Assembly of 43 diverse human Y chromosomes reveals extensive complexity and variation

Pille Hallast[1,*], Peter Ebert[2,3,*], Mark Loftus[4], Feyza Yilmaz[1], Peter A. Audano[1], Glennis A. Logsdon[5], Marc Jan Bonder[6], Weichen Zhou[7], Wolfram Höps[8], Kwondo Kim[1], Chong Li[9], Philip Dishuck[5], David Porubsky[5], Fotios Tsetsos[1], Jee Young Kwon[1], Qihui Zhu[1], Katherine M. Munson[5], Patrick Hasenfeld[8], William T. Harvey[5], Alexandra P. Lewis[5], Jennifer Kordosky[5], Kendra Hoekzema[5], The Human Genome Structural Variation Consortium (HGSVC), Jan O. Korbel[8], Chris Tyler-Smith[10], Evan E. Eichler[5,11], Xinghua Shi[9], Christine R. Beck[1,12], Tobias Marschall[2], Miriam K. Konkel[4], Charles Lee[1]

[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]Clemson University, Department of Genetics & Biochemistry, Clemson, SC, USA

[5]University of Washington School of Medicine, Department of Genome Sciences, Seattle, WA, USA

[6]German Cancer Research Center (DKFZ), Division of Computational Genomics and Systems Genetics, Heidelberg, Germany

[7]University of Michigan Medical School, Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, USA

[8]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany

[9]Temple University, Department of Computer and Information Sciences, Philadelphia, PA, USA

[10]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK

[11]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

[12]The University of Connecticut Health Center, Farmington, CT, USA

*These authors contributed equally to this work

Correspondence to: Charles Lee charles.lee@jax.org

Abstract

The prevalence of highly repetitive sequences within the human Y chromosome has led to its incomplete assembly and systematic omission from genomic analyses. Here, we present long-read *de novo* assemblies of 43 diverse Y-chromosomes, three contiguously assembled including two from deep-rooted African Y lineages. Examination of the full extent of genetic variation between Y chromosomes across 180,000 years of human evolution reveals its remarkable complexity and diversity in size and structure, in contrast with its low level of base substitution variation. The size of the Y chromosome assemblies vary extensively from 45.2 to 84.9 Mbp, with individual repeat arrays showing up to 6.7-fold difference in length across samples. Half of the male-specific euchromatic region is subject to large (up to 5.94 Mbp) inversions with a >2-fold higher recurrence rate compared to the rest of the human genome. The Y centromere, composed of 171 bp α-satellite monomer units, appears to have evolved from tandem arrays of a 36-mer ancestral higher order repeat (HOR), which has been predominantly replaced by a 34-mer HOR, and reveals a pattern of higher sequence variation towards the short-arm side. The Yq12 heterochromatic region is ubiquitously flanked by approximately 649 kbp and 472 kbp inversions that maintain the alternating arrays of *DYZ1* and *DYZ2* repeat units in between. While the sizes and the distribution of the *DYZ1* and *DYZ2* arrays vary considerably, primarily due to local expansions and contractions, the copy number ratio between the *DYZ1* and *DYZ2* monomer repeat units remains consistently close to 1:1. In addition, we have identified on average 65 kbp of novel sequence per Y chromosome. The availability of sequence-resolved Y chromosomes from multiple samples provides a basis for identifying new associations of specific traits with the Y chromosome and garnering novel evolutionary insights.

Introduction

The mammalian sex chromosomes evolved from a pair of autosomes, gradually losing their ability to recombine over increasing lengths, leading to degradation and accumulation of large proportions of repetitive sequences[1]. The resulting sequence composition of the human Y chromosome is rich in complex repetitive regions, including highly similar segmental duplications (SDs)[2,3]. 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 via a laborious approach of mapping and Sanger sequencing of bacterial artificial chromosome (BAC) clones, which provided a high quality but incomplete sequence (~30.8/57.2 Mbp unresolved in GRCh38)[3]. Less than half (~25 Mbp) of the GRCh38 Y chromosome is composed of euchromatin which contains two pseudoautosomal regions, PAR1 and PAR2 (~3.2 Mbp in total), that actively recombine with homologous regions on the X chromosome and are therefore not considered as part of the male-specific Y region (MSY)[3]. The remainder of the Y-chromosomal euchromatin (~22 Mbp) has been divided into three main classes according to their sequence composition and evolutionary history[3]: (i) the X-degenerate regions (XDR, ~8.6 Mbp) are remnants of the ancient autosomes from which the X and Y chromosomes evolved, (ii) the X-transposed regions (XTR, ~3.4 Mbp) resulted from a duplicative transposition event from the X chromosome followed by an inversion, and (iii) the ampliconic regions (~9.9 Mbp) that contain sequences having up to 99.9% intra-chromosomal identity across tens or hundreds of kilobases (**Fig. 1a**). The rest of the Y chromosome is largely composed of repetitive centromeric and heterochromatic sequences, including the (peri-)centromeric *DYZ3* α-satellite and *DYZ17* arrays, *DYZ18* and *DYZ19* arrays, and the large Yq12 block, which is known to be highly variable in size[3,4,5]. All these heterochromatic regions are thought to be predominantly satellites, simple repeats and segmental duplications[3,6].

The current 57.2 Mbp GRCh38 Y reference assembly is a patched version of the 2003 Sanger assembly and is still structurally incomplete, as it is composed of 53.8% missing sequence (N's) (**Fig. 1a**). Past attempts have been made to assemble the human Y chromosome using Illumina short-read[7] and Oxford Nanopore Technologies (ONT) long-read data[8], but a contiguous assembly of the ampliconic and heterochromatic regions was not achieved.

In April 2022, the first complete *de novo* assembly of a human Y chromosome (from individual HG002/NA24385, carrying a rare J1a-L816 Y lineage found among Ashkenazi Jews and Europeans[9]), was deposited in GenBank by the Telomere-to-Telomere (T2T) Consortium[10]. However, understanding the composition and appreciating the complexity of the Y chromosomes in the human population requires access to assemblies from many diverse individuals. Here, we have combined PacBio HiFi and ONT long-read sequence data to assemble the Y chromosomes from 43 males, representing the five continental groups from the 1000 Genomes Project. While both the GRCh38 (mostly R1b-L20 haplogroup) and the T2T Y represent European Y lineages, 21/43 (49%) of our Y chromosomes represent African lineages and include most of the deepest-rooting human Y lineages. This newly assembled dataset of 43 Y chromosomes thus provides a more comprehensive view of genetic variation at the nucleotide level across over 180,000 years of human Y chromosome evolution.

Results

Sample Selection

We selected 43 genetically diverse males from the 1000 Genomes Project that had accompanying data recently generated by the Human Genome Structural Variation Consortium (HGSVC) (n=28)[11] and the Human Pangenome Reference Consortium (HPRC) (n=15)[12] (**Table S1**). These 43 males include three samples carrying the deepest-rooting African Y lineages present among the 1000 Genomes Project (HG01890, HG02666 and NA19384, which carry A0b-L1038, A1a-M31 and B2b-M112, respectively)[13] (**Fig. 1b**). The time to the most recent common ancestor (TMRCA) among our 43 Y chromosomes and the Y assembly from HG002/NA24385 (J1a-L816 haplogroup, termed as T2T Y) was estimated to be approximately 183 thousand years ago (kya) (95% HPD interval: 160-209 kya) (**Fig. S1; Methods**), consistent with previous reports[14,15]. Additionally, a pair of closely-related African Y chromosomes, representing the E1b1a1a1a-CTS8030 lineage (NA19317 and NA19347), were included for assembly validation, as these Y chromosomes are expected to be highly similar (TMRCA 200 ya [95% HPD interval: 0 - 500 ya]). Taken together, the 43 samples we analyzed represent 21 largely African (haplogroups A, B and E)[16] and 22 non-African Y haplogroups (**Table S1**). Notably, there is an African Y lineage (A00) older than the lineages in our dataset (TMRCA 254 kya; 95% CI 192-307 kya[14,17]) that we could not include due to sample availability issues. Nevertheless, our diverse samples cover genetic variation across a substantial period of modern human evolution (**Fig. 1b,d; Fig. S1**).

Constructing De Novo Assemblies

We employed the hybrid assembler Verkko[18] to generate Y chromosome assemblies including the ampliconic and heterochromatic regions (**Methods**). Verkko leverages the high accuracy of PacBio HiFi reads (99.5% base pair calling accuracy) with the length of Oxford Nanopore Long/Ultra Long Reads (median read length N50 134 kbp) to produce highly accurate and contiguous assemblies (**Table S2**). Using this approach, we generated high-quality (median QV 48; **Table S3**) whole-genome (median length 5.9 Gbp; **Table S4**) assemblies for 43 male samples. The chromosome Y sequences exhibit a high degree of completeness (median length 55.6 Mbp, 79% to 148% assembly length relative to GRCh38 Y; **Fig. 1; Fig. S2; Table S5**), contiguity (median NG50 9.6 Mbp, median LG50 2) and base-pair quality (median QV 46, **Table S3**). The Verkko assembly process was robust (sequence identity for NA19317/NA19347 pair of 99.9959%, **Fig. S3**; **Table S6; Supplementary Results '*De novo***

assembly evaluation') and generated the complete Y chromosome assembly, spanning from PAR1 to PAR2, for three individuals (HG01890 haplogroup A0b-L1038, HG02666 haplogroup A1a-M31, HG00358 haplogroup N1c-Z1940; **Figs. 1b, 2; Table S7**). This study presents the first dataset where deep-rooting African Y chromosomes have been contiguously assembled to high quality. These three samples are among nine samples with an increased HiFi coverage of at least 50× ("high-coverage samples", **Tables S1-S2**). The other six high-coverage samples were not completely assembled, indicating that increased HiFi coverage alone is not sufficient to ensure complete Y-chromosomal assembly (on average 2.5/24 Y-chromosomal subregions completely assembled, **Figs. 1c,e; Supplementary Results 'Effect of input read characteristics on assembly contiguity'**).

Following established procedures[11,19], we computed error rate estimates ranging from 0.04 errors per kbp assembled Y sequence up to 7.6 errors per kbp (**Table S8**; **Methods**). The upper range of the annotated errors is dominated by a few outlier samples as indicated by a median and mean of 0.77 and 1.28 (± 1.52 s.d.) errors per kbp, respectively. Although the error rate is increased for the lower-coverage assemblies, increasing the HiFi coverage beyond 50× has limited effect on the error rate (**Figs. S4-S5**).

We further annotated each of the Y-chromosomal assemblies with respect to the 24 Y-chromosomal subregions originally proposed by Skaletsky and colleagues (**Fig. 1a-c; Fig. S2; Table S9; Methods**)[3] and looked in more detail at the assembly outcome of each of these subregions. In addition to the three complete Y chromosomes, we have contiguously assembled the MSY (excluding Yq12) for 10/43 samples and the MSY (excluding Yq12 and the (peri-)centromeric region) for 17/43 samples (**Tables S7, S10-S11**). Overall, 17/24 subregions were contiguously assembled across 41/43 samples (**Figs. 1b-c; Fig. S2**).

Genomic and epigenetic variation of assembled Y chromosomes

The assembled Y chromosomes showed extensive variation both in size and structure (**Figs. 2a-c, 3a and 4; Figs. S6-S17; Methods**). The sizes of the Y assemblies ranged from 45.2 to 84.9 Mbp (mean 57.6 and median 55.7 Mbp, **Fig. S15**; **Table S5; Methods**), a 1.88-fold difference in size. The three complete Y chromosomes varied from 46.4 Mbp (our deepest-rooting A0b Y represented by HG01890) to 58.9 Mbp (HG00358; haplogroup N1c). In comparison, the T2T Y assembly from HG002 (haplogroup J1a) is 62.5 Mbp in size, primarily due to expansion of the Yq12 heterochromatic subregion. In contrast, the MSY (excluding Yq12 subregion) for the 10 contiguously assembled individuals and the T2T Y varies by less than 2 Mbp (from 24.1 to 26.1 Mbp, mean 25.4 Mbp) (**Tables S10-S11**).

Among the contiguously assembled Y-chromosomal subregions the largest variation in size was seen in the heterochromatic Yq12 (17.6 to 37.2 Mbp, mean 27.6 Mbp), the (peri-)centromeric region (2.0 to 3.3 Mbp, mean 2.7 Mbp) and the *DYZ19* repeat array (63.5 to 428 kbp, mean 305.4 kbp) (**Figs. 2a, 4f; Figs. S15-S21; Tables S10-S11**). Phylogenetically, a relatively shorter size of the *DYZ19* subregion was observed in ten samples representing the E1b1a1a haplogroup (from 217 to 247 kbp, mean 236 kbp), and an increased size (from 283 to 428 kbp, mean 369 kbp) among 17 phylogenetically-related haplogroup N, O, Q and R samples (**Figs. S17, S19-S22**).

The euchromatic regions show comparatively little variation in size (**Figs. 2a; Tables S10-S11)**. The exception is the ampliconic subregion 2 which contains a highly copy-number variable repeat array, composed of approximately 20.3 kbp long repeat units and each containing a copy of the *TSPY* (testis specific protein Y-linked 1) gene (**Fig. 3c**), which accounts for up to 467 kbp size difference between samples (**Fig. S23; Tables S12-S13; Methods**). The *TSPY* repeat array was also found to be shorter in haplogroup QR samples (from 567 to 648 kbp, mean 603 kbp) compared to the rest of the samples (from 465 to 932 kbp, mean 701 kbp) (**Figs. S17, S23**). Such phylogenetic consistency offers support to the high quality of our assemblies even across homogeneous tandem arrays, as more closely related Y chromosomes are expected to be more similar, and consequently allows investigation of mutational dynamics across well-defined timeframes.

We produced a comprehensive set of variant calls using contig length to span across GRCh38 euchromatin and heterochromatin (including 165 kbp of the Yq12 subregion present in GRCh38) and the fidelity of HiFi to resolve small variants and structural variants (SVs). In the MSY, we report on average 88 insertion and deletion structural variants (SVs, ≥ 50 bp), 3 large inversions (>1 kbp), 2,168 indels (< 50 bp), and 3,228 single nucleotide variants (SNVs) (**Fig. S24; Table S14; Methods**). Variants were merged across all 43 samples to produce a nonredundant callset of 413 SVs, 10 inversions, 16,216 indels, and 34,764 SNVs (**Tables S15-S19; Supplementary Results 'Orthogonal support to Y-chromosomal SVs'**). The average SNV density on the MSY is 0.09 SNV / kbp, which is significantly less than any other chromosome including chromosome X and the Y-chromosomal PARs ($p < 1.87 \times 10^{-17}$, Welch's t-test). The next lowest density is chromosome X (0.73 SNV / kbp), and all other chromosomes including the Y-chromosomal PAR average 1.42 SNV / kbp (1.94 – 1.62 SNV / kbp) (**Table S20**). Based on insertion calls (≥ 50 bp in size), we have identified from 30 to 140 kbp (mean 65 kbp; or an average of 16 kbp after exclusion of mobile elements and simple repeats) of inserted sequences per Y chromosome that is not present in the GRCh38 Y reference sequence (**Table S21**).

While we identified no SVs that directly intersect known exons, a 47 kbp duplication in 15 of 43 samples (35%) contains an additional copy of *RBMY1B*, a functional copy of RNA binding motif protein Y-linked family 1 (RBMY1). Duplicate copies contain two missense variants that do not appear to disrupt the gene. Previous studies have shown that fewer than six RBMY1 copies are associated with male infertility[20] and that low expression of *RBMY1B* in high-risk infertility cases can be upregulated by hormonal treatment with improved outcomes[21]. Taken together, this suggests that the observed *RBMY1B* duplication may be protective against male infertility. The results from variant calling overlap well with the gene annotation of the Y-chromosomal assemblies and showed that all protein-coding genes in the GRCh38 Y reference were present in the 43 Y chromosomes studied, except for 14 genes in PAR1, 1 gene in XDR1 and 1 gene in PAR2 in a total of 14 individuals, overlapping with poorly assembled regions in those individuals (**Tables S22-S26; Supplementary Results 'Gene annotation'**).

Additional large inversions were identified using Strand-seq and manual inspection of assembly alignments, which yielded a total of 14 inversions in the euchromatic regions of the Y chromosome and two inversions within the Yq12 subregion (**Figs. 3a, 4c; Figs. S25-S26; Tables S27-S28**; **Methods; Supplementary Results 'Y-chromosomal Inversions'**). Seven of these matched the 10 inversions identified by variant calling. We have defined the breakpoint regions for 8/14 of the euchromatic inversions to DNA intervals as small as 500 bp (**Fig. 3b**; **Fig. S26-S28; Table S29**; **Methods**). All of these inversions are flanked by highly similar (up to 99.97%) and large (up to 1.45 Mbp) inverted segmental duplications, and while determination of the molecular mechanism generating Y-chromosomal inversions remains challenging, most are likely a result of non-allelic homologous recombination (NAHR). 12/14 (85%) of the euchromatic inversions are recurrent, occurring from 2 to 13 times in the Y phylogeny and translate to an inversion rate estimate ranging from 3.68×10^{-5} (95% C.I.: 3.25 - 4.17×10^{-5}) to 2.39×10^{-4} (95% C.I.: 2.11 - 2.71×10^{-4}) per father-to-son Y transmission (**Table S27**), with the highest inversion recurrence seen among the 8 Y-chromosomal palindromes (called P1-P8, **Fig. 3a; Fig. S22**). Taken together, we calculate a rate of one recurrent inversion per 603 (95% C.I.: 533 - 684) father-to-son Y transmissions. The per site per generation rate estimates for 12 Y-chromosomal recurrent inversion are significantly higher (>2-fold difference between median estimates, two-tailed Mann-Whitney-Wilcoxon test, n=44, p-value<0.0001) than the rates previously estimated for 32 autosomal and X-chromosomal recurrent inversions[22].

There are two fixed inversions on either side of the Yq12 subregion (**Fig. 4c; Fig. S29; Table S28; Supplementary Results 'Y-chromosomal Inversions'**). The proximal inversion, which was observed in 10/11 individuals analyzed but completely deleted in HG01106, ranged from 358.9 to 820.7 kbp in size (mean 649.0 kbp) (**Table S28**). The distal inversion, on other hand, was observed in all 11 individuals and ranged from 259.5 to 641.4 kbp in size (mean 472.5 kbp). We resolved the exact breakpoints for these two inversions and found them to be identical among all individuals in which they were present. This suggests that the consistent presence of these two inversions at either end of the Yq12 subregion, may prevent unequal sister chromatid exchange from occurring, restricting expansion and contraction of the repeat units to the region between these two inversions.

We also identified 25 transposable elements in the 43 Y-chromosomal assemblies that are not present in the GRCh38 Y, including 18 *Alu* elements (4/18 within the Yq12 heterochromatic region) and 7 LINE-1 elements (no significant difference compared to the whole-genome distribution reported in[11]) (**Fig. 4f**; **Tables S30-S31**; **Methods; Supplementary Results 'Yq12 heterochromatic subregion'**). No novel SVA (SINE-VNTR-Alu) or HERV (human endogenous retroviruses) elements were observed within these 43 Y chromosome sequences. Three out of seven LINE-1 insertions are reported as full-length, including one with two intact ORFs within an intron of the *PCDH11Y* gene, suggesting that at least one potential retrotransposition-competent polymorphic LINE-1 element resides on the Y chromosome. Among the 25 identified transposable elements, six were shared between phylogenetically related individuals (including the *Alu* insertion known as the YAP marker fixed in all haplogroup DE Y chromosomes[23]), while 19 were found in single individuals (**Tables S30-S31**). For the *Alu* insertions in the Yq12 subregion, we noted that *Alu*Y (denoted as A1 and an A2 in **Fig. 4f**) insertions have occurred in the proximal and distal regions, respectively, at least 180,000 years ago, and have subsequently undergone expansions of the *Alu*-containing arrays. Based on these patterns, we can ascertain arrays and/or repeat units with the same *Alu* insertion are related to each other (**Fig. 4f**). While the intra-repeat array expansions may be caused by replication slippage, non-allelic homologous recombination may cause both intra- and inter-array expansion [24,25], although gene conversion can not be excluded.

Furthermore, the ONT data provides a means to explore the base level epigenetic landscape of the Y chromosome across these 43 individuals (**Fig. S30**). Here, we focused on DNA methylation at CpG sites, hereafter referred to as DNAme. We found 2,861 DNAme segments (**Methods**) that vary across these Y chromosomes (**Fig. S31a; Table S32**). 21% of the variation in DNAme levels is associated with haplogroups (Permanova p=0.003, (n=41), while only 4.8% of the expression levels (Permanova p=0.005 (n=210), leveraging the Geuvadis RNA-seq expression data[26]) is associated with haplogroups (**Methods**; **Supplemental Results 'Functional analysis'**). There is a significant association of Y haplogroup with both DNAme and gene expression particularly for five genes (*BCORP* (**Fig. S32**), *LINC00280*, *LOC100996911*, *PRKY*, *UTY*). Lastly, we find 194 Y-chromosomal genetic variants, including a 171 base-pair insertion (SV) and one inversion, that impact DNAme levels on chromosome Y (**Table S33**; **Supplementary Results 'Functional analysis'**). Taken together, this suggests that the genetic background, either on the Y chromosome or elsewhere in the genome, can impact the functional outcome (the epigenetic and transcriptional profiles) of specific genes on the Y chromosome.

Genetic variation and evolution of the Y-chromosomal heterochromatic regions

Variation in the size and structure of centromeric/pericentromeric repeat arrays. Our analysis of 21 chromosome Y centromeres (17 contiguously assembled centromeres, 3 centromeres with a single break within the *DYZ3* α-satellite array without unplaced contigs, and the T2T Y centromere) allowed the investigation of its diversity and evolution in detail (**Methods**). In general, the chromosome Y centromeres are composed of 171-bp *DYZ3* α-satellite repeat units[3], organized into a higher-order repeat (HOR) array, flanked on either side by short stretches of monomeric α-satellite. The monomeric α-satellite transitions into a unique sequence on the p-arm and an array of human satellite III (*HSat3*) on the q-arm.

Analysis of each α-satellite HOR array revealed that it ranges in size from 264 kbp to 1.361 Mbp (mean 667 kbp), with the largest arrays found in samples of African ancestry (mean 900 kbp) and smaller arrays found in samples of American, European, East Asian, or South Asian ancestry (means

664, 488, 264, and 565 kbp, respectively; **Figs. S17, S33; Table S11**; **Methods**)[27,28]. The *DYZ3* α-satellite HOR array is mostly composed of a 34-monomer repeating unit and is the most prevalent HOR type found in all samples (**Figs. 3e,f**). However, we identified two other HORs that were present at high frequency among the analyzed Y chromosomes: a 35-monomer HOR found in 14/21 samples and a 36-monomer HOR found in 11/21 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 (**Fig. S33)**. Analysis of the sequence composition of these HORs revealed that the 36-monomer HOR likely 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 (**Fig. 3f; Methods**).

The overall organization of the *DYZ3* α-satellite HOR array is similar to that found on other human chromosomes, with highly 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 to the block of diverged monomers juxtaposed to the q-arm.

Adjacent to the *DYZ3* α-satellite HOR array is an *HSat3* repeat array, which ranges in size from 372 to 488 kbp (mean 378 kbp), followed by a *DYZ17* repeat array, which ranges in size from 858 kbp to 1.740 Mbp (mean 1.085 Mbp). Comparison of the sizes of these three repeat arrays reveals no significant correlation among their sizes (**Fig. 3e; Figs. S34-S36; Table S11**).

The *DYZ19* repeat array is located on the long arm, flanked by X-degenerate regions (**Fig. 1a**) and composed of 125-bp repeat units (fragment of an LTR) in head-to-tail fashion. It is one of the subregions which has been completely assembled across all 43 Y chromosomes. It shows the highest variation in size compared to other chromosome Y subregions, ranging from 65 to 410 kbp (a 6.7-fold difference). The HG02492 individual (haplogroup J2a) with the smallest-sized *DYZ19* repeat array has an approximately 200 kbp deletion in this subregion (**Table S11**). In 43/44 Y chromosomes (including T2T Y), there appears to be evidence of at least two rounds of mutation/expansion (**Fig. 3d**, green and red colored blocks, respectively, **Figs. S19-21**) leading to directional homogenization of the central and distal parts of the region in all Y chromosomes. Finally, we have observed a recent ~80 kbp duplication event shared by the 11 phylogenetically related haplogroup QR samples (**Figs. S19-S21**) which must have occurred approximately 36,000 years ago (**Figs. 1b, S1**), resulting in substantially larger overall *DYZ19* subregion in these Y chromosomes.

Between the Yq11 euchromatin and the Yq12 heterochromatic subregion, lies the *DYZ18* subregion. We have found that this subregion comprises 3 distinct repeat arrays: a *DYZ18* repeat array, a 3.1-kbp repeat array and a 2.7-kbp repeat array (**Figs. S37-S44**). The 3.1-kbp repeat array appears to be 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-kbp repeat array appears to have originated from both the *DYZ18* (23% of the 2.7-kbp repeat unit shows 86.3% sequence identity to *DYZ18*) and *DYZ1* (77% of the 2.7-kbp repeat unit shows 97% sequence identity to *DYZ1*) repeat units (**Fig. S37**). All three repeat arrays (*DYZ18*, 3.1-kbp and 2.7-kbp) show a similar pattern and level of methylation to the *DYZ1* repeat arrays (**Fig. S45**), in that we observe constitutive hypermethylation.

The Yq12 subregion is composed of two alternating repeat arrays that expand and contract considerably but retain a 1:1 monomer repeat unit ratio. The Yq12 subregion is the most challenging portion of the Y chromosome to assemble contiguously due to its highly repetitive nature and size. In this study, we completely assembled the Yq12 subregion for six individuals (HG01890, HG02666, HG00358, HG01106, HG01952 and HG02011) and compared it to the Yq12 subregion of the T2T Y chromosome (**Figs. 1a, 4a,f; Tables S10-S11; Supplementary Results 'Yq12 heterochromatic subregion'**). The largest completely assembled Yq12 subregion is the 7th largest Yq12 subregion observed among the 44 samples analyzed (**Fig. S15b**). Therefore, the assembly outcome is likely determined not only by the size of the region. This subregion is composed of alternating arrays of repeat units: *DYZ1* and *DYZ2*[3,5,33–36]. The *DYZ1* repeat unit is approximately 3.5 kbp and consists mainly of simple repeats and pentameric satellite sequences, and it has been recently referred to as HSat3A6[4]. The *DYZ2* repeat (which has also been recently referred to as HSat1B[31]), is approximately 2.4 kbp and consists mainly of a tandemly repeated AT-rich simple repeat fused to a 5′ truncated *Alu* element followed by an HSATI satellite sequence (**Fig. S37**). The *DYZ1* repeat unit showed more variation in size (range from 1,165 to 3,608 bp, with 95% of all *DYZ1* repeat units longer than 3,000 bp with a mean length of 3,543 bp) compared to the *DYZ2* repeat units (range from 1,275 to 3,719 bp, with 93.7% of all *DYZ2* repeats 2,420 bp in size) (**Methods**).

The *DYZ1* repeat units are tandemly arranged into larger *DYZ1* repeat arrays as are the *DYZ2* repeat units (**Fig. 4**). The total number of *DYZ1* and *DYZ2* arrays (range from 34 to 86, mean: 61) were significantly positively correlated (Spearman Correlation=0.90, p-value=0.0056, n=7, alpha=0.05) with the total length of the analyzed Yq12 region (**Fig. S46**). Whereas the length of the individual *DYZ1* and *DYZ2* repeat arrays were found to be widely variable (**Fig. 4b; Fig. S47**). The *DYZ1* arrays were significantly longer (range from 50,420 to 3,599,754 bp, mean: 535,314 bp) than the *DYZ2* arrays (range from 11,215 to 2,202,896 bp, mean: 354,027 bp, two-tailed Mann-Whitney U test (n=7) p-value < 0.05) (**Fig. 4b**). The *DYZ1* and *DYZ2* arrays alternate with one another but interestingly the total number of *DYZ1* and *DYZ2* repeat units is nearly equal within each individual Y chromosome assembly (*DYZ1* to *DYZ2* ratio ranges from 0.88 to 1.33, mean: 1.09, SD: 0.17) (**Fig. 4b; Table S34**). From ONT data, we have observed a consistent hypermethylation of the *DYZ2* repeat arrays compared to 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 number of CpG dinucleotides (1 CpG/150 bp *DYZ2* versus 1 CpG/35 bp *DYZ1*) potentially explaining the marked DNA methylation differences (**Fig. S30**).

Sequence analysis of the repeat units in 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, when examining repeat units within a given repeat array, the repeat units near the middle of the repeat array show a higher level of sequence similarity to each other than to the repeat units at the distal regions of the repeat arrays (**Fig. 4d; Fig. S48**). This suggests that expansion and contraction tends to occur in the middle of the repeat arrays, homogenizing these units but allowing divergent repeat units to accumulate towards the periphery. Similarly, when looking at the entire Yq12 subregion, we observed that entire 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. 4e; Figs. S48-S49**). This observation is supported by results from the *DYZ2* repeat divergence analysis and the inter-*DYZ2* array profile comparison (**Methods**).

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 males from the 1000 Genomes Project dataset and a comprehensive analysis of their genetic and epigenetic variation and composition. While both the GRCh38 Y and the T2T Y represent relatively recently emerged (TMRCA 54.5 kya (95% HPD interval: 47.6 - 62.4 kya), **Fig. S1**) European Y lineages, 49% of our Y chromosomes carry African Y lineages, including two of the deepest rooting human Y lineages (A0b and A1a, TMRCA 183 kya (95% HPD interval: 160-209 kya)) which we have assembled contiguously allowing us to investigate how the Y chromosome has changed over 180,000 years of human evolution.

For the first time, we have been able to comprehensively and precisely examine the extent of genetic variation down to the nucleotide level across multiple human Y chromosomes. The male-specific region of the Y chromosome can be roughly divided into two portions: the euchromatic and the heterochromatic regions. Within the euchromatic region, the single-copy protein-coding Y-chromosomal genes, present in the GRCh38 Y reference sequence, are conserved in all 43 Y assemblies with few single nucleotide polymorphisms. 5/8 copy-number variable protein-coding gene families located in the ampliconic subregions showed variation in terms of copy number, with the highest variation determined in the *TSPY* gene family (from 24 to 40 copies, **Table S23**).

The euchromatic region harbors considerable structural variation across the 43 individuals. Most notably, we identified 14 inversions that affect half of the Y-chromosomal euchromatin, with only the most closely related pair of African Ys (from NA19317 and NA19347) showing the exact same inversion composition. We have been able to narrow down the breakpoints for all of the inversions, and for 8 of 14 inversions have refined the breakpoints down to a 500-bp region. The determination of the molecular mechanism causing the inversions remains challenging; however, the increased recurrent inversion rate on the Y chromosome compared to the rest of the human genome may be in part due to DNA double-strand breaks being repaired by intra-chromatid recombination[37]. Since inversions generally suppress interchromosomal recombination events[38], and the Y chromosome is paired with the X chromosome during meiosis, the widespread presence of inversions on the Y chromosome is consistent with limiting synaptonemal complexes to a small portion at the termini of the Y chromosome (i.e., PAR1 and PAR2). A neutral evolutionary view of the ubiquitousness of the inversions on the Y chromosome would be that inversions can arise anywhere in the genome but often lead to the formation of disadvantageous variants at chromosome regions that are normally involved in recombination. Therefore, most inversions would be lost by selection over time except for those in the non-recombining portions of the Y chromosome, where they are more tolerated and can therefore accumulate.

There are 4 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 4 heterochromatic subregions, the high content of the repetitive sequences and the high level of methylation (**Figs. S30, S45**) observed is consistent with the definition of heterochromatin. Furthermore, resolving the complete structural variation in the heterochromatic regions of the human Y chromosome provides novel molecular archeological evidence for evolutionary mechanisms. For example, in this study we have shown how the higher order structure at the centromeric region of the Y chromosome has evolved from an ancestral 36-mer HOR to a 34-mer HOR which predominates in the centromeres of current human males[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 following two *Alu* insertions, over 180,000 years ago, led to a subsequent *Alu*-*Alu* recombination that deleted the region in between and removing the diverged centromeric sequence block[41]. In the Yq12 subregion, there appear to be localized expansions and contractions of the *DYZ1* and *DYZ2* repeat units; however, evolutionary constraints seem to dictate a need to preserve the nearly 1:1 ratio of these two repeat units among all males studied by an unknown mechanism. These alternating repeat units are confined between two inversions that are fixed among modern-day humans.

In this study, we have fully sequenced and analyzed 43 diverse Y chromosomes and identified the full extent of variation of this chromosome across more than 180,000 years of human evolution. For the first time, sequence-level resolution across multiple human Y chromosomes has revealed new DNA sequences, new elements of conservation and provided molecular data that give us important insights into genomic stability and chromosomal integrity. 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 provides the starting point to associate Y-chromosomal sequences to specific human traits and more thoroughly study human evolution.

Main figures

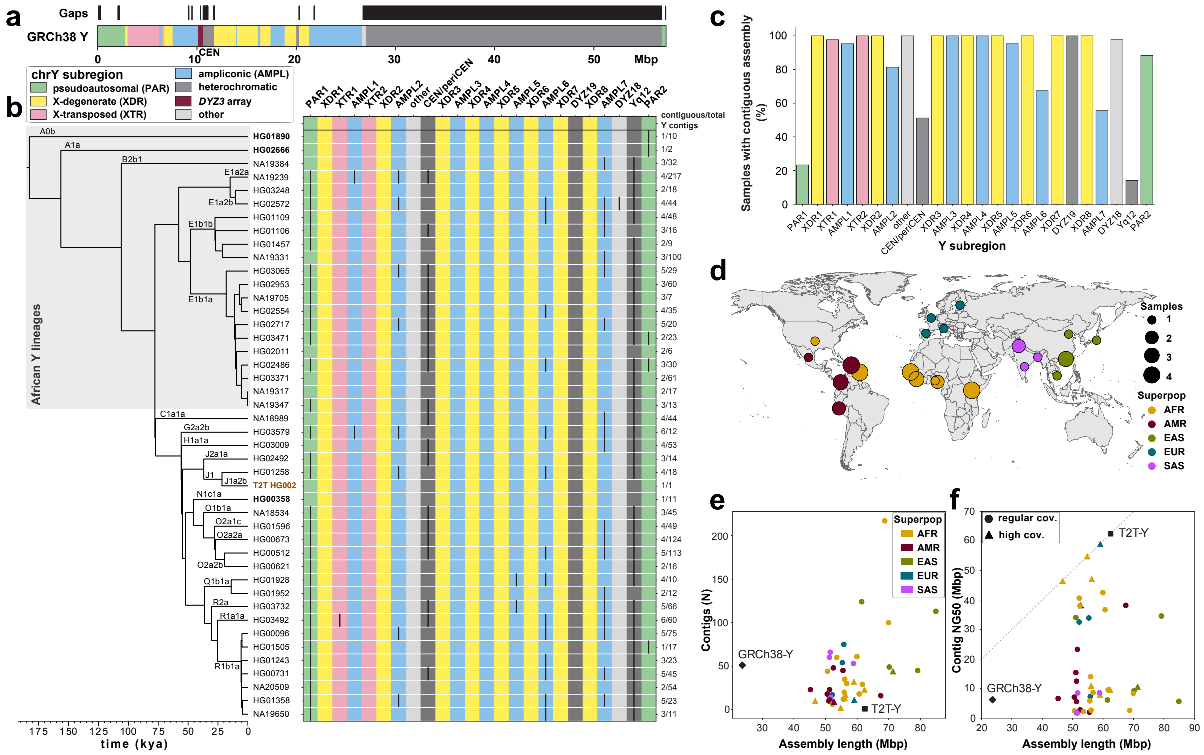

Figure 1. Assemblies capture more diversity and are more complete than the GRCh38 Y sequence.

a. Human Y chromosome structure based on the GRCh38 Y reference sequence.

b. Phylogenetic relationships (left) with haplogroup labels of the analyzed Y chromosomes with branch lengths drawn proportional to the estimated times between successive splits (see **Fig. S1** and **Table S1** for additional details). Summary of Y chromosome assembly completeness (right) with black lines representing non-contiguous assembly of that region (**Methods**). 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 - centromere - includes the *DYZ3* α-satellite array and the pericentromeric region. Three contiguously assembled Y chromosomes are in bold (assemblies for HG02666 and HG00358 are contiguous from telomere to telomere, while HG01890 assembly has a break approx. 100 kbp before the end of PAR2) and the T2T Y for HG002 in brown. Note - GRCh38 Y sequence mostly represents haplogroup R1b.

c. The proportion of contiguously assembled Y-chromosomal subregions across 43 samples.

d. Geographic origin and sample size of the included 1000 Genomes Project samples colored according to the continental groups (AFR, African; AMR, American; EUR, European; SAS, South Asian; EAS, East Asian).

e. Y-chromosomal assembly length vs number of Y contigs. Gap sequences (N's) were excluded from GRCh38.

f. Y-chromosomal assembly length vs Y contig NG50. High coverage defined as >50X genome-wide PacBio HiFi read depth. Gap sequences (N's) were excluded from GRCh38.

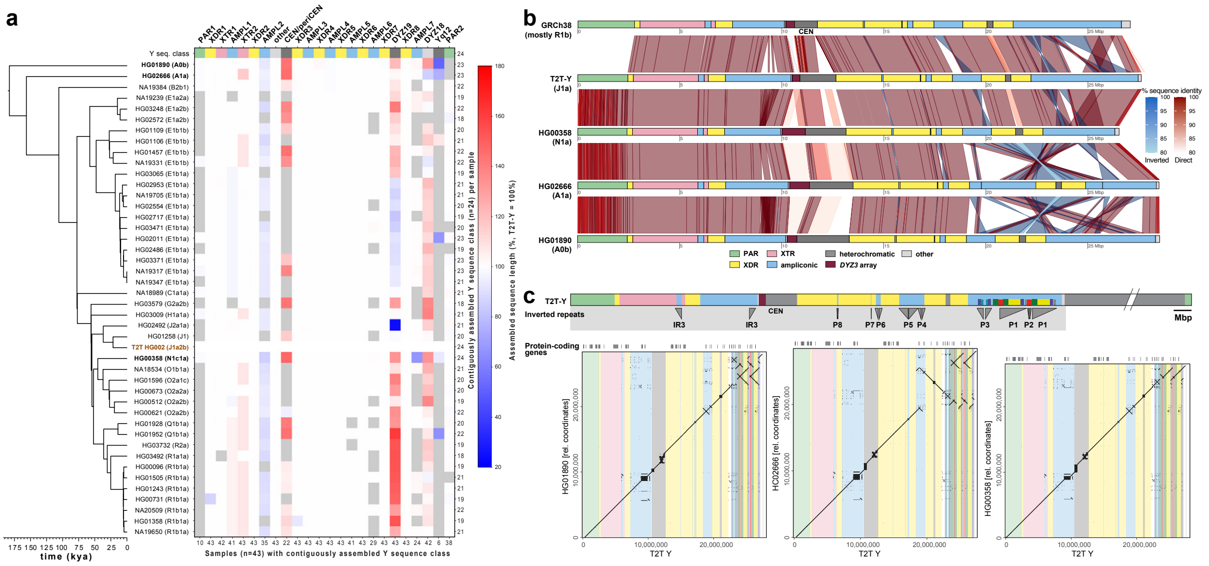

Figure 2. Size and structural variation of Y chromosomes.

a. Size variation of contiguously assembled Y-chromosomal subregions shown as a heatmap relative to the T2T Y size (as 100%). Boxes in gray indicate regions not contiguously assembled (**Methods**). Numbers on the bottom indicate contiguously assembled samples for each subregion out of a total of 43 samples, and numbers on the right indicate the contiguously assembled Y subregions out of 24 regions for each sample.

b. Comparison of the three contiguously assembled Y chromosomes to GRCh38 and the T2T Y (excluding Yq12 and PAR2 subregions).

c. Dotplots of three contiguously assembled Y chromosomes vs the T2T Y (excluding Yq12 and PAR2), annotated with Y subregions and segmental duplications in ampliconic subregion 7 (see **Fig. S22** for details).

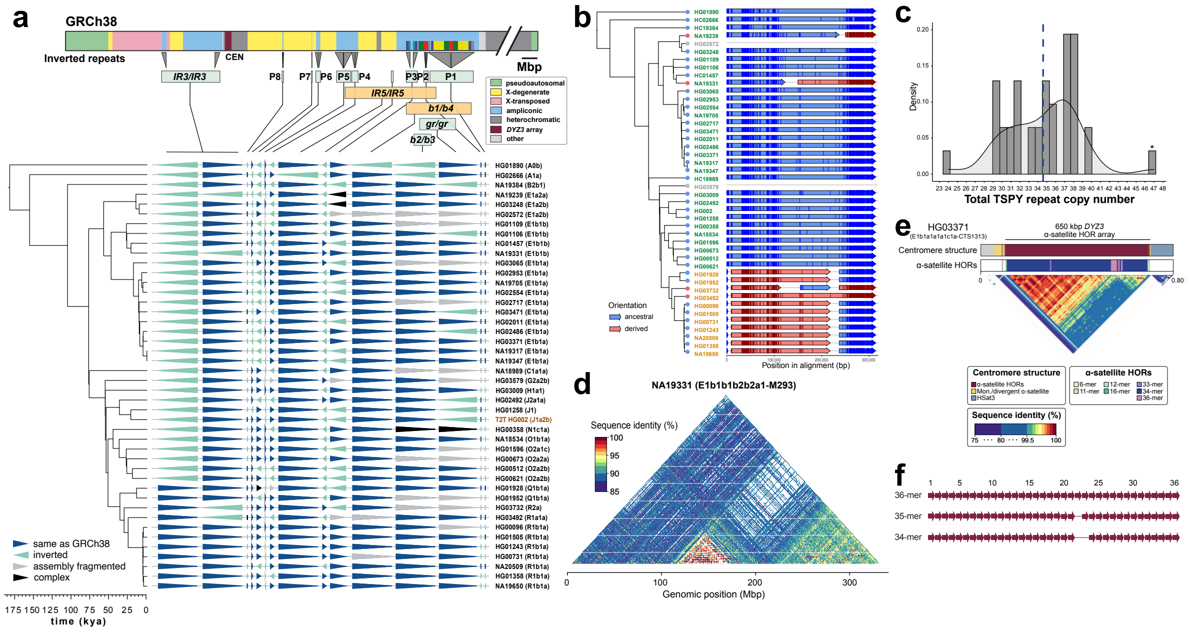

Figure 3. Characterization of large SVs.

a. Distribution of 14 euchromatic inversions in phylogenetic context, with the schematic of the GRCh38 Y structure shown above, annotated with Y subregions, inverted repeat locations and segmental duplications in ampliconic subregion 7 (see **Fig. S22** for details). Inverted segments are indicated below as green (recurrent) and orange (singleton events) boxes. **b.** Inversion breakpoint identification in the IR3 repeats. Samples highlighted in orange color have undergone two inversions (**Fig. S57, Supplementary Results 'Y-chromosomal Inversions'**). The red tip colors in the phylogenetic tree indicate samples which have undergone an additional inversion and therefore carry the region between IR3 repeats in inverted orientation compared to samples with blue tip. Informative PSV positions are shown as vertical lines with darker color in each of the arrows. The orange dotted line indicates the start of the unique 'spacer' region. Any information that is not available is indicated by gray. **c.** The total copy number distribution of the TSPY gene across 39 samples (T2T Y is marked with an asterisk). **d.** Sequence identity heatmap of the *DYZ19* repeat array from NA19331 (using 1 kbp window size) highlighting the higher sequence similarity within central and distal regions. **e.** Genetic landscape of the chromosome Y centromeric region from HG03371. This centromere harbors the newly identified ancestral 36-monomer HORs, from which the canonical 34-monomer HOR is derived. **f.** The 34-monomer α-satellite HOR was formed via two sequential steps in which a single α-satellite monomer residing at the 22nd position was deleted. The 34-monomer α-satellite HOR dominates all chromosome Y centromeres.

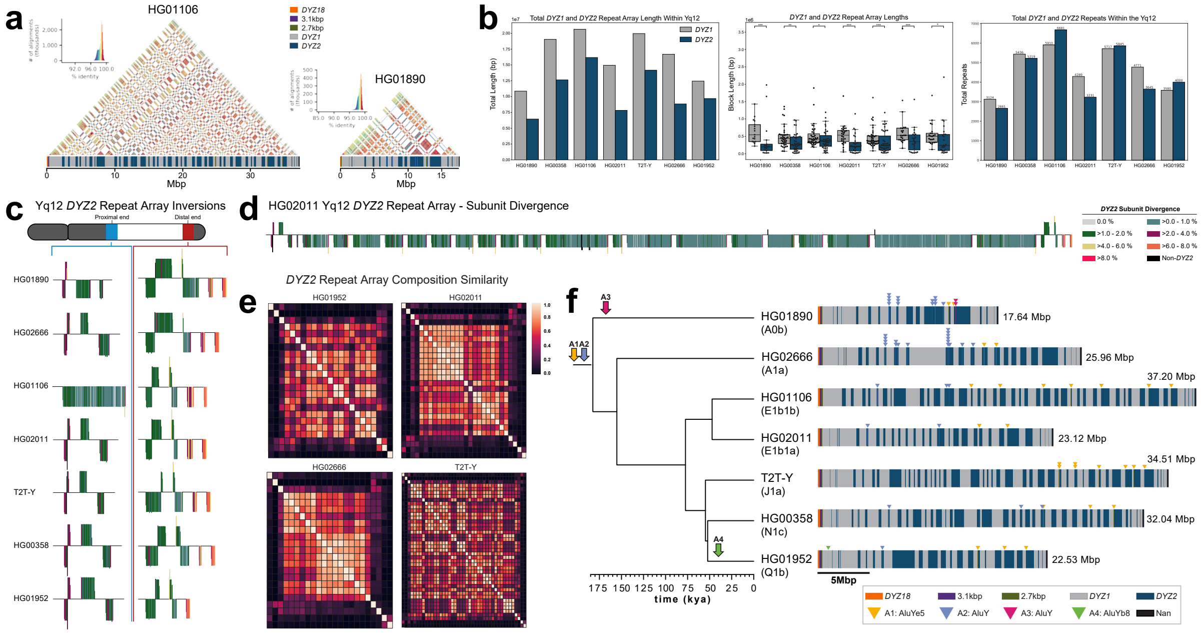

Figure 4. Yq12 heterochromatic region.

a. Yq12 heterochromatic subregion sequence identity heatmap in 5-kbp windows for HG01106 and HG01890 with repeat array annotations.

b. Bar plot of the total length of *DYZ1* and *DYZ2* repeat arrays for each sample (left), boxplots of individual array lengths (middle) and the total number of *DYZ1* and *DYZ2* repeat units (right) within contiguously assembled genomes. Black dots represent individual arrays and stars (*) denote a statistically significant difference between *DYZ1* and *DYZ2* array lengths (two-sided Mann-Whitney U test: p-value < 0.05, alpha=0.05, Methods).

c. *DYZ2* repeat array inversions in the proximal and distal ends of the Yq12 region. *DYZ2* repeats are colored based on their divergence estimate (see panel d) and visualized based on their orientation (sense - up, antisense - down).

d. Detailed representation of *DYZ2* subunit divergence estimates for HG02011. Length of each line is a function of the subunit length. Orientation (sense - up, antisense - down).

e. Heatmaps showing the inter-*DYZ2* repeat array subunit composition similarity within a sample. Similarity is calculated using the Bray-Curtis index (1 – Bray-Curtis Distance, 1.0 = exactly 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 highlighting four putative *Alu* insertions, their locations, and insertion occurrences across the seven complete genomes. The total size of Yq12 region is indicated on the right.

Methods

1. Sample selection

Samples were selected from the 1000 Genomes Project Diversity Panel[42] and at least one representative was selected from each of 26 populations (**Table S1**). 13/28 samples were included from the Human Genome Structural Variation Consortium (HGSVC) Phase 2 dataset, which was published previously[11]. In addition, for 15/28 samples data was newly generated as part of the HGSVC efforts (see the section 'Data production' for details'). We also included 15 samples from the Human Pangenome Reference Consortium (HPRC) (**Table S1**).

2. Data production

a. PacBio HiFi sequence production

University of Washington - Sample HG00731 data have been previously described[11]. 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 kbp. The sheared material was subjected to 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 kbp High Pass" and a cutoff of 14-15 kbp. Size-selected libraries were checked via fluorometric quantitation (Qubit) and pulse-field sizing (FEMTO Pulse). All cells were sequenced on a Sequel II instrument (PacBio) using 30-hour movie times using version 2.0 sequencing chemistry and 2-hour pre-extension. HiFi/CCS analysis was performed using SMRT Link v10.1 using an estimated read-quality value of 0.99.

The Jackson Laboratory - High-molecular-weight (HMW) DNA was extracted from 30M frozen pelleted cells using the Gentra Puregene extraction kit (Qiagen). Purified gDNA was assessed using fluorometric (Qubit, Thermo Fisher) assays for quantity and FEMTO Pulse (Agilent) for quality. For HiFi sequencing, samples exhibiting a mode size above 50 kbp were considered good candidates. Libraries were prepared using SMRTBell Express Template Prep Kit 2.0 (Pacbio). Briefly, 12 µl of DNA was first sheared using gTUBEs (Covaris) to target 15-18 kbp fragments. Two 5 µg of sheared DNA were used for each prep. DNA was treated to remove single strand overhangs, followed by DNA damage repair and end repair/ A-tailing. The DNA was then ligated 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 that has a size >10 kbp. The size selected and purified >10 kbp fraction of libraries were used for sequencing on Sequel II (Pacbio).

b. ONT-UL sequence production

University of Washington - High-molecular-weight (HMW) DNA was extracted from 2 aliquots of 30 M frozen pelleted cells using phenol-chloroform approach as described in[43]. Libraries were prepared using Ultra long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendation. Briefly, DNA from ~10M cells was incubated with 6 μl of fragmentation mix (FRA) at room temperature (RT) for 5 min and 75°C for 5 min. This was followed by an addition of 5 μl of adaptor (RAP-F) to the reaction mix and incubated for 30 min at RT. The libraries were cleaned up using Nanobind disks (Circulomics) and Long Fragment Buffer (LFB) (SQK-ULK001, ONT) and eluted in Elution Buffer (EB). Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT) for 96 hrs. A library was split into 3 loads, with each load going 24 hrs followed by a nuclease wash (EXP-WSH004, ONT) and subsequent reload.

The Jackson Laboratory - High-molecular-weight (HMW) DNA was extracted from 60 M frozen pelleted cells using phenol-chloroform approach as previously described [44]. Libraries were prepared using Ultra long DNA Sequencing Kit (SQK-ULK001, ONT) according to the manufacturer's recommendation. Briefly, 50ug of DNA was incubated with 6 μl of FRA at RT for 5 min and 75°C for 5 min. This was followed by an addition of 5 μl of adaptor (RAP-F) to the reaction mix and incubated for 30 min at RT. The libraries were cleaned up using Nanodisks (Circulomics) and eluted in EB. Libraries were sequenced on the flow cell R9.4.1 (FLO-PRO002, ONT) on a PromethION (ONT) for 96 hrs. A library was generally split into 3 loads with each loaded at an interval of about 24 hrs or when pore activity dropped to 20%. A nuclease wash was performed using Flow Cell Wash Kit (EXP-WSH004) between each subsequent load.

c. Bionano optical genome maps production

Optical mapping data were generated at Bionano Genomics, San Diego, USA. Lymphoblastoid cell lines were obtained from Coriell Cell Repositories and grown in RPMI 1640 media with 15% FBS, supplemented with L-glutamine and penicillin/streptomycin, at 37°C and 5% CO_2. Ultra-high-molecular-weight DNA was extracted according to the Bionano Prep Cell Culture DNA Isolation Protocol(Document number 30026, revision F) using a Bionano SP Blood & Cell DNA Isolation Kit (Part #80030). In short, 1.5 M 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 via 1hr end-over-end rotation at 15 rpm, followed by an overnight rest at RT. Isolated DNA was fluorescently tagged at motif CTTAAG by the enzyme DLE-1 and counter-stained using a Bionano Prep™ DNA Labeling Kit – DLS (catalog # 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 (Part #60325) and Instrument Control Software (ICS) version 4.9.19316.1.

d. Strand-seq data generation and data processing

Strand-seq data were generated at EMBL and the protocol is as follows. EBV-transformed lymphoblastoid cell lines from the 1000 Genomes Project (Coriell Institute; **Table S1**) were cultured in BrdU (100 uM final concentration; Sigma, B9285) for 18 or 24 hrs, and single isolated nuclei (0.1% NP-40 substitute lysis buffer [45] 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 0-cell negative control were deposited. Strand-specific single-cell DNA sequencing libraries were generated using the previously described Strand-seq protocol[45,46] and automated on the Beckman Coulter Biomek FX P liquid handling robotic system[47]. Following 15 rounds of PCR amplification, 288 individually barcoded libraries (amounting to three 96-well plates) were pooled for sequencing on the Illumina NextSeq500 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 (version 0.7.15-0.7.17) for standard library selection. Aligned reads were sorted by genomic position using SAMtools (version 1.10) and duplicate reads were marked using sambamba (version 1.0). Low-quality libraries were excluded from future analyses if they showed low read counts (<50 reads per Mbp), 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[45]. Aligned BAM files were used for inversion discovery as described in[22].

e. Hi-C data production

Lymphoblastoid cell lines were obtained from Coriell Cell Repositories and cultured in RPMI 1640 supplemented with 15% FBS. Cells were maintained at 37°C in an atmosphere containing 5% CO_2. Hi-C libraries using 1.5 M human cells as input were generated with Proximo Hi-C kits v4.0 (Phase Genomics, Seattle, WA) following the manufacturer’s protocol with the following modification: in brief, cells were crosslinked, quenched, lysed sequentially with Lysis Buffers 1 and 2, and liberated chromatin immobilized on magnetic recovery beads. A 4-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. Following fragmentation and fill-in with biotinylated nucleotides, fragmented chromatin was proximity ligated for 4 hrs at 25°C. Crosslinks were then reversed, DNA purified and biotinylated junctions 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 an Illumina Novaseq 6000 instrument using 2x150 bp cycles.

f. RNAseq data production

Total RNA of cell pellets were isolated using QIAGEN RNeasy Mini Kit according to the manufacturer's instructions. Briefly, each cell pellet (10 M cells) was homogenized and lysed in Buffer RLT Plus, supplemented with 1% β-mercaptoethanol. The lysate-containing RNA was purified using an RNeasy spin column, followed by an in-column DNase I treatment by incubating for 10 min at RT, and then washed. Finally, total RNA was eluted in 50 uL 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 a second strand synthesis and A-tailing at 16°C for 30 min, 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 real-time qPCR (Thermo Fisher). Sequencing was performed on an Illumina NovaSeq platform generating paired end reads of 100 bp at The Jackson Laboratory for Genomic Medicine.

g. Iso-seq data production

Iso-seq data were generated at The Jackson Laboratory. Total RNA was extracted from 10 M human cell pellets. 300 ng total RNA were used to prepare Iso-seq libraries according to Iso-seq Express Template Preparation (Pacbio). First, full-length cDNA was generated using NEBNext Single Cell/ Low Input cDNA synthesis and Amplification Module in combination with 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 a DNA damage repair, end repair, and A-tailing and finally ligated with Overhang Barcoded Adapters. Libraries were sequenced on Pacbio Sequel II. Iso-seq reads were processed with default parameters using the PacBio Iso-seq3 pipeline.

3. Construction and dating of Y phylogeny

The genotypes were jointly called from the 1000 Genomes Project Illumina high-coverage data from [48] using the ~10.4 Mbp of chromosome Y sequence previously defined as accessible to short-read sequencing[49]. BCFtools (v1.9)[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. Additionally, 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, i.e., missing in more than 3 out of 44 samples, were removed using VCFtools (v0.1.16)[52]. After filtering, a total of 10,406,108 sites remained, including 12,880 variant sites. Since Illumina short-read data was 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, v15.73, accessed in August 2021). We used the coalescence-based method implemented in BEAST (v1.10.4[53] to estimate the ages of internal nodes in the Y phylogeny. A starting maximum likelihood phylogenetic tree for BEAST was constructed with RAxML (v8.2.10[54]) with the GTRGAMMA substitution model. Markov chain Monte Carlo samples were based on 200 million iterations, logging every 1000 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% confidence interval: 0.67×10^{-9}– 0.86×10^{-9}) single-nucleotide mutations per bp per year was used[55]. A prior with a normal distribution based on the 95% confidence interval of the substitution rate was applied. A summary tree was produced using TreeAnnotator (v1.10.4) and visualized using the FigTree software (v1.4.4).

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 ya (95% HPD interval: 0 - 500 ya).

A separate phylogeny (see **Fig. 4f**) was reconstructed using seven samples (HG01890, HG02666, HG01106, HG02011, T2T Y from NA24385/HG002, HG00358 and HG01952) with contiguously assembled Yq12 region following 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.

4. *De novo* Assembly Generation

a. Reference assemblies

We used the GRCh38 (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 unlocalised GRCh38 contig “chrY_KI270740v1_random” (37,240 bp, composed of 289 *DYZ19* primary repeat units) in any of the analyses presented in this study.

b. Constructing *de novo* assemblies

All 28 HGSVC and 15 HPRC samples were processed with the same Snakemake[56] 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 v1.0[18] with default parameters, combining all available PacBio HiFi and Oxford Nanopore data per sample to create a whole-genome assembly:

```
verkko -d work_dir/ --hifi {hifi_reads} --nano {ont_reads}
```

We note here 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 v1.0 resulting in an empty output sequence for contig "0000598". The Verkko development team suggested removing the affected record, i.e. 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 50X HiFi input coverage (termed high-coverage samples, **Tables S1-S2**), we generated alternative assemblies using hifiasm v0.16.1-r375[57] for quality control purposes. Hifiasm was executed with default parameters using only HiFi reads as input, thus producing partially phased output assemblies "hap1" and "hap2" (cf. 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, i.e., the assembled Y sequence contigs have to be identified in a subsequent process that we implemented as follows.

We employed 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, i.e. to either chrY or chrX. The sequence alignments were produced with minimap2 v2.24 [58]:

```
minimap2 -t {threads} -x asm20 -Y --secondary=yes -N 1 --cs -c --paf-
no-hit
```

Rule 2: the assembled sequence has mixed primary alignments, i.e. not only to the target sequence of interest, but exhibits Y-specific sequence motif hits for any of the following motifs: *DYZ1*, *DYZ18* and the secondary repeat unit of *DYZ3* from[3]. The motif hits were identified with HMMER v3.3.2dev (commit hash #016cba0)[59]:

```
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, i.e. not only to the target sequence of interest, but exhibits more than 300 hits for the Y-unspecific repeat unit *DYZ2* (see Section '**Yq12 *DYZ2* Consensus and Divergence**' 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 cutoff of 1.6e-15 and a score threshold of 1700.

Rule 4: the assembled sequence has no alignment to the chrY reference sequence, but exhibits Y-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 chrX and chrY.

After identification of all assembled chrY and chrX 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.

c. Assembly evaluation and validation

Error detection in *de novo* assemblies

Following established procedures[11,18], we implemented two independent approaches to identify regions of putative misassemblies for all 43 samples. First, we used VerityMap (v2.1.1-alpha-dev #8d241f4)[19] that generates and processes read-to-assembly alignments to flag regions in the assemblies that exhibit spurious signal, i.e., 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 (v1.3.0)[60] and the PEPPER-Margin-DeepVariant pipeline (v0.8, DeepVariant v1.3.0, [61]) to identify heterozygous (HET) 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 HET SNVs remaining after quality filtering (bcftools v1.15 "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
```

The output of all error detection steps was merged using custom code (see "Code Availability" statement in main text; **Table S8**).

Assembly QV estimates were produced with yak v0.1 (github.com/lh3/yak) following the examples in its documentation (see readme in referenced repository). The QV estimation process requires an independent sequence data source to derive a (sample-specific) reference k-mer set to compare the k-mer content of the assembly. 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 Y assemblies or to the T2T-Y, which we added during the alignment step to capture reads that would align to Y sequences missing from our assemblies.

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 (v3.5.1) 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_saphy
r_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 (v3.5.1)); 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_conf
ig.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_sa
phyr_human.xml
```

Inconsistencies between optical mapping data and Verkko assemblies were identified based on variant calls from approach 1 using "exp_refineFinal1_merged_filter_inversions.smap" output file. Variants were filtered out based on the following criteria: a) variant size smaller than 500 base pairs; b) variants labeled as "heterozygous"; c) 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); d) 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 "conflicts_cut_status.txt" output file from approach 3 were used to evaluate if inconsistencies identified above based on optical mapping variant calls overlap with conflict sites (i.e. sites identified by hybrid scaffolding pipeline representing inconsistencies between sequencing and optical mapping data) (**Table S35**). 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 kbp bins and counting the number single molecules covering each 10 kbp window (**Table S36**).

d. *De novo* assembly annotation

Annotation of Y-chromosomal subregion

The 24 Y-chromosomal subregion coordinates (**Table S9**) relative to the GRCh38 reference sequence were obtained from[7]. Since Skov et al. 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 pseudoautosomal regions we used the coordinates from the UCSC Genome Browser for GRCh38.p13 as they slightly differed. Additionally, Y-chromosomal amplicon start and end coordinates were edited according to more recent annotations from[62], and the locations of *DYZ19* and *DYZ18* repeat arrays were adjusted based on the identification of their locations using HMMER3 (v3.3.2)[63] with the respective repeat unit consensus sequences from[3].

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 (v2.24, 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 (**Table S9**). 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[3]. Only tandemly organized matches with HMMER3 score thresholds higher than 1700 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 as 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 PAR regions where they were defined as >95% length of the T2T Y PAR regions 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 kbp unplaced PAR2 contig aligning best to a central region of PAR2). The assembly of HG01890 however has a break approximately 100 kbp before the end of PAR2. Assembly of PAR1 remains especially challenging due to its sequence composition and sequencing biases[8,10], and among our samples was contiguously assembled for 10/43 samples, while PAR2 was contiguously assembled for 39/43 samples.

Annotation of centromeric and pericentromeric regions

To annotate the centromeric regions, we first ran RepeatMasker (v4.1.0) 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 (v3.3.2; https://github.com/fedorrik/HumAS-HMMER_for_AnVIL) to identify the location of α-satellite higher-order repeats (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 (v4.1.0) 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, we used the α-satellite HOR annotations generated via HumAS-HMMER (v3.3.2; described above) to determine the location of *DYZ3* α-satellite HORs, focusing on only those HORs annotated as "live" (e.g. S4CYH1L). Live HORs are those that have a clear higher-order pattern and are highly (>90%) homogenous [64]. This analysis was conducted on 21 centromeres (including the T2T Y), excluding 5/26 samples (NA19384, HG01457, HG01890, NA19317, NA19331), where, despite a contiguously assembled pericentromeric subregion, the assembly contained unplaced centromeric contig(s).

To annotate the human satellite III (*HSat3*) and *DYZ17* arrays within the pericentromere, we ran StringDecomposer (v1.0.0) on each assembly centromeric contig using the *HSat3* and *DYZ17* consensus sequences described in Altemose, 2022, Seminars in Cell and Developmental Biology [65] and available at the following URL: 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.

5. Downstream analysis

a. Effect of input read depth on assembly contiguity

We explored 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 v1.1.1, see "Code Availability" statement in main text). The models were trained following standard procedures with 5-fold nested cross-validation (see scikit-learn documentation for "ElasticNetCV"). We note that we did not use the haplogroup information due 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 kbp in length, i.e., the so-called ultralong fraction of ONT reads; **Table S38**) as model features and assembly contig NG50, assembly length or number of assembled contigs as target variable.

b. 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 kbp 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 analyzed these sequences to compare the GC content, segmental duplication content, and Y subregion. Read depth for each bin was calculated using mosdepth[66] and the flags `-n -x`. GC content for each bin was calculated using BedTools `nuc` function[67]. Segmental duplication locations for GRCh38 Y were obtained from the UCSC genome browser, and for the CHM13 plus T2T Y from[2]. Y-chromosomal subregion locations were determined as described in Methods section '*De novo* assembly annotation with Y-chromosomal subregions'. The bin read depth and GC content statistics were merged into matrices and visualized using *matplotlib* and *seaborn*[68,69].

c. Effect of read depth on assembly contiguity

Read depth statistics of both PacBio HiFi and ONT raw reads as mapped to the *de novo* assemblies were calculated using samtools bedcov (version 1.15.1) [50]. We investigated normality through histograms and qq-plotting of the read depth distribution, and proceeded to use the mean read depth in our analyses. Average read depth for the whole Y- chromosomal assembly was regressed against contig N50 and L50, total contig number, total assembly length, and largest contig length. Regressions were calculated using the OLS function in *statsmodels*, and visualized using *matplotlib* and *seaborn*[68–70].

d. Comparison of assembled Y subregion sizes across samples

Sizes for each chromosome's (peri-)centromeric regions were obtained as described in Methods section 'Annotation of pericentromeric regions'. 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 heatmap, 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*[68].

e. 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[71] with sequence identity threshold of 80% (95% threshold was used for PAR1 subregion) (**Fig. 2b**) and excluding non-specific alignments (i.e. showing alignments between different Y subregions), followed by visualization with genoPlotR (0.8.11)[72]. 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 in queries.

```
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.

f. Sequence identity heatmaps

Sequence identity within repeat arrays was investigated by running StainedGlass[73]. For the centromeric regions, StainedGlass was run with the following configuration: `window = 5785 and mm_f = 30000`. We adjusted the color scale in the resulting plot using a custom R script that redefines the breaks in the histogram and its corresponding colors. This script is publicly available here: https://eichlerlab.gs.washington.edu/help/glogsdon/Shared_with_Pille/StainedGlass_adjustedScale.R. The command used to generate the new plots is: `StainedGlass_adjustedScale.R -b {output_bed} -p {plot_prefix}. For the` *`DYZ19`* `repeat array, window = 1000 and mm_f = 10000` were used, 5 kbp of flanking sequence was included from both sides, followed by adjustment of color 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 kbp 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 kbp in size) which was not included. For the Yq11/Yq12 transition region, 100 kbp proximal to the *DYZ18* repeat and 100 kbp of the first *DYZ1* repeat array was included in the StainedGlass runs, using `window = 2000 and mm_f = 10000`.

g. Dotplot generation

Dotplot visualizations were created using the NAHRwhals package version 0.9 which provides visualization utilities and a custom pipeline for pairwise sequence alignment based on minimap2 (v2.24) . Briefly, NAHRwhals initiates pairwise alignments by splitting long sequences into chunks of 1 - 10 kbp which are then aligned to the target sequence separately, enhancing the capacity of minimap2 to correctly capture inverted or repetitive sequence alignments. Subsequently, alignment pairs are concatenated whenever the endpoint of one alignment falls in close proximity to the startpoint of another (base pair distance cutoff: 5% of the chunk length). Pairwise alignment dotplots 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 https://github.com/WHops/nahrchainer, and dotplot views of selected regions can be accessed at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/20221020_Dotplots_chrY_ASMS

h. Inversion analyses

Inversion calling using Strand-seq data

The inversion calling from Strand-seq data, available for 30/43 samples and the T2T Y, using both the GRCh38 and the T2T Y sequences as references was performed as described previously[22].

Note on 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[3] 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 on **Fig. 3a**, inverted relative to GRCh38).

Inversion calling from the *de novo* assemblies

In order to determine the inversions from the *de novo* assemblies, we aligned the Y-chromosomal repeat units/segmental duplications as published by [62] to the *de novo* assemblies as described above (see Section: 'Annotation with Y-chromosomal subregions'). Inverted alignment orientation of the unique sequences flanked by repeat units/segmental duplications 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 dotplots generated against both GRCh38 and T2T Y sequences (see Methods section: Dotplot generation), followed by merging with the Strand-seq calls (**Table S26**).

Inversion rate estimation

In order 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 males and Y chromosomal substitution rate from above (see Methods section 'Construction and dating of Y phylogeny') 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[74]. 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 focussed 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 (i.e. proximal/distal) of paralogous sequence variants (PSVs) across the segmental duplications 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 (v7.487)[75,76] with default parameters. From the alignment, we only selected informative sites (i.e. not identical across all repeats and samples), excluding singletons and removing sites within repetitive or poorly aligned regions as determined by Tandem Repeat Finder (v4.09.1)[77] and Gblocks (v0.91b)[78], 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 a) that the reference (i.e. same as GRCh38 Y) state was ancestral and b) that the inverted (i.e. inverted compared to 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 (17/21) and non-African Y lineages have more reference states (17/23)), 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 (i.e. with the single ~20.3 kbp 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 (e.g. 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 where phase switching occurred between two segments, and the coordinate was converted to the T2T Y coordinate based on the multiple sequence alignment 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.

i. Variant calling

Variant calling using *de novo* assemblies

Variants were called from assembly contigs using PAV (v2.1.0)[11] with default parameters using minimap2 (v2.17) 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 (v2.1.0) using LRA[79] alignments (commit e20e67) with assemblies, PBSV (v2.8.0) (https://github.com/PacificBiosciences/pbsv) with PacBio HiFi reads, SVIM-asm (v1.0.2)[80] with assemblies, Sniffles (v2.0.7)[81] with PacBio HiFi and ONT, DeepVariant (v1.1.0)[60,82] with PacBio HiFi, Clair3 (v0.1.12)[83] with ONT, CuteSV (v2.0.1)[84] with ONT, and LongShot (v0.4.5)[85] with ONT. A validation approach based on the subseq command was used to search for raw-read support in PacBio HiFi and ONT[11].

A merged callset was created from the PAV calls with minimap2 alignments across all samples with SV-Pop[11,86] 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 & position, then 50% reciprocal overlap, then 50% overlap by size and within 200 bp), "`nr::exact:ro(0.2)`" for inversions (Exact size & 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 (i.e. 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 dotplots.

To search for likely duplications within insertion calls, insertion sequences were re-mapped to the reference with minimap2 (v2.17) with parameters "`-x asm20 -H --secondary=no -r 2k -Y -a --eqx -L -t 4`".

To evaluate whether the identified additional *RBMY1B* copies were functional the insertion sequence containing the *RBMY1B* duplicate copy was aligned to the reference with minimap2 using default parameters. Small variants between the duplicate and reference copy were identified using the alignment (CIGAR string parsing). A VCF was generated for these variants and run with VEP (version 107). Variants VEP annotated with MODIFIER were discarded.

In order to compare the SNV densities between chromosomes the following approach was used: for autosomes and chrX, we obtained a set of filtered SNV calls and PAV callable regions from a set of 32 samples derived from long-read phased assemblies[11]. We used a similar callset generated for chrY from this study and separated the psedoautosomal regions (PAR), which recombine with chrX during meiosis, and the MSY, which does not recombine with chrX. For each chromosome, we generated globally callable loci by taking a union of all the PAV callable loci (regions where variants could be called). To further guarantee that SNVs were not a result of alignment artifacts, we removed simple repeats, segmental duplications, N-gaps, and centromeres using UCSC browser tracks. We also excluded unreliable regions used to filter the Ebert callset (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/filter/20210127_LowConfidenceFilter/), which is mostly covered by the other region filters from UCSC tracks. We then counted the number of SNVs in these callable regions and divided by the callable size in kbp to obtain the number of SNVs per kbp. We tested the significance of the difference in means using Welch's t-test by comparing the MSY to all other regions (including PAR and chrX) and by comparing chrX to all other regions (including PAR and MSY). The maximum p-value reported for each of these two tests was Bonferroni-corrected by multiplying the p-value by the number of other chromosomes tested.

Validation of large SVs using optical mapping data

Orthogonal support for merged PAV calls were evaluated by using optical mapping data (**Table S39**). Molecule support was evaluated using local *de novo* assembly maps which aligned to GRCh38 reference assembly. This evaluation included all 10 called inversions, and insertions and deletions at least 5 kbp or larger in size. Although variants <5 kbp could be resolved by optical mapping technique, 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 (v1.7). 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, known to be highly variable in copy number [87], 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 (v7.487)[75,76] with default parameters. A consensus sequence was generated using EMBOSS cons (v6.6.0.0) command line version with default parameters, followed by manual editing to replace sites defined as 'N's 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 v3.3.2[63], 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 (**Table S13**).

j. Mobile element insertion analysis

Mobile element insertion (MEI) calling

We leveraged an enhanced version of PALMER (Pre-mAsking Long reads for Mobile Element inseRtion,[88]) to detect MEIs across the long-read sequences. Reference-aligned (to both GRCh38 Y and T2T Y) Y contigs from Verkko assembly were used as input. Putative insertion sequences of non-reference repetitive elements (L1s, Alus or SVAs) were identified based on a library of mobile element sequences after a pre-masking process. PALMER then identifies the hallmarks of retrotransposition events for the putative insertion signals, including TSD motifs, transductions, and poly(A) tract sequences, and etc. Further manual inspection was carried out based on the information of large inversions, structural variations, heterochromatic regions, and concordance with the Y phylogeny. Low confidence calls overlapping with large SVs or discordant with the Y phylogeny were excluded, and high confidence calls were annotated with further genomic content details.

In order to compare the ratios of non-reference mobile element insertions 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 Mbp was used, while the total GRCh38 reference sequence length is 3.2 Gbp. At the whole genome level, this results in a ratio for non-reference Alu of 0.459 per Mbp (1470/3.2 Gbp) and for non-reference LINE-1 of 0.066 per Mbp (210/3.2 Gbp)[11]. In chromosome Y, the ratio for non-reference Alu and LINE-1 is 0.315 per Mbp (18/57.2 Mbp) and 0.122 per Mbp (7/57.2 Mbp), respectively. The ratios within the MEI category were compared using the Chi-square test.

k. Gene annotation

Genome Annotation - liftoff

Genome annotations of chromosome Y assemblies were obtained by using T2T Y and GRCh38 Y gff annotation files using liftoff [89].

```
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 which were labeled as "protein_coding" from the GENCODEv41 annotation file (i.e., a total of 63 protein-coding genes).

l. Methylation analysis

Read-level CpG DNA-methylation (DNAme) likelihood ratios were estimated using nanopolish version 0.11.1. Nanopolish (https://github.com/jts/nanopolish) was run on the alignment to GRCh38, for the three complete assemblies (HG00358, HG01890, HG02666) we additionally mapped the reads back to the assembled Y chromosomes and performed a separate nanopolish run. Based on the GRCh38 mappings we first performed sample quality control (QC). We find four samples with genome wide methylation levels below 50%, which were QCed out. 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 QC we leverage pycoMeth to *de novo* identify interesting methylation segments on chromosome Y. pycoMeth (version 2.2) [90] Meth_Seg is a Bayesian changepoint-detection algorithm that determines regions with consistent methylation rate from the read-level methylation predictions. Over the 139 QCed flowcells of the 41 samples, we find 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 leveraged the permanova test, implemented in the R package vegan [91,92], to identify the impact of "aggregated" haplotype group on the DNAme levels over the segments. Because of the low sample numbers per haplotype group we aggregated haplogroups to meta groups based on 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. Additionally we tested for individual segments with differential meta-haplogroup methylation differences using the Kruskal Wallis test. Regions with FDR<=0.2, as derived from the Benjamini-Hochberg procedure, are reported as DMRs. For follow up tests on the regions that are found to be significantly different from the Kruskal Wallis test we used a one versus all strategy leveraging a Mann–Whitney U test.

Next to assessing the effects of haplogroup to DNAme we also tested for local methylation Quantitative Trait Loci (*cis*-meQTL) using limix-QTL [93,94]. Specifically, we tested the impact of the genetic variants called on GRCh38 (see Methods **“Variant calling using *de novo* assemblies”**), versus the DNAme levels in the 2,861 segments discovered by pycoMeth. For this we leveraged an LMM implemented in limixQTL, methylation levels were arcsin transformed and we leveraged population as a random effect term. Variants with a MAF >10% and a call rate >90%, leaving 11,226 variants to be tested. 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 gene and P values were corrected for this estimated number of tests using the Bonferroni procedure. To account for the number of tested segments we leveraged a Benjamini-Hochberg procedure over the top variants per segment to correct for this.

m. Expression analysis

Gene expression quantification for the HGSVC [11] and the Geuvadis dataset [26] was derived from the [11]. In short, RNA-seq QC was conducted using Trim Galore! (v0.6.5) [95] and reads were mapped to GRCh38 using STAR (v2.7.5a) [96], followed by gene expression quantification using FeatureCounts (v2) [97]. After quality control gene expression data is available for 210 Geuvaids males and 21 HGSVC males.

As with the DNAme analysis we leveraged 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 signal character haplotype groups, specifically “E”, “G”, “I”, “J”,”N”,”R” and “T”. Additionally we tested for single gene effects using the Kruskal Wallis test, and the Mann–Whitney U test. For *BCORP1* we leveraged the HGSVC expression data to assess the link between DNAme and expression variation.

n. Iso-Seq data analysis

Iso-Seq reads were aligned independently with minimap v2.24 (-ax splice:hq -f 1000) to each chrY Verkko assembly, as well as the T2T v2.0 reference including HG002 chrY, and GRCh38. Read alignments were compared between the HG002-T2T chrY reference and each *de novo* Verkko chrY assembly. Existing testis Iso-seq data from seven individuals was also analyzed (SRX9033926 and SRX9033927).

o. Hi-C data analysis

We analyzed 40/43 samples for which Hi-C data was available (Hi-C data is missing for HG00358, HG01890 and NA19705). For each sample, GRCh38 reference genome was used to map the raw reads and Juicer software tools (version 1.6) [98] with the default aligner BWA mem (version: 0.7.17) [99] was utilized 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 applied to create a Hi-C contact map for each sample. To leverage the collected chrY Hi-C data from these 40 samples with various resolutions, we combined the chrY Hi-C contact maps of these 40 samples using the *mega.sh* script [98] given by Juicer to produce a "mega" map. Knight-Ruiz (KR) matrix balancing was applied to normalize Hi-C contact frequency matrices [100].

We then calculated Insulation Score (IS) [101], which was initially developed to find TAD boundaries on Hi-C data with a relatively low resolution, to call TAD boundaries at 10 kilobase resolution for the merged sample and each individual sample. For the merged sample, the FAN-C toolkit (version 0.9.23b4) [102] with default parameters was applied to calculate IS and boundary score (BS) based on the KR normalized "mega" map at 10 kb resolution and 100 kb window size (utilizing the same setting as in the 4DN domain calling protocol) [103]. For each individual sample, the KR normalized contact matrix of each sample served as the input to the same procedure as in analyzing the merged sample. The previous merged result was treated as a catalog of TAD boundaries in lymphoblastoid cell lines (LCLs) for chrY 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 as 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.

The average and variance (maximum difference between any of the two samples) insulation scores of our 40 chrY samples were calculated to show the differences among these samples and were plotted aligned with methylation analysis and chrY assembly together. Due to the limited Hi-C sequencing depth and resolution, some of the chrY 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) test (SciPy v1.7.3 scipy.stats.kruskal) was performed on the insulation scores (10 kb resolution) of each sample with the same 6 meta haplogroups classified in the methylation analysis to detect any associations between differentially insulated regions (DIR) and differentially methylated regions (DMR). 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.

p. Yq12 heterochromatin analyses

Yq12 partitioning

RepeatMasker (v4.1.0) was run using the default Dfam library to identify and classify repeat elements within the sequence of the Yq12 region[104]. 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. Therefore, five-hundred *DYZ2* repeats 2,413 bp in length were randomly selected from each assembly, and their sequences retrieved using Pysam (version 0.19.1) [105], (https://github.com/pysam-developers/pysam). Next, a multiple sequence alignment (MSA) of these five-hundred sequences was performed using Muscle (v5.1) [106]. Based on 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 HSAT I sequence of all *DYZ2* repeat units were retrieved, an MSA performed using Muscle (v5.1)[106], and a consensus sequence 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 re-screened using HMMER (v3.3.2) and the 2,420 bp *DYZ2* consensus sequence.

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 with 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 (v1.8.1) library was used to calculate the Bray-Curtis

Distance/Dissimilarity (as implemented in scipy.spatial.distance.braycurtis) between *DYZ2* array composition profiles [107]. 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. 4e, S49**).

Yq12 *DYZ1* array analysis

Initially, RepeatMasker (v4.1.0) 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 (Skaletsky et al 2003). In light of these findings, *DYZ1* array sequences were extracted with Pysam, and then each sequence underwent a virtual restriction digestion with HaeIII using the Sequence Manipulation Suite [108]. HaeIII, which has a 'ggcc' restriction cut site, was chosen based on previous research of the *DYZ1* repeat in monozygotic twins [109]. 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 this length were in the greatest abundance in 6 out of 7 analyzed genomes), performing a MSA using Muscle (v5.1), 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-kbp, 2.7-kbp, *DYZ1*) were computed. A k-mer of length 5 was chosen as *DYZ1* is likely ancestrally derived from a pentanucleotide [4,110]. Next, the Bray-Curtis dissimilarity between k-mer abundance profiles of each fragment and consensus sequence was computed, and fragments were classified based on their similarity to the consensus sequence k-mer profile (using a 75% similarity minimum) **(Fig. S38).** Afterwards, the sequence fragments with the same classification adjacent to one another were concatenated, and the fully assembled sequence was provided to HMMER (v3.3.2) to detect repeats and partition fragment sequences into individual repeat units [63]. The HMMER output was filtered by E-value (only E-value of zero was kept). Once individual repeat units (*DYZ18*, 3.1-kbp, 2.7-kbp, and *DYZ1*) were characterized (**Fig. S39**), 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 **(Fig. S40).** A two-sided Mann-Whitney U test (SciPy v1.7.3 scipy.stats.mannwhitneyu[107]) was utilized 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, pvalue=0.000786; HG02011:MWU=169.0, pvalue=0.000167; HG01106:MWU=617.0, pvalue=0.038162; HG01952:MWU=172.0, pvalue=0.042480; HG01890:MWU=51.0, pvalue=0.000867; HG02666:MWU=144.0, pvalue=0.007497; HG00358:MWU=497.0, pvalue=0.008068;) (**Fig. 4b**). A two-sided Spearman rank-order correlation coefficient (SciPy v1.7.3 scipy.stats.spearmanr[107]) was calculated using all samples with a completely assembled Yq12 (n=7) to measure the relationship between the total length of the analyzed Yq12 region and the total *DYZ1* plus *DYZ2* arrays within this region (correlation=0.90093, p-value=0.005620) (**Fig. S46**). All statistical tests performed were considered significant using an alpha=0.05.

Yq12 Mobile Element Insertion Analysis

RepeatMasker output was screened for the presence of additional transposable elements, in particular mobile element insertions (MEIs). Putative MEIs (i.e., elements with a divergence <4%) plus 100 nt of flanking were retrieved from the respective assemblies. Following an MSA using Muscle, the ancestral sequence of the MEI was determined and utilized for all downstream analyses, (This step was necessary as some of the MEI duplicated multiple times and harbored substitutions). The divergence, and subfamily affiliation, were determined based on 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 (TPRT) hallmarks (i.e., presence of an A-tail, target site duplications, and endonuclease cleavage site) [111].

6. 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 running text. Figures were generated using R and Python's Matplotlib (https://matplotlib.org), seaborn [68] and the "Turtle" graphics framework (https://docs.python.org/3/library/turtle.html).

7. Data Availability

All data generated are available via the HGSVC data portal at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/ and https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/. HPRC year 1 data files, PacBio HiFi, Oxford Nanopore (ONT) long-read sequencing and Bionano Genomics optical mapping and data files were downloaded from the following url: https://humanpangenome.org/year-1-sequencing-data-release/.

8. Code Availability

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.

Author contributions

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

Acknowledgements

We thank Dr. Yali Xue for discussions and advice throughout the project; Jonathan Wood and the Genome Reference Informatics Team at the Wellcome Sanger Institute for suggestions and feedback on assembly evaluation; the Human Pangenome Reference Consortium (HPRC, https://humanpangenome.org) for making their data publicly available; the Centre 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 expert assistance with the work described herein and Research IT for providing computational infrastructure and support and the Phillippy Lab (NIH/NHGRI) for their comprehensive Verkko support. We are grateful to the people who generously contributed samples as part of the 1000 Genomes Project.

Funding

Funding was provided by National Institutes of Health (NIH) grants U24HG007497 (to C.L., E.E.E., J.O.K., T.M.), U01HG010973 (to T.M., E.E.E., and J.O.K.), and R01HG002385 and R01HG010169 (to E.E.E.); 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 (J.O.K. and P.Hasenfeld); the EMBL International PhD Programme (W.H.); the Jackson Laboratory Postdoctoral Scholar Award (K.K.); NIH National Institute of General Medical Sciences (NIGMS R35GM133600 to C.R.B.; 1P20GM139769 to M.K.K. and M.L.) and National Cancer Institute (NCI) (P30CA034196 to C.R.B. and P.A.A.); U24HG007497 (P. H., F.Y., Q.Z., F.T., J.Y.K.); NIGMS K99GM147352 (G.A.L.) and Wellcome grant 098051 (to C.T.-S.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

Competing interests

E.E.E. is a scientific advisory board (SAB) member of Variant Bio. C.L. is an SAB 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.

References

1. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**, 1563–1572 (2000).
2. Vollger, M. R. *et al.* Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
3. 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).
4. 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).
5. 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 Research* vol. 14 7569–7580 Preprint at https://doi.org/10.1093/nar/14.19.7569 (1986).
6. Cooke, H. Repeated sequence specific to human males. *Nature* **262**, 182–186 (1976).
7. Skov, L., 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).
8. 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).
9. Sahakyan, H. *et al.* Origin and diffusion of human Y chromosome haplogroup J1-M267. *Sci. Rep.* **11**, 6659 (2021).
10. Rhie, A. *et al.* The complete sequence of a human Y chromosome. bioRxiv 2022.
11. Ebert, P. *et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, (2021).
12. Liao, W.-W. *et al.* A Draft Human Pangenome Reference. *bioRxiv* 2022.07.09.499321 (2022) doi:10.1101/2022.07.09.499321.
13. 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).
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. Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339–348 (2002).
17. 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).
18. Rautiainen, M. *et al.* Verkko: telomere-to-telomere assembly of diploid chromosomes. *bioRxiv*

2022.06.24.497523 (2022) doi:10.1101/2022.06.24.497523.

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. Yan, Y. *et al.* Copy number variation of functional RBMY1 is associated with sperm motility: an azoospermia factor-linked candidate for asthenozoospermia. *Hum. Reprod.* **32**, 1521–1531 (2017).
21. Gegenschatz-Schmid, K., Verkauskas, G., Stadler, M. B. & Hadziselimovic, F. Genes located in Y-chromosomal regions important for male fertility show altered transcript levels in cryptorchidism and respond to curative hormone treatment. *Basic Clin Androl* **29**, 8 (2019).
22. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
23. Hammer, M. F. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Mol. Biol. Evol.* **11**, 749–761 (1994).
24. 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).
25. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).
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. Lange, J. *et al.* Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**, 855–869 (2009).
38. Sturtevant, A. H. Genetic Factors Affecting the Strength of Linkage in Drosophila. *Proc. Natl. Acad. Sci. U. S. A.* **3**, 555–558 (1917).
39. Verma, R. S. *Heterochromatin: Molecular and Structural Aspects*. (Cambridge University 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).
42. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
43. Logsdon, G. HMW gDNA purification and ONT ultra-long-read data generation v3. (2022) doi:10.17504/protocols.io.b55tq86n.
44. 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.* (2019) doi:10.3791/58954.
45. 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).
46. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
47. Sanders, A. D. *et al.* Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).
48. Byrska-Bishop, M. *et al.* High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426–3440.e19 (2022).
49. Poznik, G. D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).
50. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (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. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
53. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
54. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
55. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
56. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
57. 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).
58. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
59. 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).
60. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
61. 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).
62. 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).
63. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
64. 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).
65. Altemose, N. A classical revival: Human satellite DNAs enter the genomics era. *Semin. Cell Dev. Biol.* **128**, 2–14 (2022).
66. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
69. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
70. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). doi:10.25080/majora-

92bf1922-011.

71. 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).

72. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).

73. Vollger, M. R., Kerpedjiev, P., Phillippy, A. M. & Eichler, E. E. StainedGlass: Interactive visualization of massive tandem repeat structures with identity heatmaps. *Bioinformatics* (2022) doi:10.1093/bioinformatics/btac018.

74. 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).

75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

76. 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).

77. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

78. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552 (2000).

79. Ren, J. & Chaisson, M. J. P. lra: A long read aligner for sequences and contigs. *PLoS Comput. Biol.* **17**, e1009078 (2021).

80. Heller, D. & Vingron, M. SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1034.

81. Smolka, M. *et al.* Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *bioRxiv* 2022.04.04.487055 (2022) doi:10.1101/2022.04.04.487055.

82. 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).

83. Zheng, Z. *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *bioRxiv* 2021.12.29.474431 (2021) doi:10.1101/2021.12.29.474431.

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

85. Edge, P. & Bansal, V. Longshot enables accurate variant calling in diploid genomes from single-molecule long read sequencing. *Nat. Commun.* **10**, 4660 (2019).

86. Audano, P. A. *et al.* Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663–675.e19 (2019).

87. 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).

88. 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).

89. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1016.
90. Snajder, R., Leger, A., Stegle, O. & Bonder, M. J. pycoMeth: A toolbox for differential methylation testing from Nanopore methylation calls. *bioRxiv* 2022.02.16.480699 (2022) doi:10.1101/2022.02.16.480699.
91. The R Project for Statistical Computing. https://www.R-project.org/.
92. Community Ecology Package [R package vegan version 2.6-4]. (2022).
93. Cuomo, A. S. E. *et al.* Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* **22**, 188 (2021).
94. 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).
95. Krueger. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for …. *URL http://www. bioinformatics. babraham. ac. uk*.
96. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
97. 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).
98. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
99. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
100. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2012).
101. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
102. 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).
103. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
104. 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).
105. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
106. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
107. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*

Methods **17**, 261–272 (2020).

108. Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**, 1102, 1104 (2000).

109. Yadav, S. K., Kumari, A., Javed, S. & Ali, S. DYZ1 arrays show sequence variation between the monozygotic males. *BMC Genet.* **15**, 19 (2014).

110. Prosser, J., Frommer, M., Paul, C. & Vincent, P. C. Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* **187**, 145–155 (1986).

111. Konkel, M. K., Walker, J. A. & Batzer, M. A. LINEs and SINEs of primate evolution. *Evol. Anthropol.* **19**, 236–249 (2010).