated multiple times and contained substitutions). The divergence, and subfamily affiliation, were determined on the basis of the MEI with the lowest divergence from the respective consensus sequence. All MEIs were screened for the presence of characteristics of target-primed reverse transcription hallmarks (that is, the presence of an A-tail, target-site duplications and endonuclease cleavage site)[118].

Repeat annotation. Application of T2T-CHM13-derived repeat annotation pipeline on 43 assembled Y chromosomes. Repeat discovery and annotation were performed on the T2T Y[10] following the pipeline described previously[119]. Subsequently, this repeat-annotation compilation pipeline was applied to the 43 assembled Y chromosomes presented here for annotation using RepeatMasker (v.4.1.2-p1) of all known repeats using the Dfam3.3 library[92] as well as CHM13 and T2T Y derived repeat models (noted previously[10,119]). Repeat annotations were summarized for all 44 assembled chromosomes at the repeat class level using the RepeatMasker script buildSummary.pl[92] with a corresponding genome file denoting the size of the assembly (in bp) and reported in Supplementary Table 31.

Regional repeat assessment of four fully assembled Y chromosomes (including T2T Y). Four Y chromosomes contiguously assembled from PAR1 to PAR2 (HG01890, HG02666, HG00358 and the T2T Y; Supplementary Table 9) provided an opportunity to compare repeat variation within and between Y-chromosomal subregions. Therefore, Y-chromosomal subregions were extracted from the RepeatMasker compilation output (containing known and new repeat models) and summarized using buildSummary.pl[93] with a corresponding genome file denoting the length of the region to be summarized. Repeat classes were summarized per region on the basis of base pair composition, rather than counts, and similar regions were combined (for example, PAR1 + PAR2 = PARs) and presented in Supplementary Fig. 30 and reported in Supplementary Table 30. PAR1 pairwise comparison and satellite-size variation results are reported in Supplementary Tables 60 and 61.

BLAST estimates of *DAZ* and *RBMY1* composite repeat copy number. BLAST custom databases were generated from all Y assemblies and were used to detect instances of the *DAZ* and *RBMY1* composite repeat units per assembly. The consensus sequences for these two composite repeat units were derived from the T2T Y and are reported in Supplementary Table 44 and in ref. 10. The *RBMY1* composite repeat unit contains the whole gene, whereas that of *DAZ* lies within the gene. Owing to the fact that composite repeats are composed of three or more repeating sequences (that is, TEs, satellites, composite subunits, simple/low complexity repeats) as defined previously[119], which are scattered throughout the genome, we required at least an 85% length match to detect predominantly full-length copies while still allowing for variation in the ends. While this requirement for length matching prevents the detection of individual repeats within the composite from being counted as a composite, it does have the limitation of not detecting a full-length copy, as polymorphic transposable element insertions may interfere. Copy-number estimation results for all 44 Y chromosomes are reported in Supplementary Tables 44 and 45.

Statistical analysis and plotting

All statistical analyses in this study were performed using R (http://CRAN.R-project.org/) and Python (http://www.python.org). The respective test details, such as program or library version, sample size, resulting statistics and *P* values, are stated in the text. Figures were generated using R and Python's Matplotlib (https://matplotlib.org), seaborn[70] and the Turtle graphics framework (https://docs.python.org/3/library/turtle.html).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data generated and used in this study were derived from lymphoblast lines available from the Coriell Institute for Medical Research for research purposes (https://www.coriell.org/). Further details are provided in Supplementary Table 1. All data generated by the HGSVC (PacBio HiFi, ONT-UL, Hi-C, RNA-seq, Iso-seq and Bionano Genomics optical genome maps) are available at the International Nucleotide Sequence Database Collaboration (INSDC) under the following NCBI project IDs: PRJEB58376, PRJNA988114, PRJEB41077, PRJEB39684 and PRJEB39750. The HPRC (https://humanpangenome.org/) year 1 PacBio HiFi, ONT long-read sequencing and Bionano Genomics optical mapping data files are available at INSDC (PRJNA701308). Further details are provided in Supplementary Table 1. Existing testis Iso-seq data from seven individuals are available from the European Nucleotide Archive (ENA) under accessions SRX9033926 and SRX9033927. The GEUVADIS expression data are available through ArrayExpress under accession E-GEUV-3. The Genome in a Bottle (GIAB) data can be downloaded from ENA (PRJNA200694). Large supplementary data files such as the assembled genomes are available online (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20230412_sigY_assembly_data).

Code availability

Project code implemented to produce the assemblies and the basic quality control/evaluation statistics is available at GitHub (https://github.com/marschall-lab/project-male-assembly). All scripts written and used in the study of the Yq12 subregion are available at GitHub (https://github.com/Markloftus/Yq12).

42. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
43. Mendez, F. L. et al. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am. J. Hum. Genet.* **92**, 454–459 (2013).

44. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440 (2022).
45. Logsdon, G. HMW gDNA purification and ONT ultra-long-read data generation v3. *Protocols.io* https://doi.org/10.17504/protocols.io.b55tq86n (2022).
46. Gong, L., Wong, C.-H., Idol, J., Ngan, C. Y. & Wei, C.-L. Ultra-long read sequencing for whole genomic dna analysis. *J. Vis. Exp.* https://doi.org/10.3791/58954 (2019).
47. Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C. J. & Lansdorp, P. M. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nat. Protoc.* **12**, 1151–1176 (2017).
48. Sanders, A. D. et al. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).
49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
50. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
51. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
52. Poznik, G. D. et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).
53. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
54. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
56. Fu, Q. et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
57. Mölder, F. et al. Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
58. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
59. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
60. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
61. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
62. 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).
63. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
64. 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).
65. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
66. Shepelev, V. A. et al. Annotation of suprachromosomal families reveals uncommon types of alpha satellite organization in pericentromeric regions of hg38 human genome assembly. *Genom Data* **5**, 139–146 (2015).
67. Altemose, N. A classical revival: human satellite DNAs enter the genomics era. *Semin. Cell Dev. Biol.* **128**, 2–14 (2022).
68. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
69. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
70. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
71. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
72. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
73. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
74. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btac018 (2022).
75. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
76. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
77. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
78. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
79. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).
80. Helgason, A. et al. The Y-chromosome point mutation rate in humans. *Nat. Genet.* **47**, 453–457 (2015).
81. Ren, J. & Chaisson, M. J. P. lra: a long read aligner for sequences and contigs. *PLoS Comput. Biol.* **17**, e1009078 (2021).
82. Heller, D. & Vingron, M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1034 (2020).
83. Smolka, M. et al. Comprehensive structural variant detection: from mosaic to population-level. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.04.487055 (2022).
84. Zheng, Z. et al. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci.* **2**, 797–803 (2022).
85. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* **21**, 189 (2020).
86. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).
87. Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675 (2019).
88. 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).
89. Cunningham, F. et al. Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).
90. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).
91. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
92. Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A. & Minh, B. Q. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res.* **44**, W232–W235 (2016).
93. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proc. 7th Python in Science Conference (SciPy2008)* (eds Varoquaux, G. et al.) 11–15 (SciPy, Pasadena, 2008).
94. Raghavan, U. N., Albert, R. & Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Phys. Rev. E* **76**, 036106 (2007).
95. Zhou, W. et al. Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.* **48**, 1146–1163 (2020).
96. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1016 (2020).
97. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
98. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
99. Snajder, R., Leger, A., Stegle, O. & Bonder, M. J. pycoMeth: a toolbox for differential methylation testing from Nanopore methylation calls. *Genome Biol.* **24**, 83 (2023).
100. Cuomo, A. S. E. et al. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* **22**, 188 (2021).
101. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015).
102. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
103. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
104. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
105. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2012).
106. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
107. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).
108. Dekker, J. et al. The 4D nucleome project. *Nature* **549**, 219–226 (2017).
109. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
110. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
111. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102–1104 (2000).
112. Yadav, S. K., Kumari, A., Javed, S. & Ali, S. DYZ1 arrays show sequence variation between the monozygotic males. *BMC Genet.* **15**, 19 (2014).
113. Prosser, J., Frommer, M., Paul, C. & Vincent, P. C. Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187**, 145–155 (1986).
114. 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).
115. Nurk, S. et al. The complete sequence of a human genome. *Science* https://doi.org/10.1101/2021.05.26.445798 (2021).
116. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
117. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).
118. Konkel, M. K., Walker, J. A. & Batzer, M. A. LINEs and SINEs of primate evolution. *Evol. Anthropol.* **19**, 236–249 (2010).
119. Hoyt, S. J. et al. From telomere to telomere: the transcriptional and epigenetic state of human repeat elements. *Science* **376**, eabk3112 (2022).

Acknowledgements Funding was provided by National Institutes of Health (NIH) grants U24HG007497 (to C. Lee, E.E.E., J.O.K. and T.M.), U01HG010973 (to T.M., E.E.E. and J.O.K.), R01HG002385 and R01HG010169 (to E.E.E.), and GM123312 (to S.J.H. and R.J.O.); the German Federal Ministry for Research and Education (BMBF 031L0184 to J.O.K. and T.M.); the German Research Foundation (DFG 391137747 to T.M.); the German Human Genome-Phenome Archive (DFG (NFDI 1/1) to J.O.K.); the European Research Council (ERC Consolidator grant 773026 to J.O.K.); the EMBL (to J.O.K. and P. Hasenfeld); the EMBL International PhD Programme (to W.H.); the Jackson Laboratory Postdoctoral Scholar Award (to K.K.); NIH National Institute of General Medical Sciences (NIGMS R35GM133600 to C.R.B.; 1P20GM139769 to M.K.K. and M.L.) and the National Cancer Institute (NCI) (P30CA034196 to C.R.B. and P.A.A.); U24HG007497

(P. Hallast, F.Y., Q.Z., F.T. and J.Y.K.); NIGMS K99GM147352 (to G.A.L.); and Wellcome grant 098051 (to C.T.-S.). This work was also supported, in part, by the P30 CA034196 grant from the NCI. E.E.E. is an investigator of the Howard Hughes Medical Institute. We thank A. Rhie and A. Phillippy for coordination and discussions; Y. Xue for discussions and advice throughout the project; J. Wood and the members of the Genome Reference Informatics Team at the Wellcome Sanger Institute for suggestions and feedback on assembly evaluation; L. Skov for advice and sharing his scripts for gene conversion detection; the members of the HPRC (https://humanpangenome.org) for making their data publicly available; the staff at Clemson University for their allotment of compute time on the Palmetto Cluster; staff at the Center for Information and Media Technology at Heinrich Heine University Düsseldorf and the Scientific Services at the Jackson Laboratory, including the Genome Technologies Service for their assistance with the work described herein and Research IT for providing computational infrastructure and support and the members of the Phillippy laboratory (NIH/NHGRI) for their Verkko support; and the people who contributed samples as part of the 1000 Genomes Project.

Author contributions PacBio production sequencing: Q.Z., K.M.M., A.P.L. and J.K. ONT production: Q.Z. and K.H. Strand-seq production: P. Hasenfeld. and J.O.K. ONT re-basecalling and methylation calling: P.A.A. and W.T.H. Genome assembly: P.E., F.Y. and T.M. Assembly analysis and evaluation: P.E., P. Hallast, F.Y., W.H. and F.T. Assembly-based variant calling: P.E., P.A.A., P. Hallast and C.R.B. Variant quality control, merging and annotation: P.A.A. and P. Hallast. Short-read calling, phylogeny construction and dating: P. Hallast. Analysis of Bionano Genomics optical maps: F.Y. Strand-seq inversion detection and genotyping: D.P. MEI discovery and integration: W.Z., M.L. and M.K.K. Inversion analysis: P. Hallast, D.P., K.K., M.L. and M.K.K. Gene conversion and evolutionary rate: K.K., P. Hallast and M.K.K. Gene families: M.L., F.Y. and M.K.K. Analyses on Y subregions: P.E., P. Hallast, M.L., F.Y., G.A.L., P.A.A., W.H., K.K., F.T., M.K.K., E.E.E. and C.Lee. RNA-seq analysis: M.J.B. Methylation and meQTL analysis: M.J.B. HiC analysis: C. Li. and X.S. Repeat annotation: S.J.H. and R.J.O. Iso-seq analysis: P.C.D. and E.E.E. Gene annotations F.Y. and P.C.D. Supplementary materials: P. Hallast, P.E., M.L., F.Y., P.A.A., G.A.L., M.J.B., W.Z., W.H., K.K., C. Li, S.J.H., P.C.D., F.T., J.Y.K., Q.Z., K.M.M., P. Hasenfeld, X.S. and M.K.K. Display items: P. Hallast, P.E., M.L., F.Y., G.A.L., W.H., K.K., F.T. and M.K.K. Manuscript writing: P. Hallast, P.E., M.L., P.A.A., G.A.L., M.J.B., W.Z., M.K.K., C.Lee with contributions from all of the other authors. All of the authors contributed to the final interpretation of data. HGSVC co-chairs: C. Lee, J.O.K., E.E.E. and T.M.

Competing interests E.E.E. is a scientific advisory board member of Variant Bio, Inc. C. Lee is a scientific advisory board member of Nabsys and Genome Insight. The following authors have previously disclosed a patent application (no. EP19169090) relevant to Strand-seq: J.O.K., T.M. and D.P. The other authors declare no competing interests.

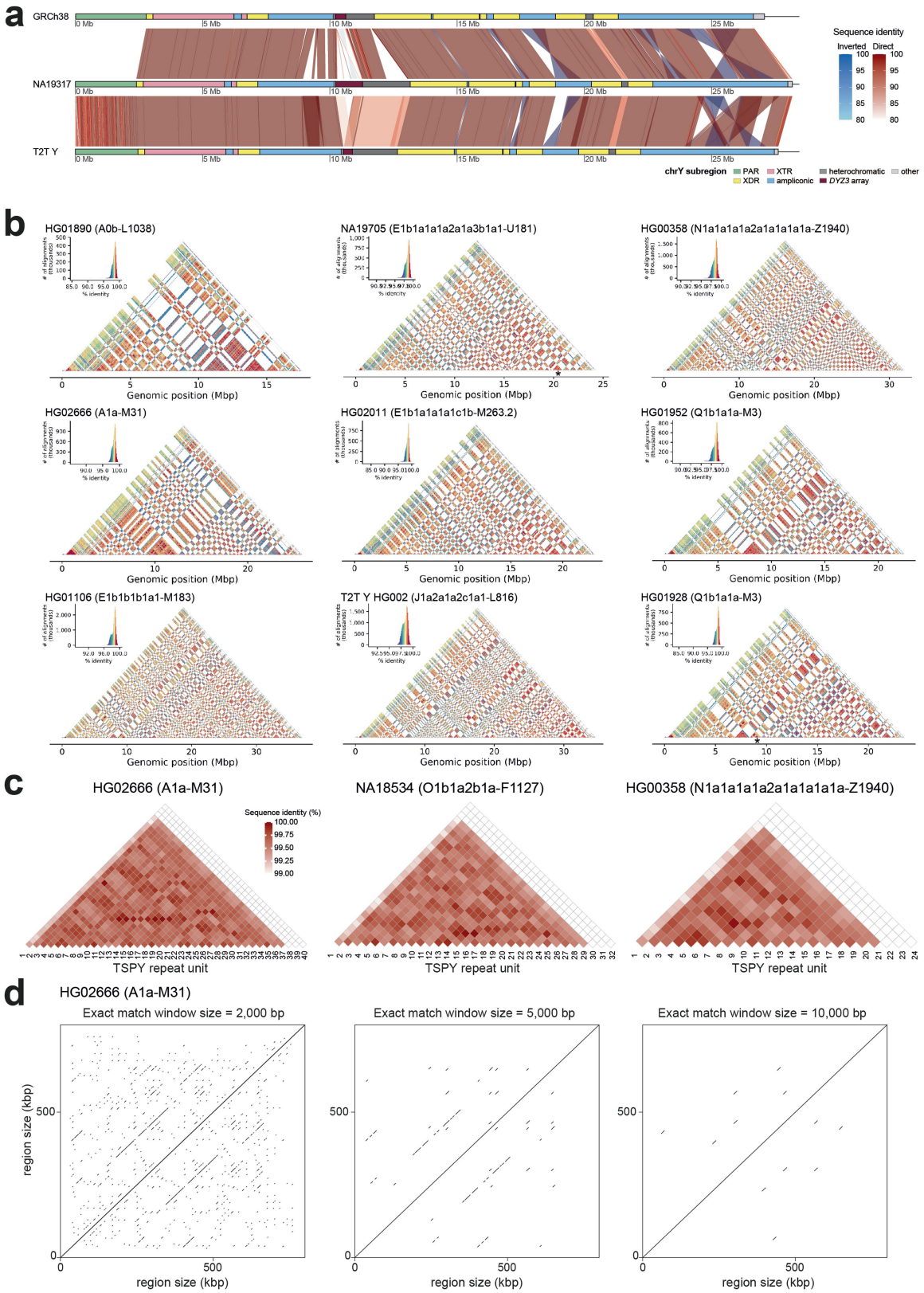
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06425-6.

Correspondence and requests for materials should be addressed to Charles Lee.

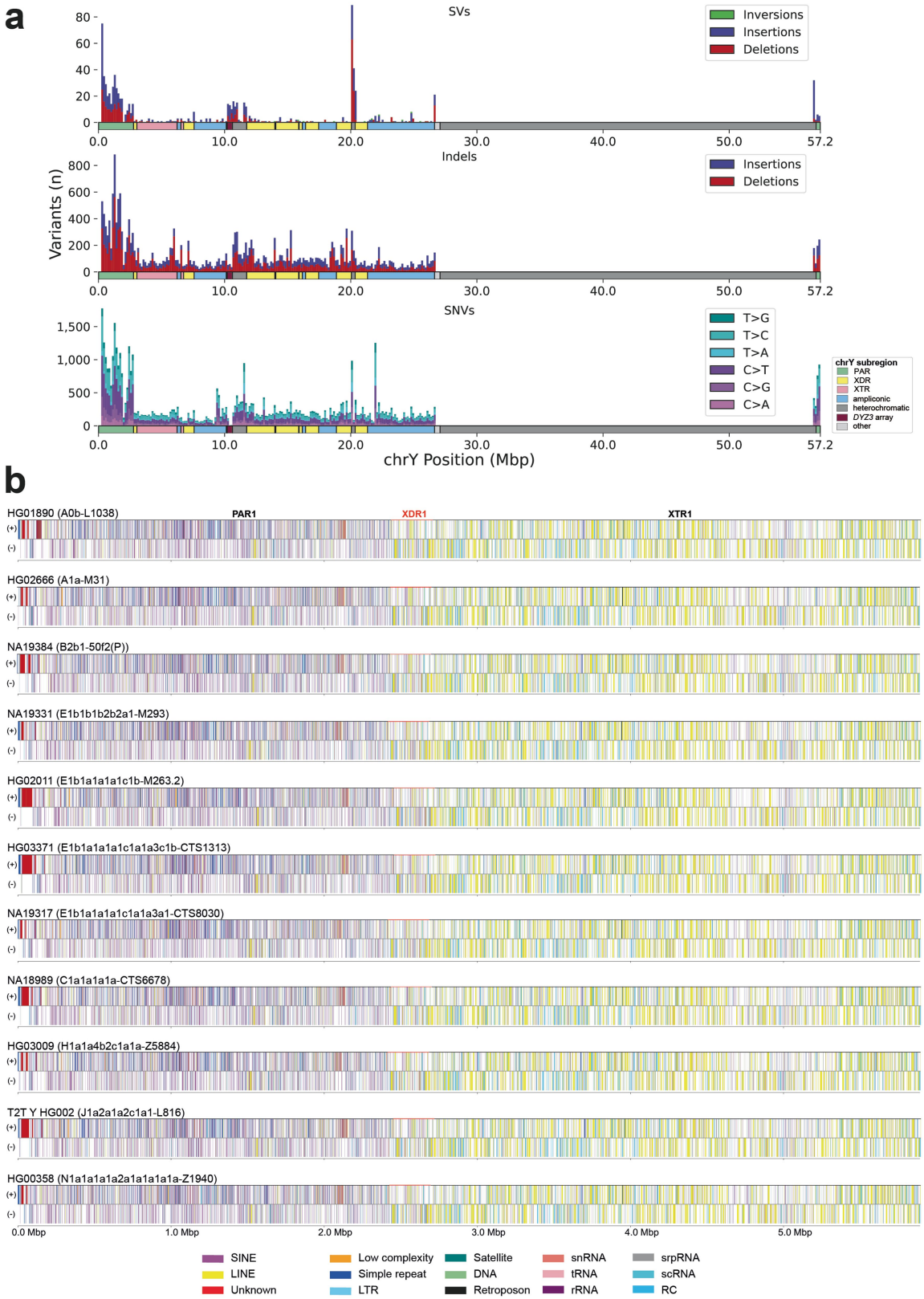
Peer review information *Nature* thanks Mikkel Heide Schierup, John Lovell and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at http://www.nature.com/reprints.



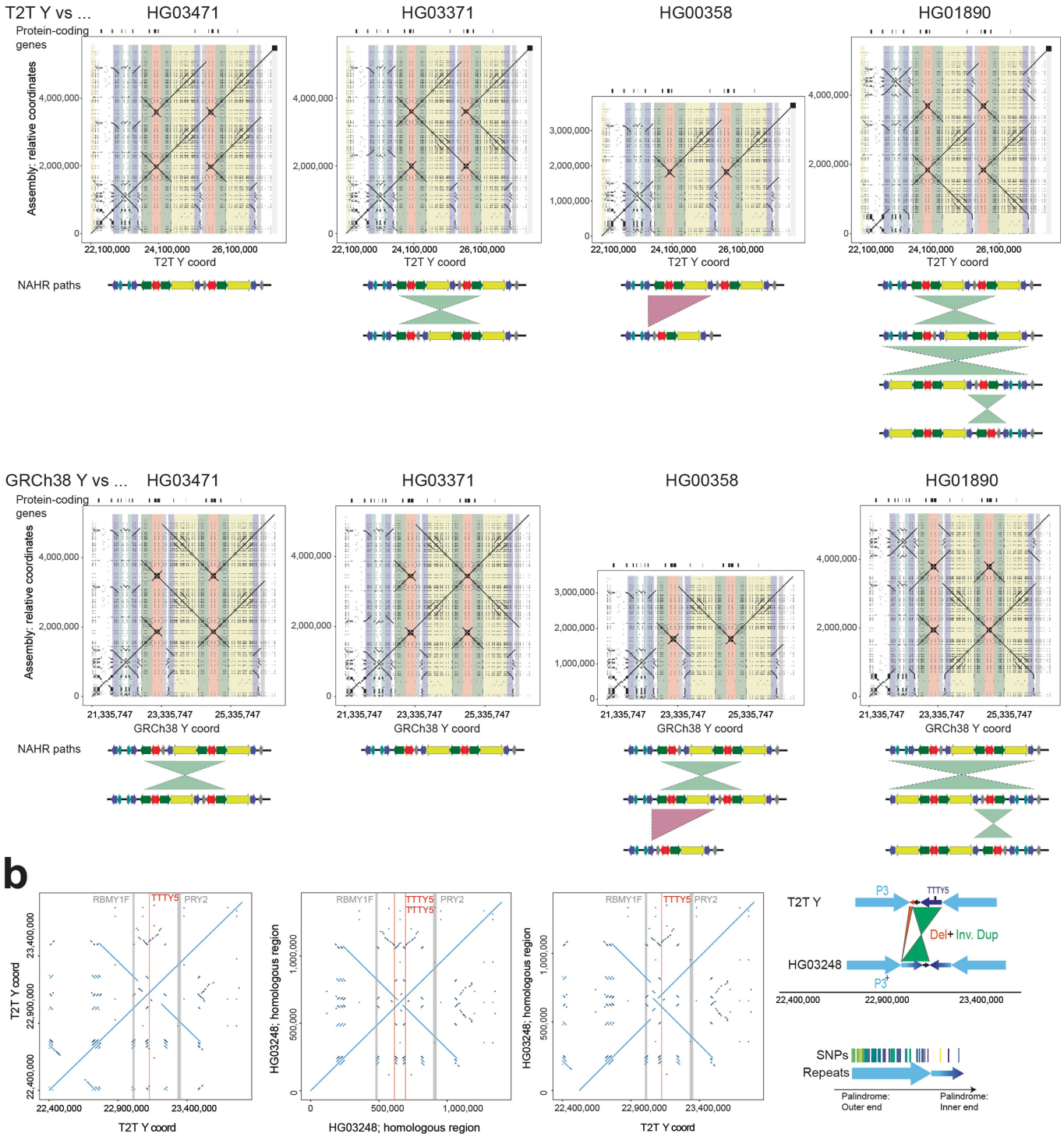
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Variation in structure and composition across Y-chromosomal subregions. a. Overview of the Y chromosome. A three-way comparison of sequence identity between GRCh38 Y, NA19317 (E1b1a1a1a1c1a1a3a1-CTS8030) and the T2T Y (excluding Yq12 and PAR2 subregions), highlighting substantial differences in the size and orientation of some subregions. **b**. Focus on Yq12. Sequence identity heatmaps of the Yq12 subregion for six contiguously assembled samples (HG01890, HG02666, HG01106, HG02011, HG00358 and HG01952), two samples (NA19705 and HG01928) with a single gap in the Yq12 subregion (gap location marked with asterisk) and the T2T Y using 5kb window size. **c**. Focus on TSPY repeat array. Sequence identity heatmaps of ~20.3-kbp TSPY repeat units for three males highlighting putative expansion events harbouring both single and multiple repeat units. Red shades from lighter to darker indicate sequence identity from 99–100%, respectively, while white fill indicates sequence identity <99%. The last copy on the right is the single separate repeat unit containing the *TSPY2* gene. See Fig. S22 for heatmaps of all samples. **d**. Dotplots of the TSPY repeat array for HG02666 with 5 kbp of flanking regions showing identical matches of 2, 5, and 10 kbp in size indicating regions with high sequence identity. See Fig. S25 for additional examples.

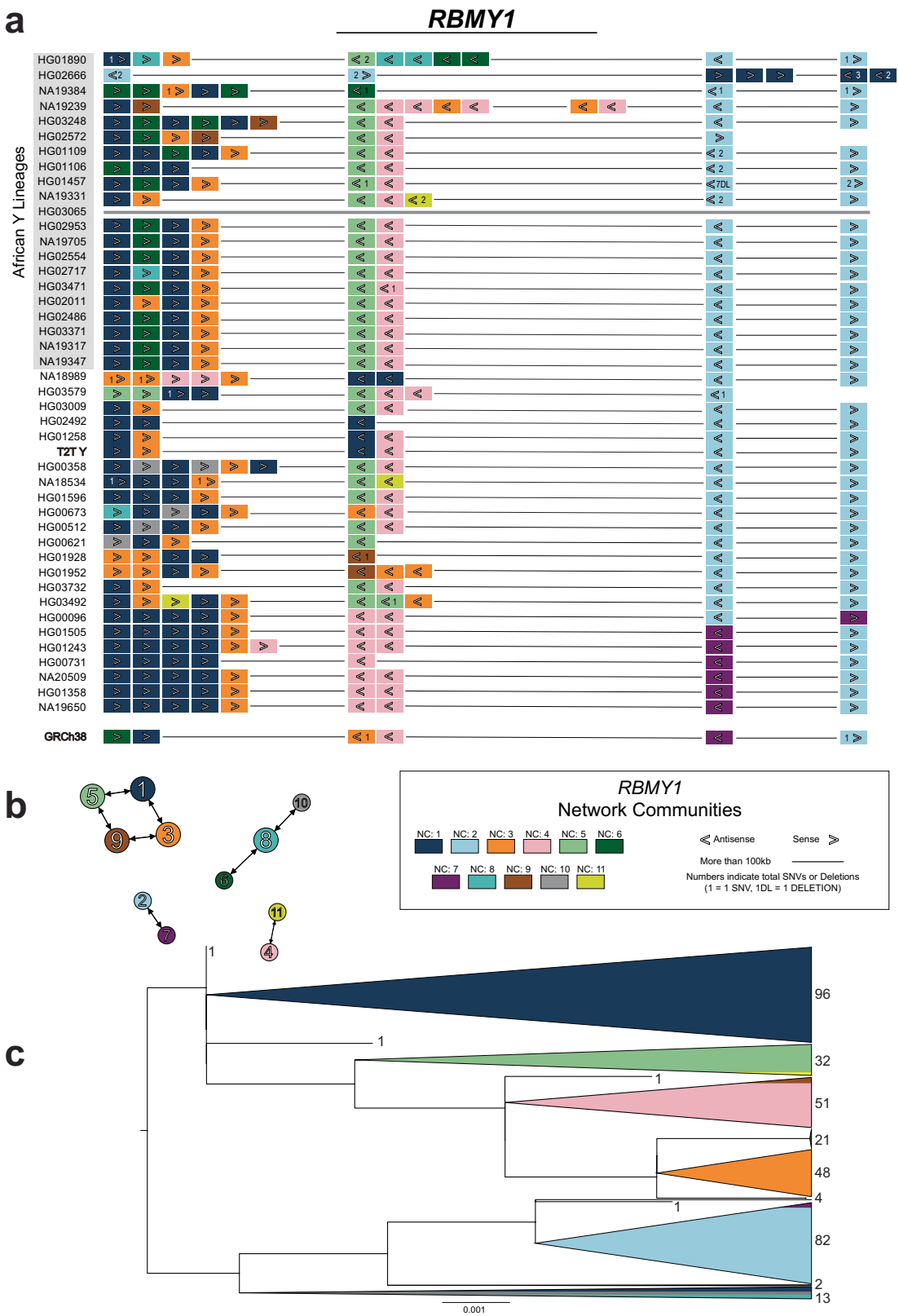


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Distribution of genetic variants across the Y chromosome and repeat elements in PAR1, XDR1 and XTR1 subregions. **a**. Distribution of variant sizes for SVs (≥ 50 bp, top), Indels (< 50 bp, middle), and SNVs (bottom) with the Y chromosome coloured by subregion. High peaks in heterochromatin are apparent for SVs, but are absent in SNVs and indels. **b**. Repeat element distribution across 10 samples with contiguously assembled PAR1 regions and the T2T Y. Repeat elements on sense (+) and antisense (−) strand are shown, coloured according to repeat class. Extensive differences in size can be seen between samples, especially in the satellite arrays located close to the telomere (in dark red), and substantial differences in repeat element composition in PAR1 vs. the male-specific XDR1 and XTR1 regions. The locations of PAR1, XDR1 and XTR1 subregions in each individual are shown in black, red and black, respectively. Please note that the maroon colour of the "Unknown" elements close to the telomere is caused by significant clustering of those elements. DNA: DNA repeat elements, snRNA: small nuclear RNA, tRNA: transfer RNA, rRNA: ribosomal RNA, srpRNA: signal recognition particle RNA, scRNA: small conditional RNA, RC: rolling circle.

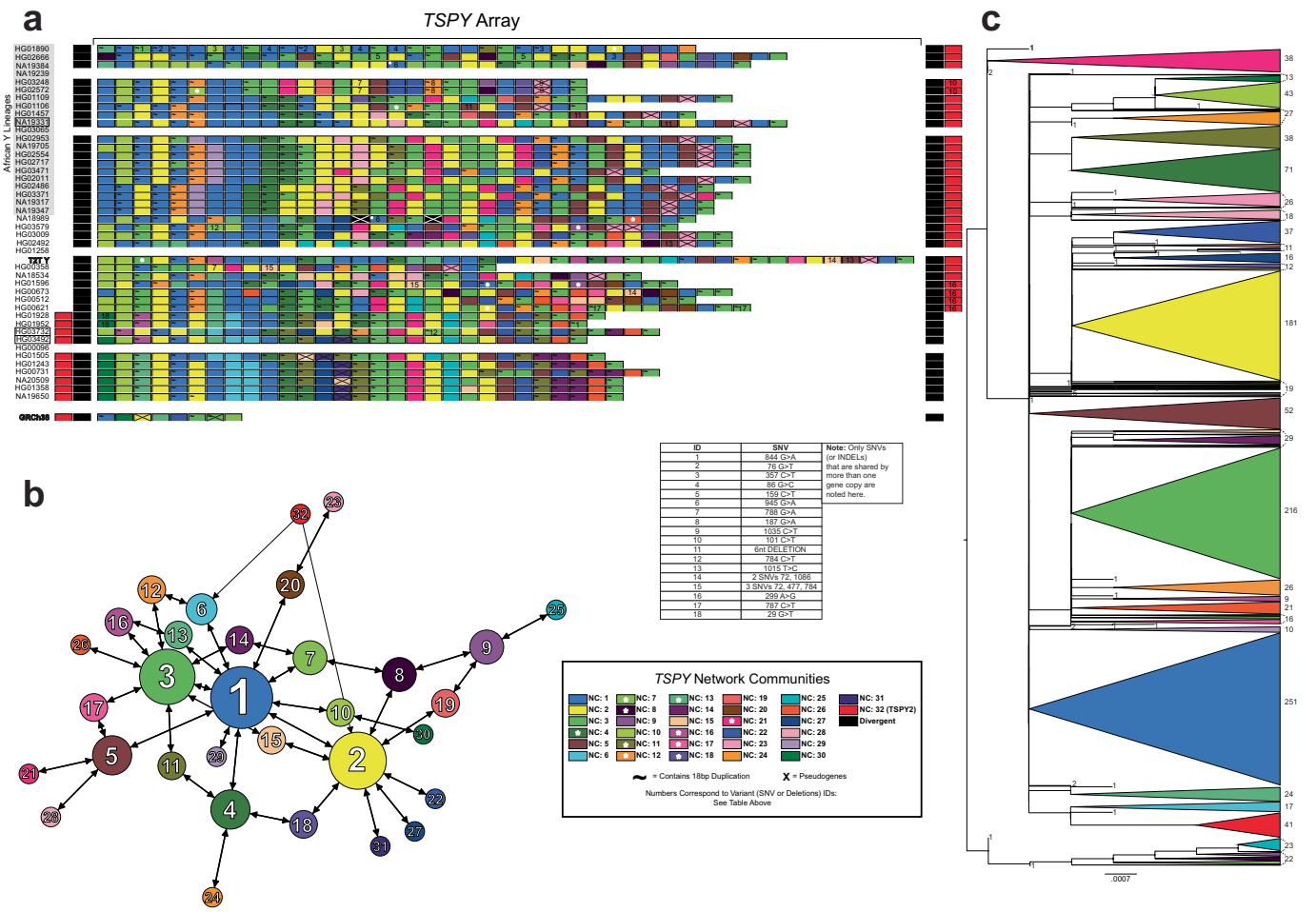


Extended Data Fig. 3 | Examples of structural variation identified in the assembled Y chromosomes. a. Inversions identified in the *AZFc*/ampliconic 7 subregion. Top - comparison between the T2T Y and select *de novo* assemblies, bottom - GRCh38 Y and the *de novo* assemblies (see Fig. S34 for details on *AZFc*/ampliconic 7 subregion composition). Potential NAHR path is shown below the dotplot. **b.** Inverted duplication affecting roughly two thirds of the 161 kbp unique 'spacer' sequence in the P3 palindrome, spawning a second copy of the *TTTY5* gene and elongating the LCRs in this region. A detailed sequence view reveals a high sequence similarity between the duplication and its template, and its placement in Y phylogeny supports emergence of this variant in the common ancestor of haplogroup E1a2 carried by NA19239, HG03248 and HG02572 (Fig. 3a).

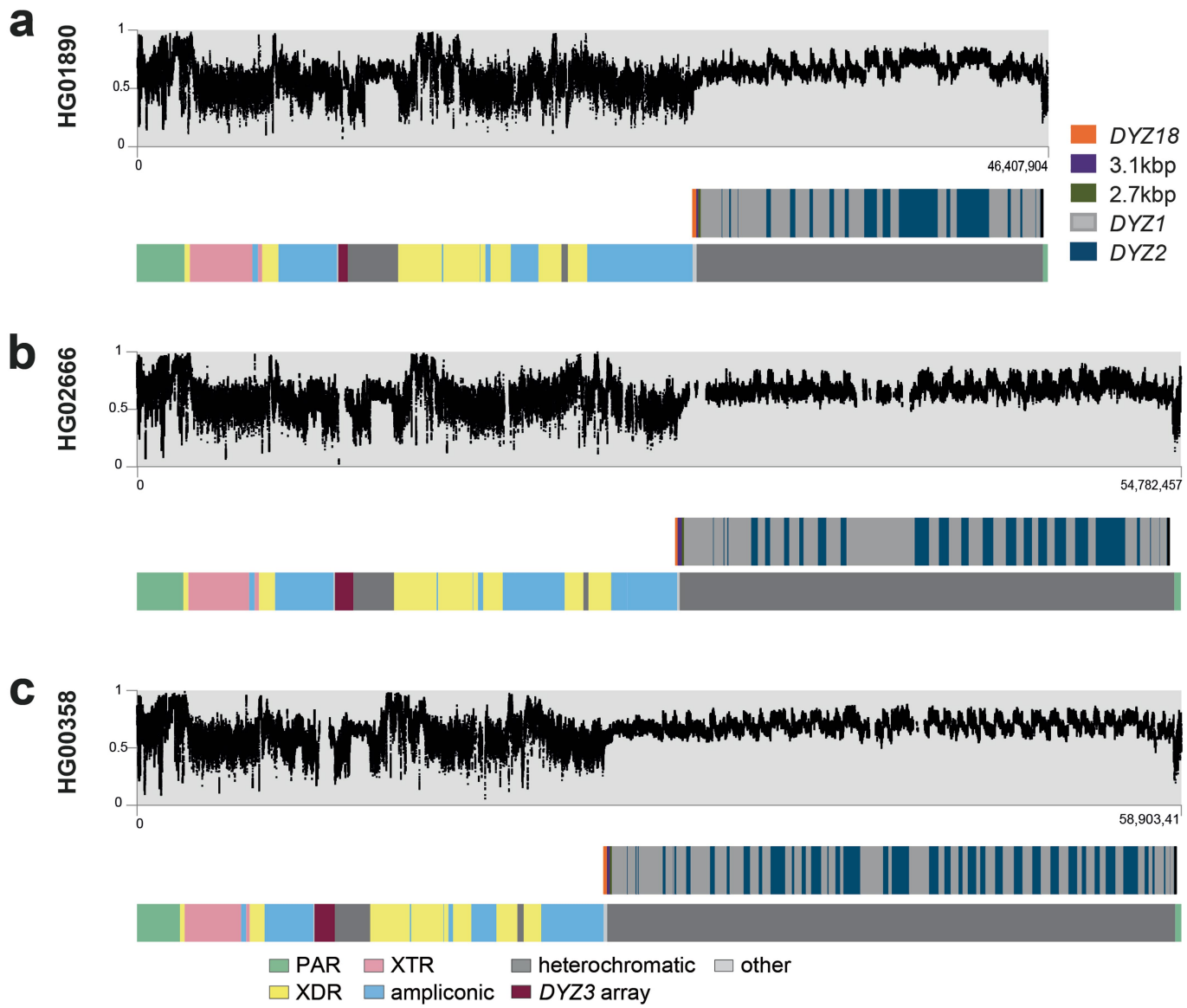


Extended Data Fig. 4 | See next page for caption.

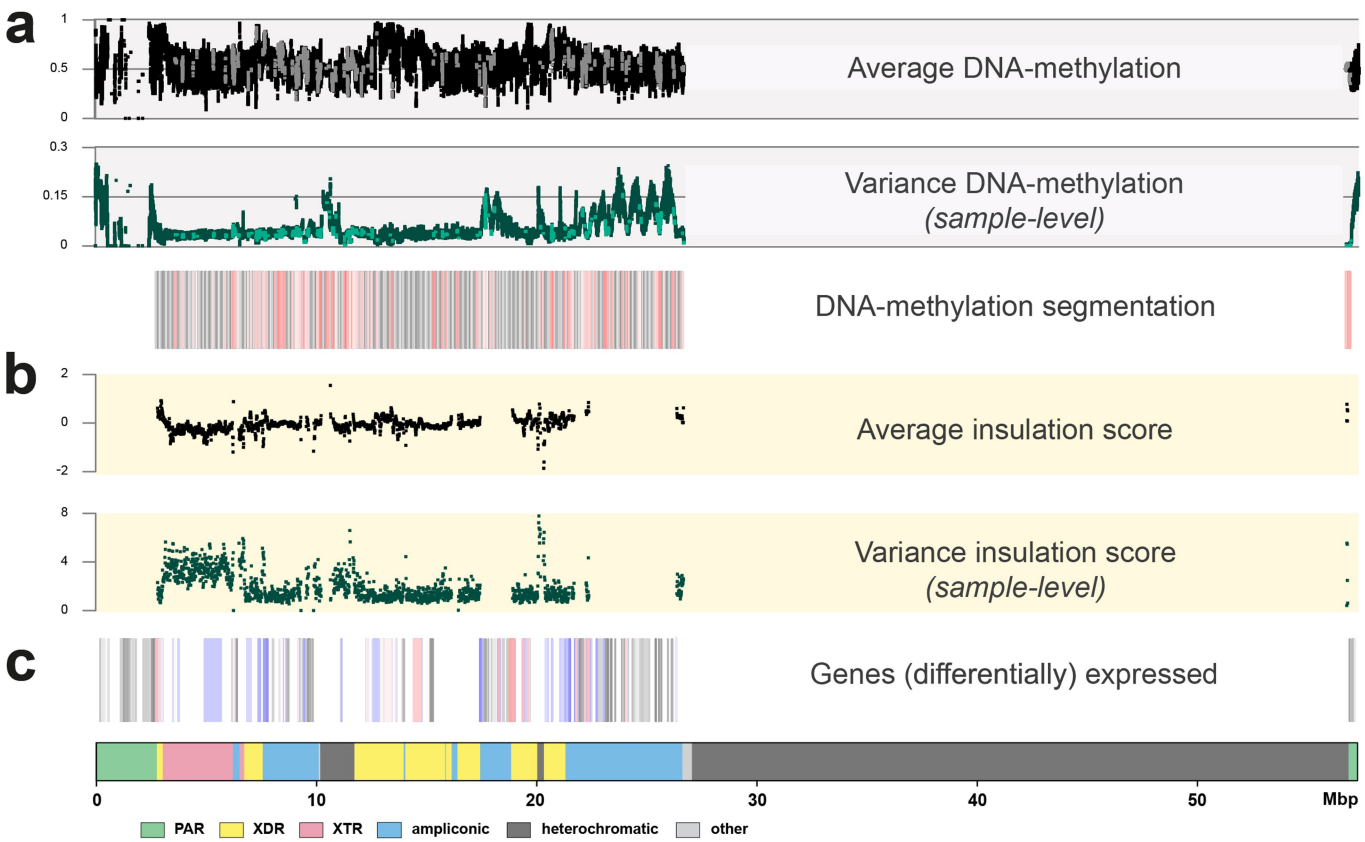
Extended Data Fig. 4 | *RBMY1* gene similarity and architecture. a. A schematic distribution of individual *RBMY1* gene copies (filled rectangles) within analysed Y chromosome assemblies (42 + T2T + GRCh38). The *RBMY1* gene copies are located in four primary regions (NA19239 carries a partial duplication of gene region 2 and the composition of HG02666 suggests at least one inversion within the RBMY regions). Fill colours refer to the assigned network community (NC) and indicates a similar sequence (**Methods**). Assembly of this region was not contiguous in HG03065 (brown line) and was not included in the analysis. **b**. A secondary directed network showing connections between NCs with the most similar consensus sequences. An edge pointing from one node to a second node indicates that the second node was the first's closest match (i.e., most similar sequence; ties are allowed and shown as multiple edges stemming from a node). The width of the edge represents the sequence similarity between two nodes (i.e., NC consensus sequence similarity; thicker means fewer SNVs). The node size is representative of the total edges pointing to the node. **c**. *RBMY1* phylogenetic analysis of exonic nucleotide sequences. Shown is the unrooted phylogenetic tree of *RBMY1* genes constructed using a maximum likelihood approach (**Methods**). This tree is rooted at the midpoint with the total count of *RBMY1* copies shown on the right. The scale bar represents the average number of substitutions per site. *RBMY1* copies located in regions 1 and 2 (primarily dark blue, orange, dark/light green, and pink) distinguish themselves from those located downstream in regions 3 and 4 (primarily light blue and purple copies).



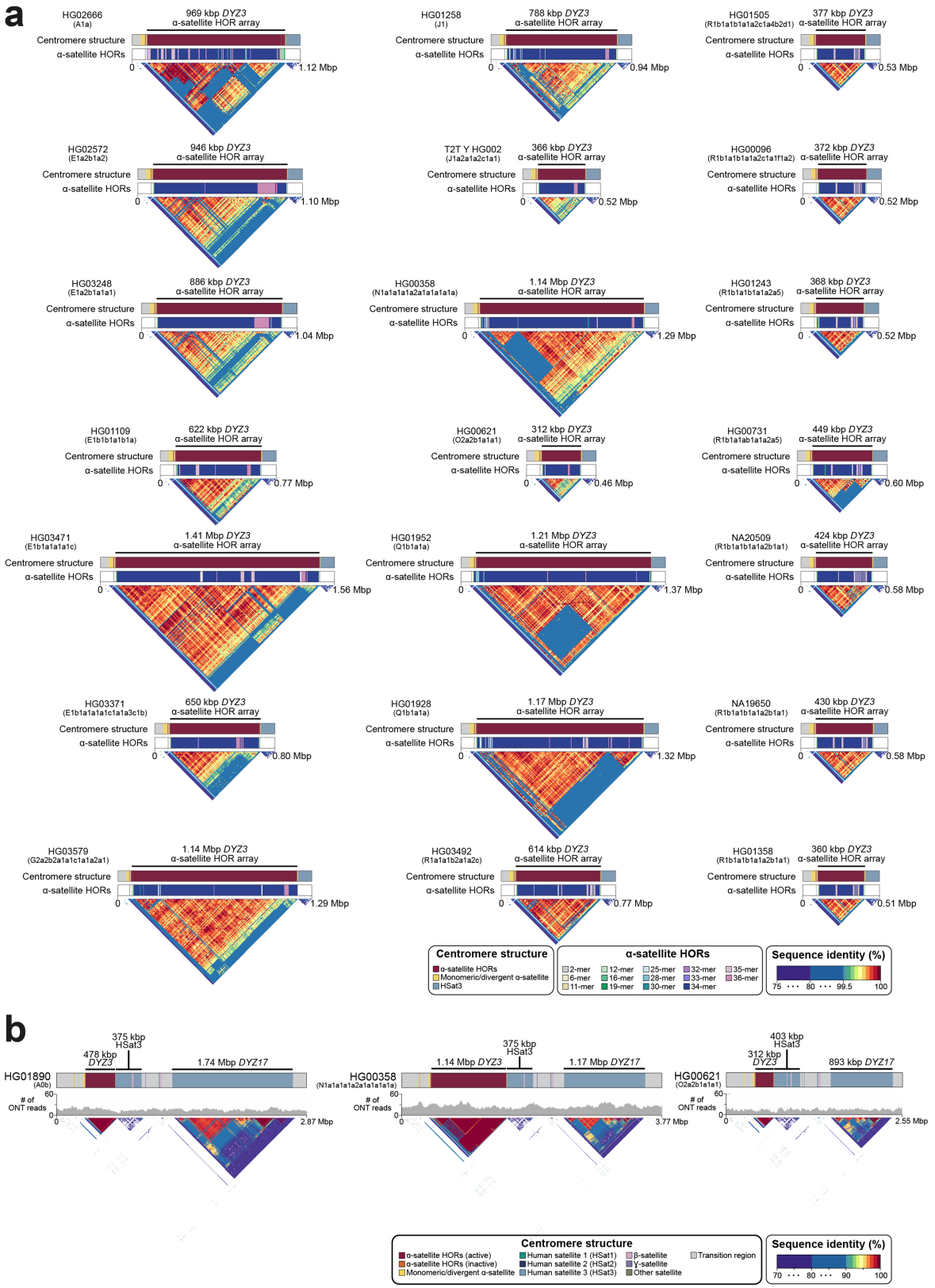
Extended Data Fig. 5 | *TSPY* gene similarity and architecture. a. TSPY array visualization of each sample with contiguous assembly in this region. Individual *TSPY* gene copies are shown (rectangles), and their colour is based on the assigned network community (NC) (**Methods**). Sample names with black rectangles (NA19331, HG03732 and HG03492) carry the IR3/IR3 inversion and were re-oriented for visualization. Asterisks within individual gene copies indicate possible gene conversion (GC) or recombination (R) events unique to that gene copy. If a GC/R event is shared by an NC an asterisk is shown in the NC legend rectangle. The *TSPY2* gene copy is shown as a red rectangle. **b**. A secondary directed network showing the sequence similarity between NC consensus sequences. An edge pointing from one node to a second node indicates that the second node was the first's closest match (i.e., most similar sequence; ties are allowed and shown as multiple edges stemming from a node). The width of the edge represents the sequence similarity between two nodes (i.e., NC consensus sequence similarity; thicker means fewer SNVs). The node size is representative of the total edges pointing to the node. **c**. *TSPY* phylogenetic analysis of exonic nucleotide sequences. Shown is the unrooted phylogenetic tree of *TSPY* genes constructed using a maximum likelihood approach (**Methods**). This tree is rooted at the midpoint and the total count of *TSPY* copies is shown on the right. The scale bar represents the average number of substitutions per site. The early split/rise of NC1 within the tree, in conjunction with the secondary directed network and manual comparison of *TSPY* sequences (as well as their presence across all lineages) suggests that NC1 *TSPY* copies represent the ancestral *TSPY* gene sequence.



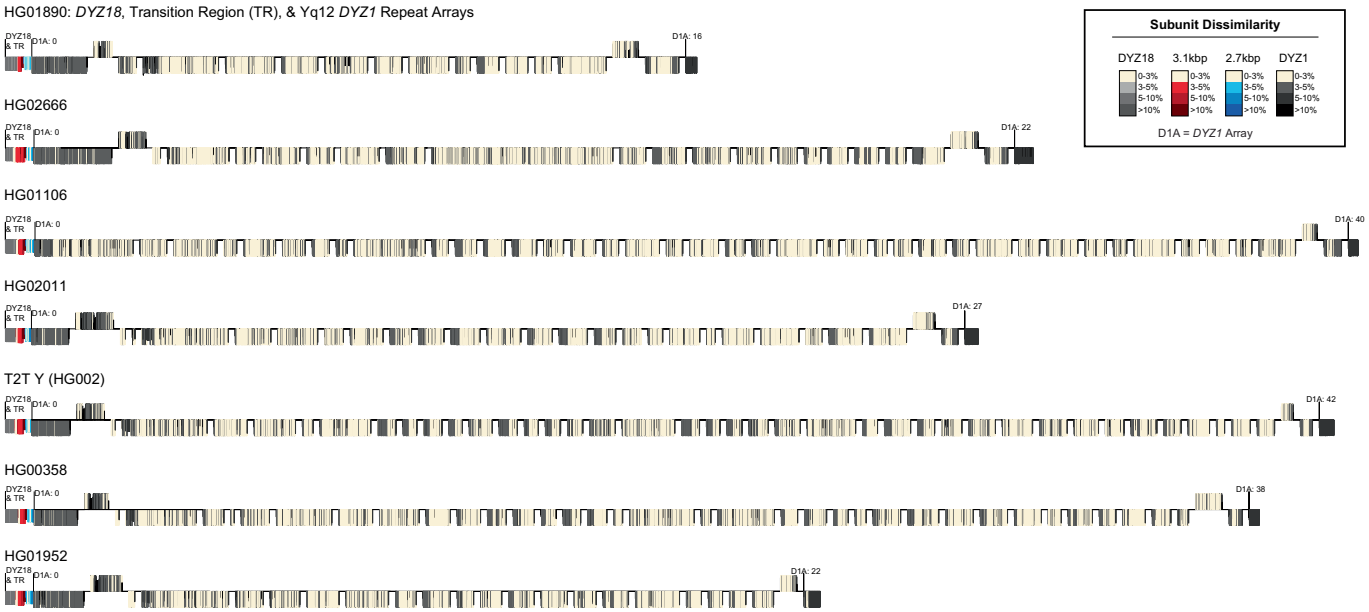
Extended Data Fig. 6 | DNA methylation patterns as determined from the ONT data across the three contiguously assembled Y chromosomes. Methylation patterns for samples: **a**. HG1890, **b**. HG02666 and **c**. HG00358. The three dot plots (in grey) show the smoothed DNAme levels, in 5 kbp windows for visualization, in beta-scale ranging from 0 (not methylated) to 1 (methylated). The locations of Yq12 repeat arrays (*DYZ18*, 2.7kb-repeat, 3.1kb-repeat, *DYZ1* and *DYZ2*) and the Y-chromosomal subregions are shown below as bar plots.



Extended Data Fig. 7 | Functional analyses on the Y chromosome with DNA-methylation, RNA expression and HiC information as anchored to GRCh38 Y. **a**. The top three panels show DNA-methylation levels and variation over the studied chromosomes (n = 41). In black (top dot plot) the average methylation is shown, in green (middle dot plot) the variation in DNAme levels across the studied genomes. The bottom boxplot represents the DNA methylation segmentation using PycoMeth-seg (**Methods**). In grey shades 2,861 methylation segments, and in red shades the 340 significantly differentially methylated segments (DMS). The CpG sites that fall in a DMS are coloured in a lighter shade in the top two dot plots, the dot plots are in beta-scale, ranging from 0 (not methylated) to 1 (methylated). **b**. Average insulation scores (top) and variance of insulation scores between any two samples (bottom) across 40 samples with Hi-C data with 10 kbp resolution. Regions with lower insulation scores are more insulated and more likely to be topologically associating domain (TAD) boundaries, while regions with higher scores are more likely to stay inside TADs (the regions between the two adjacent TAD boundaries). The y-axis represents the average insulation scores ranging from −2 (most insulated) to 2 (least insulated) and the variance insulation scores ranging from 0 (no variance) to 8 (more variance). **c**. The Geuvadis-based gene-expression analysis, shown are the 205 genes on the Y chromosome (grey shades), the 64 genes expressed in the Geuvadis LCLs (blue shades), of which 22 are differentially expressed (red shades, Supplementary Results 'Functional analysis' for additional details).



Extended Data Fig. 8 | Composition of the Y-chromosomal (peri-)centromeric regions. a. Organization of the chromosome Y centromeric region from 21 genomes representing all major superpopulations. The structure (top), α-satellite HOR organization (middle), and sequence identity heat map (bottom) for each centromere is shown and reveals the presence of novel HORs in over half of the centromeres. Note - the sizes of the *DYZ3* α-satellite array are shown on top as determined using RepeatMasker (**Methods**). **b.** Genetic landscape of the Y-chromosomal pericentromeric region for three select samples (see Figs. S47–S48 for all samples). The top panel shows locations and composition of the pericentromeric region with repeat array sizes shown for each Y chromosome (the *DYZ3* α-satellite array size as determined using RepeatMasker, **Methods**). The middle panel shows (UL-)ONT read depth and bottom sequence identity head maps generated using the StainedGlass pipeline[74] (using a 5 kbp window size).



Extended Data Fig. 9 | Divergence of *DYZ18*, Yq11/Yq12 transition region and *DYZ1* repeat units. An overview of the Bray-Curtis distance/dissimilarity of k-mer abundance profiles for individual *DYZ18* (grey), 3.1-kbp (red), 2.7-kbp (blue) and *DYZ1* (black) repeats versus their consensus sequence. The Yq11/transition region/Yq12 are shown for each of the seven samples with a completely assembled Yq12 subregion. Lighter colours indicate less distance/dissimilarity (i.e., more similar) k-mer abundance profiles compared to their consensus sequence. Results indicate that arrays located on the proximal and distal boundaries of the Yq12 subregion contain repeats with k-mer abundance compositions less similar to their consensus sequence (i.e., more diverged). The size of individual lines is a function of the length of the repeat. The repeat unit orientation (above = sense, below = antisense) was determined based on RepeatMasker annotations of satellite sequences within repeats (**Methods**).

Extended Data Fig. 10 | Divergence of Yq12 *DYZ2* repeat units. An overview of the divergence of individual *DYZ2* subunits for **a**. samples with completely assembled Yq12 subregion (HG01890, HG02666, HG01106, HG02011, T2T Y, HG00358, HG01952), and **b**. the two most closely related genomes (NA19317 and NA19347) with incompletely assembled Yq12 subregions. The size of individual lines is a function of the length of the repeat. The repeat unit orientation (above = sense, below = antisense) was determined based on RepeatMasker annotations of satellite sequences within repeats (**Methods**). A higher divergence was observed within the subunits located in arrays at the proximal and distal ends of the Yq12 subregion. Additionally, *DYZ2* subunits located near the boundaries of individual arrays tend to be more diverged than those located centrally. Between the closely related genomes, the divergence of *DYZ2* repeats within the shared *DYZ2* arrays are highly similar.

nature portfolio

Corresponding author(s): Charles Lee

Last updated by author(s): Jun 29, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

HiFi/CCS read analysis was performed using SMRT Link v10.1
BioNano optical mapping data collection was performed using Saphyr 2nd generation instruments (Part #60325) and Instrument Control Software (ICS) version 4.9.19316.1
Iso-seq reads were processed with default parameters using the PacBio Iso-seq3 pipeline
ONT read-level CpG DNA-methylation (DNAme) likelihood ratios were estimated using nanopolish version 0.11.1

Data analysis	BWA aligner (version 0.7.15-0.7.17); SAMtools (version 1.10); sambamba (version 1.0); BCFtools (v1.9); VCFtools (v0.1.16); BEAST (v1.10.4); RAxML (v8.2.10); TreeAnnotator (v1.10.4); FigTree software (v1.4.4); Verkko pipeline (v1.0); hifiasm (v0.16.1-r375); minimap2 (v2.24); HMMER (v3.3.2dev); VerityMap (v2.1.1-alpha-dev #8d241f4); DeepVariant (v1.3.0); PEPPER-Margin-DeepVariant pipeline (v0.8, DeepVariant v1.3.0); yak (v0.1) (github.com/lh3/yak); Bionano Solve (v3.5.1): pipelineCL.py; refAligner; hybridScaffold.pl; HMMER3 (v3.3.2); RepeatMasker (v4.1.0); HumAS-HMMER (v3.3.2); StringDecomposer (v1.0.0); "ElasticNet" from scikit-learn v1.1.1; bedcov (version 1.15.1); blastn; genoPlotR (0.8.11); StainedGlass; NAHRwhals package version 0.9; MAFFT (v7.487); Tandem Repeat Finder (v4.09.1); Gblocks (v0.91b); LiftOver tool at the UCSC Genome Browser web page (https://genome.ucsc.edu/cgi-bin/hgLiftOver); PAV (v2.1.0); minimap2 (v2.17); LRA (commit e20e67); PBSV (v2.8.0); SVIM-asm (v1.0.2); Sniffles (v2.0.7); DeepVariant (v1.1.0); Clair3 (v0.1.12); CuteSV (v2.0.1); LongShot (v0.4.5); SV-Pop (v3.3.5); VEP (version 107); Bionano Access (v1.7); EMBOSS cons (v6.6.0.0); PALMER (Pre-mAsking Long reads for Mobile Element inseRtion, v2.0.0); pycoMeth (version 2.2); R package vegan; Trim Galore! (v0.6.5); STAR (v2.7.5a); FeatureCounts (v2); Juicer software tools (version 1.6); BWA mem (version: 0.7.17); FAN-C toolkit (version 0.9.23b4); SciPy v1.7.3; Pysam (version 0.19.1); Muscle (v5.1); NucFreq (v0.1); Rukki (packaged with Verkko v1.2); IQ-Tree (v1.6.12); NetworkX (v2.8.4); Sequence Manipulation Suite (v2), Snakemake (v6.13.1) Project code implemented to produce the assemblies and the basic QC/evaluation statistics is available at github.com/marschall-lab/project-male-assembly. All scripts written and used in the study of the Yq12 subregion are available at https://github.com/Markloftus/Yq12.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data generated and used in this study were derived from lymphoblast lines available from the Coriell Institute for Medical Research for research purposes (https://www.coriell.org/), see Table S1 for additional details. All data generated by the HGSVC (PacBio HiFi, ONT-UL, HiC, RNA-Seq, IsoSeq and Bionano Genomics optical genome maps) are available at International Nucleotide Sequence Database Collaboration (INSDC) under the following project IDs: PRJEB58376, PRJNA988114, PRJEB41077, PRJEB39684 and PRJEB39750. The Human Pangenome Reference Consortium (https://humanpangenome.org/) year 1 PacBio HiFi, ONT long-read sequencing and Bionano Genomics optical mapping data files are available at INSDC under PRJNA701308. Please see Table S1 for additional details. Existing testis Iso-seq data from seven individuals is available from the European Nucleotide Archive (ENA) under accessions SRX9033926 and SRX9033927. The GEUVADIS expression data is available via ArrayExpress under accession E-GEUV-3. The Genome in a bottle (GIAB) data can be downloaded from ENA under accession: PRJNA200694. Large supplementary data files such as the assembled genomes are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20230412_sigY_assembly_data.

Additionally, GRCh38 (GCA_000001405.15) and the CHM13 (GCA_009914755.3) plus the T2T Y assembly from GenBank (CP086569.2) released in April 2022 were used in the current study.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Since the study focuses on the male-specific chromosome Y, only male participants from the 1000 Genomes Project Diversity Panel were included.
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).* *Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)* *Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	The male participants included in the study represent 23 human populations (and 5 super populations) as defined by the 1000 Genomes Project
Recruitment	Samples from diverse human populations were included to the Human Structural Variation Consortium (HGSVC) dataset, including some samples chosen to represent a specific Y lineage. The Human Pangenome Reference Consortium (HPRC) dataset is a publicly available dataset and all 15 Year 1 male samples were included.
Ethics oversight	The genomic DNA and lymphoblastoid cell lines for 1000 Genomes Project sample are available for research purposes from Coriell Institute (https://www.coriell.org/) and covered by appropriate ethics approvals by Coriell Institute

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	A total of 43 males samples with relevant sequencing data available were included: 28 samples from the Human Genome Structural Variation Consortium (HGSVC) dataset and 15 samples from the Human Pangenome Reference Consortium (HPRC).
Data exclusions	No data was excluded
Replication	Not applicable. All computational analyses can be replicated using the provided codes and pipelines.
Randomization	Not applicable. Samples were not assigned to groups.
Blinding	Not applicable. All experiments were done computationally and do not involve a human experimenter

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Coriell Institute for Medical Research (https://www.coriell.org/). Sequence data which has been derived from the following cell lines was used in the current study: HG01890, HG02666, NA19384, NA18989, NA19239, HG03248, HG02572, HG03471, HG02486, NA19317, NA19347, HG03371, HG02011, HG02717, HG02554, HG02953, NA19705, HG03065, HG01109, HG01106, HG01457, NA19331, HG03579, HG03009, HG01258, NA24385, HG02492, HG00358, NA18534, HG01596, HG00673, HG00621, HG00512, HG01928, HG01952, HG03492, HG00731, HG01243, NA19650, NA20509, HG01358, HG00096, HG01505, HG03732.
Authentication	We used sequence data derived from cell lines, and did not authenticate the cell lines
Mycoplasma contamination	We used sequence data derived from cell lines, and did not test for mycoplasma contamination. According to information provided by Coriell Institute, all cell lines are free of bacterial, fungal or mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.