Supplementary Results and Figures

Table of contents

Supplementary Results

De novo assembly evaluation

We evaluated a series of contig alignments, in which we combined different samples and assemblies as target and query for the alignment process. We selected all contig alignments for the respective assemblies to the T2T Y reference as a baseline, and ran the following experiments (**Fig. S50**, from left to right): pairing the two closely related African samples NA19317 and NA19347 (with the TMRCA estimated to be only 200 ya [95% HPD interval: 0 - 500 ya] and therefore considered as quasi-replicates); considering the four pairs of high- and lower-coverage assemblies; aligning the Verkko hybrid assemblies to HiFi-only assemblies built with hifiasm v0.16.1-r375[1], and generating self-alignments. In all these scenarios, the collected statistics support the view that the Verkko assemblies have been robustly assembled and contain sample-specific sequences. For example, the fraction of the query sequences aligned with the maximal mapping quality (MAPQ) of 60 is highest for the self-alignments (**Fig. S50**, middle row), followed by the quasi-replicate African pair and the alignments to the HiFi-only hifiasm assemblies. The alignments of the high- and lower-coverage pairs (**Fig. S50**, middle row, blue boxes) show a drop relative to the aforementioned combinations, which is consistent with the higher error rate for the lower-coverage assemblies (**Fig. S4**). Next, we checked the overall (dis-)similarity of the Y assemblies with respect to their k-mer content (**Methods**). Relative to the GRCh38 Y assembly, all our assemblies plus the T2T Y show a coherent behavior, sharing a substantial fraction of their constituent k-mers (**Fig. S51**). Notably, the four pairs of high and lower coverage assemblies do not exhibit any inconsistencies, suggesting that, despite elevated error rate and increased fragmentation at lower coverages, the assemblies still represent a sample-specific Y chromosome.

We investigated the locations of assembly gaps via aligning all identified Y contigs to the GRCh38 and CHM13 plus T2T Y reference sequences and assessing the Y-chromosomal alignments via alignment coverage (**Figs. S52-56, Methods**). Contigs aligning well to the reference sequence were expected to show coverage of 1, while assembly gaps and poor alignments due to misalignment, misassembly or structural differences between the assembly and reference sequence should show no or >1 coverage. The results highlight the Yq12 heterochromatin as the most poorly aligning subregion (**Fig. S54**), consistent with the presence of the highest number of Y assembly breaks across samples (**Table S10**) and an overall high variation in the composition of this region across samples (**Fig. 4f**). In the (peri-)centromeric region majority of the poor alignments were localized to the highly repetitive centromeric *DYZ3* α-satellite array (**Fig. S56**). In euchromatic regions the PAR1 and ampliconic subregion 7 were the most challenging to contiguously assemble. Majority of the poorly aligning regions in ampliconic region 7 overlap with the P1 palindrome (**Figs. S53**) composed of ~1.45 Mbp inverted segmental duplications 99.97% sequence identity[2], while in the PAR1 the poorly aligning regions are broadly distributed (**Fig. S55**).

We also compared in more detail the assemblies of the closely related pair of African Y chromosomes from NA19317 and NA19347, assembled to a similar level of contiguity (NA19317 contiguously assembled from PAR1 to Yq12 in a single contig, while NA19347 has an additional break at the (peri-)centromeric region). In agreement with the TMRCA estimate, the Y assemblies show high similarity in structure and sequence (**Fig. S3, Table S6**). Across 23.96 Mbp (PAR1, (peri-)centromeric, Yq12 and PAR2 regions were excluded as either not contiguously assembled or recombine with the X chromosome), only 233 nucleotide substitutions and 583 indels summing to a total of 976 base pairs were identified, translating to a sequence identity of 99.9959% (**Table S6**). A total of 286/583 indels represent expansions and contractions at polynucleotide tracts and short tandem repeats (STRs).

In addition, we used the Bionano optical mapping data to evaluate the quality of the Verkko *de novo* Y-chromosomal assemblies. We identified a total of 94 structural variants (i.e. inconsistencies between the Verkko assembly and Bionano optical mapping data) from the local *de novo* assembly results from 25/43 samples using the optical mapping variant calling algorithm (Bionano Solve v3.5.1) (see Methods section 'Assembly evaluation using Bionano Genomics optical mapping data' for details) (**Table S35**). For the remaining 18/43 samples, none of the detected variants from optical mapping data passed the filtering thresholds and therefore contain no inconsistencies between the Verkko assemblies and Bionano optical maps. Detailed investigation of single DNA molecules from optical mapping data suggest that the majority of detected structural variants (77/94) are correctly resolved in Verkko assemblies. However, 31/77 variant sites overlap with the hybrid scaffolding conflict sites, indicating that these sites might need to be further investigated. No single DNA molecules span the remaining 17/94 sites (from 10 samples) and therefore the accuracy of Verkko assemblies can not be evaluated at these sites. However, 6/17 sites overlap with PAR1 subregion which remains challenging to assemble due to its sequence composition and sequencing biases[3,4]. Additionally, 3/17 sites overlap with the *DYZ19* heterochromatic repeat array, composed of highly similar and repetitive sequences, 7/17 sites overlap with unplaced contigs containing sequences from AMPL6 or AMPL7 subregions, and 1/17 sites overlap with the AMPL7 subregion suggesting that these sites also need further investigation.

An additional assembly evaluation step was performed by aligning single optical mapping molecules to the Verkko assemblies (**Methods**). This approach identified a total of 2,351 10-kbp windows in 43 samples which were not covered by optical mapping molecules (**Table S36**). Majority of these windows (1,798/2,351; in 43/43 samples) overlapped with PAR1, PAR2, (peri-)centromeric and Yq12 heterochromatic subregions. Additionally, 300/2,351 windows (in 26/43 samples) overlapped with ampliconic regions (more specifically AMPL1, AMPL2, AMPL5, AMPL6 and AMPL7), all of which are challenging to contiguously assemble due to their sequence comparison. Only a small proportion of windows (253/2,351 windows) from 3/43 samples overlapped with X-transposed and Xdegenerate regions (XTR1, XDR1 and XDR3; **Table S36**), highlighting regions that would require further investigation. It is important to keep in mind that while optical mapping data offers an independent orthogonal validation for the generated assemblies, it loses resolution at heterochromatic region (due to the lack of restriction enzyme cutting sites) and can struggle to correctly characterize highly repetitive and complex genomic regions.

Effect of input read characteristics on assembly contiguity

We explored the potential effect of the varying input read set characteristics, such as genomic coverage or read length N50, on the outcome of the hybrid assembly process. First, we randomly selected four of the high-coverage samples (HG02666, HG01457, NA19384, NA18989; HiFi coverage at least 50×, in the following denoted with the prefix "HC" for high coverage where needed to disambiguate) and re-assembled those with about half of the available HiFi reads, i.e., using around 30× coverage, which is comparable to most of the HiFi datasets used in this study (**Tables S1-S2**). The lower-coverage assemblies show higher fragmentation as indicated by a considerably larger number of assembled contigs and a smaller contig NG50 statistic (**Tables S4-S5**). This observation is compatible with the assumption that higher input read coverage has a positive effect on assembly quality in terms of contiguity. However, given that not all Y chromosomes of the high-coverage samples could be assembled contiguously from telomere to telomere (**Tables S5, S7**), it is evident that this factor alone is not sufficient as an explanatory variable. Moreover, for all four lower-coverage assemblies, the total assembled length of the Y sequence is increased by two to six megabases compared to their high-coverage counterparts, which suggests that the total assembly length may be of limited value when comparing Y assemblies created with substantially different HiFi input coverage (**Table S5**). We deepened our analysis by training multivariate regression models (**Methods**) to investigate the relationship between the input read set and quality-related assembly statistics of interest such as the contig NG50. For this analysis, we augmented our dataset with the four lower-coverage assemblies described above. The results of the regression analysis confirmed that HiFi input coverage and mean ONT-UL read length are relevant factors to achieve higher contig NG50 values (**Tables S37-S38**), yet cannot be sufficient as explained above. Given the small size of our dataset from a statistical point of view, e.g., including only two samples from haplogroup A (HG02666, HG01890), and these two Y chromosomes could be assembled in a single contig from telomere to telomere, it is challenging to derive a robust statement about the factors governing overall assembly quality.

Orthogonal support to Y-chromosomal SVs

We evaluated assembly-derived structural variants called with PAV (using the GRCh38 Y references sequence, **Table S15-S17, Methods**) by using optical mapping data as an orthogonal support (**Methods**). The 31 evaluated variants included all 10 inversions, and 9 insertions and 12 deletions >=5 kbp in size called using PAV. Overall, 20/31 structural variant genotypes (7 deletions, 7 insertions and 6 inversions) were supported by optical mapping data across the majority of the samples where the variant had been called (**Table S39**). Out of the 12 remaining variants, 5 were located in the (peri-)centromeric region where optical mapping does not have sufficient resolution. For 6/12 structural variants (1 deletion, 1 insertion and 4 inversions) the called genotypes were not supported by optical mapping data indicating that inversions remain the most challenging variant type to call accurately.

Gene annotation

To annotate genomes of *de novo* assemblies of 43 male samples, we used liftoff and GENCODEv41 GRCh38 annotations, and T2T-CHM13v2.0 chrY annotations (**Table S25**). Annotation of 43 Y chromosomes presented a number of genes ranging from 580 (HG00358) to 758 (HG02666). The number of protein-coding genes ranged from 82 (HG00358 and HG03732) to 114 (HG02666), and the number of pseudogenes from 339 (HG00358) to 457 (NA18989) (**Table S25**). Majority of differences between GRCh38 Y and T2T Y annotations were due to previously unassembled regions, gaps, and ampliconic gene copy numbers (**Tables S22-S23, S26**). The single-copy protein-coding genes were present in all samples, 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**). In addition, there are 8 multi-copy protein-coding gene families located in the ampliconic regions, 5 of which showed variation in copy number across the analyzed samples. 3/8 protein-coding gene families (*VCY*, *PRY* and *HSFY*) showed a constant copy number (2 copies) across all samples. Two of the assembled samples (HG00358, haplogroup N1c-Z1940 and NA18989, haplogroup C1a-CTS6678) carry known rearrangements in the *AZFc*/ampliconic 7 subregion - ~1.8 Mbp *b2/b3* deletion and a likely *gr/rg* duplication, respectively[5–7] and show variation in copy number in genes (*DAZ*, *BPY2* and *CDY1*) affected by these rearrangements. The *BPY2* gene copy number ranges from 1 to 5 (41/44 samples show a constant copy number of 3), *CDY* from 3 to 5 copies (39/44 samples show a constant copy number of 4) and *DAZ* from 2 to 6 copies (42/44 samples show a constant copy number of 4). Note that the accuracy of gene annotation and copy number determination might be impacted by fragmented assembly in case of a few samples (**Table S22**). The highest variation in copy number across the 43 *de novo* samples was observed for *RBMY* gene (from 5 to 11 copies, 27 copies in T2T Y) and *TSPY* (from 24 to 40, 47 copies in T2T Y; **Tables S22-S23**). The *RBMY* copy number estimates for 14 samples overlapping with[8] (that used read depth information from low-coverage Illumina data) are highly concordant with 13/14 samples showing either exactly the same or plus 1 total *RBMY* copy number estimates, offering independent support to the quality of our Y assemblies.

Y-chromosomal inversions

Inversions have remained one of the most challenging structural variation types to reliably genotype, especially when flanked by large highly similar segmental duplications. The Y-chromosomal *de novo* assemblies resolved to basepair level enabled us to confidently identify a total of 16 inversions (14 in the euchromatic regions and 2 in the Yq12 heterochromatic region) from the 44 individuals (43 assembled here and the T2T Y) analyzed here, to narrow down the breakpoint locations for 10/16 inversions and improve the inversion rate estimates due to higher phylogenetic resolution compared to previous reports[9,10].

The 14 euchromatic inversions were identified from the *de novo* Verkko assemblies and independently called using Strand-seq data mapped to both GRCh38 Y and the T2T Y reference sequences, available for 31/44 samples (see Methods section **'Inversion analyses'**) (**Tables S27, S40**). In addition, 7/14 of the euchromatic inversions overlapped with inversions called using PAV (**Tables S15, S17**; see Methods section **'Variant calling using *de novo* assemblies'**). All 14 inversions are flanked by inverted repeats, showing up to 99.99% sequence similarity between the repeats and up to 1.45 Mbp in size (the P1 palindrome)[2]. The sizes of inversions range from approximately 30 kbp (the P7 palindrome) to 5.94 Mbp (the IR5/IR5 inversion in HG02666, see more details below) (**Table S27, S29**). Combining the maximum sizes of euchromatic regions affected by inversions sums to a total of approximately 12.18 Mbp or 54.6% of GRCh38 MSY euchromatic composition. 12/14 euchromatic inversions are recurrent, toggling in the Y phylogeny from two (the *blue2/blue3* or *b2/b3* inversion, **Fig. S22**) to 13 times (P3 palindrome composed of 283 kbp inverted repeats which are separated by a 170 kbp spacer region[2]). The two inversions identified in single individuals (the *blue1/blue4* or *b1/b4* in HG01890 and *IR5/IR5* in HG02666, see more details below) are the largest, approximately 4.2 Mbp and 5.94 Mbp in size. Overall, across all 44 samples included in the current study, only the most closely related pair of African Y chromosomes (carried by NA19317 and NA19347) show identical composition in terms of inversions (**Fig. 3a; Table S27**), highlighting the high structural variability of the human Y chromosome.

Taking advantage of the sequence resolution offered by the Verkko assemblies, we succeeded in determining the likely breakpoint ranges for 8 euchromatic inversions down to 500-bp region (**Fig. 3b; Figs. S26-S28; Table S29;** Methods section **'Determination of inversion breakpoint ranges'**), allowing us to determine the inversion sizes more accurately. According to the GRCh38 coordinates, the average sizes of breakpoint ranges for palindromes P3-P8 are 33,381 bp (ranging from 1,115 bp in P3 to 181,342 bp in P4) and 33,203 bp (ranging from 1,117 bp in P3 to 181,342 bp in P4) for proximal and distal copies of inverted repeats or palindrome arms. The location of breakpoint ranges tend to be located closer to the spacer region, suggesting that the distance between the breakpoints in the proximal and distal arms impacts the triggering of an inversion event (**Fig. S27**). The inversion sizes (for palindromes P3-P8 and IR3) range from 29,426 bp (palindrome P7) to 3,679,407 bp (IR3) with an average of 714,036 bp, when estimating it based on the start coordinate on the proximal repeat/arm and the end coordinate on distal repeat/arm of the breakpoint ranges. The inversion size, as well as the size of the breakpoint range, are positively correlated with the size of palindrome, except for the IR3 repeats where the unique spacer region (~3.5Mb) is substantially larger than that of any Y palindrome (Spearman's correlation coefficient between breakpoint range and proximal/distal arm size: 0.8857 (p-value 0.0333), and between inversion size and proximal/distal arm size: 1.00 (p-value 0.0028) based on GRCh38 coordinates).

Large inversions are mostly responsible for the fact that all three of our contiguously assembled Y chromosomes are structurally distinct from each other across multi-Mbp euchromatic regions (**Fig. 2b,c; Figs. S6-S8, S16, S25**), and from both GRCh38 and the T2T Y sequences, which also differ from each other due to a known >1.9 Mbp polymorphic *gr/rg* inversion carried by the T2T Y[4,5]. The structural composition of the *AZFc* region in the deepest-rooting Y chromosome (HG01890 A0b-L1038) can be explained by two inversions (between the *blue 1* and *blue 4* amplicons, and another between the *blue 2* and *blue 3* amplicons, **Fig. S22**), up to 4.1 and 1.2 Mbp in size (considering the start and end coordinates of the respective blue amplicons in the GRCh38 Y), respectively, or three inversions (additionally requires the *gr/rg* inversion) when compared to the T2T Y (**Figs. 2b-c; Figs. S6, S16 and S25a**).

The second deepest-rooting Y chromosome from HG02666 (A1a-M31) carries a P5/P1 inversion and additionally a smaller inversion between *blue 2* and *blue 3* amplicons (**Figs. 2b-c; Figs. S7, S16**). We were able to pinpoint the inversion breakpoints of the P5/P1 inversion into 504-bp intervals (**Fig. S28**) within ERV1 repeat elements in the IR5 repeats located in inverted orientations in the distal arm of P5 palindrome and in the proximal arm of the P1 palindrome. The resulting inversion is 5.941 Mbp in size relative to the Verkko assembly for HG02666, or 6.001 Mbp relative to GRCh38 Y and likely caused by non-allelic homologous recombination (NAHR). Recombination between palindromes P5 and P1 (both P5/proximal-P1 and P5/distal-P1 deletions, known as *AZFb* deletions) are known to cause massive deletions and spermatogenic failure, with most breakpoints identified within a hotspot region within 30 kbp from the center of the P5 palindrome [11]. Interestingly, the inversion breakpoints identified here do not overlap with the deletion hotspots as they are located ~81.7 kb from the center of the P5 palindrome. Closer inspection of the sequences of the *blue 2* and *blue 3* repeats from HG01890 and HG02666 indicates that these are independent inversions and were therefore counted as independent events in inversion rate calculations.

The Y assembly for HG00358 (N1c-Z1940) contains a known ~1.8 Mbp *b2/b3* deletion fixed in haplogroup N samples (**Figs. 2a-b; Figs. S8, S16 and S25a**)[5].

We detected the *gr/rg* inversion, one of the major structural differences between the GRCh38 Y and the T2T Y sequences, in seven samples (**Fig. 3a; Table S27**), including the two other haplogroup J samples (HG02492 J2a-M47 and HG01259 J1-M267) which are most closely related to the T2T Y. The presence of *gr/rg* inversions is also supported by Bionano optical mapping data. Our results on *gr/rg* phylogenetic distribution fit well with previous reports both in terms of the presence of this inversion in haplogroups B2b-M112, E1b1b1b1a-M81, and its absence in other Y lineages overlapping between the two studies, although matching the results exactly is not possible due to lower resolution of typed phylogenetically informative markers by Repping and colleagues [5]. This most likely also explains the absence of the *gr/rg* inversion in their haplogroup J samples, while indicating that the inversion is not shared by haplogroup J samples as our phylogeny might suggest, but instead occurred independently in J1-M267 and its sublineages (carried by HG01259 and the T2T Y) and J2a-M47 (carried by HG02492). However, since we were not able to determine the inversion breakpoints for the *gr/rg* inversion, we took the conservative approach and counted a total of 5 independent inversions in the phylogeny (instead of 6 in case the inversions in J1 and J2a were independent). Overall, the concordance with previous studies supports structurally correct assembly of this complex region in our dataset.

The largest recurrent inversion among our samples is found on the p-arm, mediated by the inverted IR3 repeats, each approximately 290-300 kbp in size. The IR3 inversions are known to be polymorphic and reported to be approximately 3.3-3.8 Mbp in size[9,10]. Interestingly, we discovered that most (33/44) Y chromosomes, including the T2T Y, show a distinct composition of IR3 repeats compared to the GRCh38 Y sequence (**Fig. S57**). In GRCh38 Y, the distal IR3 repeat contains a single copy of the ~20.3 kbp *TSPY* repeat (see Method section **'TSPY repeat copy number analysis'**) in direct orientation, while in the majority of samples the single TSPY repeat is located in the proximal *IR3* repeat in inverted orientation (**Fig. 3b; Fig. S57**). Analysis of the *IR3* repeat sequences revealed that the phylogenetically closely related Y haplogroup QR samples (including GRCh38 Y, mostly haplogroup R1b) have likely undergone two inversions - a ~3.67 - 3.68 Mbp (relative to GRCh38 Y sequence) inversion changing the location and orientation of the single TSPY repeat from the distal to proximal repeat, while another, ~3.24 - 3.28 Mbp inversion reverted the region located between the *IR3* repeats (**Fig. S57; Table S29**). In addition to these two events shared by all QR lineage Y chromosomes, the *IR3/IR3* inversion was identified in four samples which now carry the genomic region in between the *IR3* repeats in inverted orientation compared to other samples **(Fig. 3a**), totalling to six inversion events across all analyzed samples. The inversion breakpoint ranges were narrowed down to regions of 6.7 to 40.1 kbp in size (**Fig. 3b; Table S29**). In two samples (NA19239 and HG03492) the inversion breakpoints were located closer to the unique spacer region, leading to inversions of ~3.2 Mbp in size. Interestingly, the inversion breakpoint region in HG03492 overlaps with the second inversion region shared by all QR samples. In HG03732 and NA19331 the inversions were larger, ~3.4 Mbp in size, and inversion breakpoints were located closer to the center of IR3 repeats.

Additionally, we highlight an inverted duplication which affects roughly two thirds of the 161 kbp unique sequence in the P3 palindrome, spawns a second copy of the *TTTY5* gene and effectively elongates the segmental duplications in this region (**Fig. S25b**). A detailed sequence view reveals a high sequence similarity between the duplication and its template, and its placement in Y phylogeny supports emergence of this variant in the common ancestor of haplogroup E1a2 carried by NA19239, HG03248 and HG02572 (**Fig. 1a; Figs. S1, S25b**).

In addition to the inversions in the euchromatic regions of the Y chromosome, we also identified inversions at the proximal and distal ends of the Yq12 heterochromatic region, one at each end (**Fig. 4c**). The inversion breakpoint analyses at the nucleotide level revealed distinct breakpoints, further supporting the presence of these two inversion events (**Fig. S29; Table S28**). Alternatively, a complex rearrangement with multiple breakpoints, resulting in orientation changes of the *DYZ1* and *DYZ2* repeat units within the distal and proximal ends of the Yq12 region, could have occurred.

As some variation was noticed within the proximal inversion region across the 11 analyzed samples, breakpoint analysis was performed for each assembly separately. For 9/11 examined assemblies, the 5' breakpoint of the proximal inversion was identified within a *DYZ2* repeat unit at the 3' end of the *Alu* sequence immediately upstream of a second 'orphaned' *Alu* A-tail (Adenosine-rich sequence) segment (**Fig. S29a**). The 3' breakpoint of the proximal inversion resides within the intersection of an AT-rich simple repeat region of a *DYZ2* subunit and a *DYZ1* subunit (**Fig. S29a**). For all of the assemblies analyzed, the 5' breakpoint of the distal inversion is situated at the boundary of an AT-rich simple repeat and the 5' end of an *Alu* sequence ('head' of the *Alu*) within a *DYZ2* repeat unit (**Fig. S29b**). Finally, the 3' breakpoint of the distal inversion lies between an AT-rich simple repeat and the remaining portion of the *Alu* sequence head right before the HSATI satellite (**Fig. S29b**).

Across the eleven analyzed samples, three distinct patterns within the proximal inversion region were observed. While the majority of assemblies shared the breakpoints described above, two assemblies – HG01106, and HG01890 – showed a deviating pattern. In HG01106 the entire proximal inversion region seems deleted and additional studies are required to determine if this is shared by other closely related Y chromosomes, or is sample-specific (rearrangements having occurred in the lymphoblastoid cell line can not be excluded). To determine the ancestral state of the inversion region, the HG01890 Y assembly was further investigated. This was deemed particularly important, as HG01890 represents the deepest rooting Y chromosome lineage in the current dataset. Comparison of HG01890 with the other Y assemblies revealed the likely presence of deletions encompassing both the 5' and 3' breakpoints of the proximal inversion.

Yq12 heterochromatic subregion

A Yq12 overview

Our comparison of the Yq12 subregion of T2T Y and GRCh38 Y revealed that the distal section, situated closest to the PAR2 subregion, is structurally distinct from the rest of Yq12 and fully assembled in GRCh38 Y reference sequence. As no evidence of structural variation was found within this region, we focused on the previously incompletely assembled proximal sections of this region, including the *DYZ18* repeat array (**Fig. 1a**; **Tables S9, S11**) in our subsequent analyses of the seven samples (HG01890, HG02666, HG00358, HG01106, HG01952, HG02011 plus the T2T Y) with contiguously assembled Yq12 heterochromatic regions.

First, we assessed the previously mostly unassembled Yq12 region for its repetitive sequence composition. Within each of the analyzed genomes, we observed an alternating pattern of two distinct segments (**Methods**). One segment consists mainly of a tandemly repeated AT-rich simple repeat fused to a 5' truncated *Alu* element, followed by an HSATI satellite. Comparison with the Yq12 literature revealed that this arrangement represents a previously described ~2.4 kbp tripartite repeat element, *DYZ2*[12,13]. The subunit composition in the second segment was less well defined. We noticed that these sequences mainly contain simple repeats and pentameric satellite sequences, with over 95% (33,677 of 35,370) of all satellites identified as HSATII. Further analyses revealed an association of this sequence with a ~3.5 kbp repeat called *DYZ1*[2,14–17]. Consequently, our analyses support that the alternating repeat segments be identified as *DYZ1* and *DYZ2* arrays. Interestingly, the total number of arrays within assemblies is positively correlated to the length of the analyzed Yq12 region (two-sided Spearman: 0.90; p-value=0.0056, **Fig. S46, Methods**).

Next, we extended the *DYZ1* and *DYZ2* array analyses to the two assemblies (HG01928 and NA19705) with a single gap within the Yq12. Additionally, we included the assemblies of the two most closely related individuals (NA19317 and NA19347) with an estimated divergence time of ~200 years despite the presence of multiple contigs to gain a better understanding of the evolution of this region. For the two assemblies with multiple contigs, we focused our analyses on the arrays that are continuously assembled and reside at the proximal and distal ends of the Yq12 region. As expected, we identified copy number variation both with regard to the number of *DYZ1* and *DYZ2* arrays and *DYZ1* and *DYZ2* repeat units within the arrays in all four assemblies **(Fig. S47b)**. However, the number of *DYZ1* and *DYZ2* repeat arrays within the assembled regions was identical within the two most closely related genomes **(Figs. S47b, S58)**. Furthermore, the *DYZ2* repeat unit copy numbers within 14/20 *DYZ2* arrays between NA19317 and NA19347 were identical **(Fig. S58)**. Comparison of these 14 *DYZ2* arrays with identical repeat unit copy number (encompassing a total of 2,231,881 nucleotides) revealed only five single nucleotide variants (SNVs) – none of which represented CpG mutations – and one indel within a homopolymeric adenosine tract. Of the remaining six *DYZ2* arrays, four were located in the proximal or distal ends of the Yq12 region and showed only minor variation in the *DYZ2* repeat unit copy number (+/- 1 *DYZ2* repeat units). The last two arrays were not included in the analyses because of their immediate adjacency to an incomplete assembly region.

Yq12 *DYZ1* and *DYZ2* repeat analyses

We examined inter-individual variation with regard to subunit composition of Yq12 *DYZ2* arrays in greater detail. Across the seven assemblies with fully assembled Yq12 region, the total *DYZ2* repeat units within the Yq12 region ranged from a minimum of 2,661 *DYZ2* subunits (HG01890) to a maximum of 6,681 *DYZ2* subunits (HG01106), with a mean of 4,380 units. *DYZ2* repeat units ranged in size from a minimum of 1,275 bp to a maximum of 3,719 bp, though 98.6% (30,242 out of 30,656 of all *DYZ2* repeat units across complete assemblies) were between 2,000-2,999 bp in length, with a median length of 2,420 bp (93.7% of all *DYZ2* repeats were 2,420 bp). Sequence composition analysis suggests that this variation in sequence length is primarily caused either by expansion or contraction within the AT-rich simple repeat segment of these elements (sample collective mean: 1,415 bp, standard deviation (SD): 383 bp). The single origin *DYZ2 Alu* sequence had a consistent length (sample collective mean: 290 bp, SD: 2 bp) and was primarily identified as *Alu*Y, though at roughly 20% divergence, the sequence is too diverged to confidently exclude *Alu*S origin. The HSATI satellite portion of the *DYZ2* subunit varied somewhat in size (sample collective mean: 566 bp, SD: 16 bp).

Our comparison to the *DYZ2* consensus sequence revealed that *DYZ2* repeat units located within arrays and positioned closer to the center of the Yq12 region were, on average, less diverged (i.e., potentially younger) **(Fig. 4d; Fig. S48)**. In contrast, more divergent *DYZ2* repeats were enriched toward the proximal and distal boundaries of the Yq12 region, with the putative oldest elements detected within the arrays situated between the distal inversion and the 3' end of the *DYZ* repeat arrays. Interestingly, this divergence pattern also seemed to be partially reflected within the individual *DYZ2* arrays where the divergence of *DYZ2* repeats situated closer to the center was generally lower compared to those near the ends. To investigate ongoing mutation dynamics, we also performed the *DYZ2* divergence analysis for the two most closely related genomes (NA19317 and NA19347). As expected, based on the previous *DYZ2* array comparisons, high similarity was uncovered between both genomes, and a similar divergence pattern as observed within the other genomes **(Fig. S48b)**.

Next, we constructed a *DYZ2* repeat composition profile for each *DYZ2* array within a genome. Our inter-*DYZ2* array profile comparison (see **Methods**), performed for each genome separately, revealed a trend towards *DYZ2* arrays closely situated to one another having higher repeat composition similarity **(Fig. 4e; Fig. S49)**. Curiously, these *DYZ2* array composition similarity heatmaps (**Fig. S49**) also exhibit what appear to be signals of past waves of amplifications/duplications of *DYZ2* arrays located between the peripheral Yq12 inversions.

Next, we investigated the Yq12 *DYZ1* repeat units in greater detail. Due to the low sequence complexity of the pentameric HSATII satellite and the simple repeat, we were unable to utilize the same approaches as those performed for the *DYZ2* arrays. Furthermore, an analysis using the previously published *DYZ1* consensus sequence[2] as a query sequence revealed an overall high divergence (~25%), further confounding downstream analyses. Based on these findings, two different approaches were pursued: (1) a virtual restriction digestion of the *DYZ1* array sequences with HaeIII that cuts DNA at ggcc sites[18], and (2) a targeted HMMER analysis[19]. The HaeIII restriction enzyme was selected based on previous molecular biology experiments of the *DYZ1* repeats in the Yq12 subregion, where the enzyme was shown to cut the repeat unit once, primarily resulting in fragments with 3,564 bp in length[18]. While our virtual digestion of the putative Yq12 *DYZ1* array regions of all complete assemblies showed a similar enrichment for 3,564 bp size fragments, we also observed considerable sequence length variation (Min: <25 bp, Max: >200 kbp) **(Fig. S59)**. Visualization of the distribution of fragment lengths within *DYZ1* arrays revealed a highly similar pattern across the seven complete Yq12 assemblies **(Fig. S59)**.

To explore the repeat composition of restriction fragments, we performed a k-mer profile similarity analysis. Considering that the first *DYZ1* array is adjoining the Yq11 *DYZ18*, 3.1-kbp, and 2.7-kbp repeat transition region, each digestion fragment was classified as being a unit, or a composition, of either *DYZ18*, 3.1-kbp repeat, 2.7-kbp repeat, or *DYZ1*. Compellingly, the findings of the *DYZ18* and transition region analysis within the Yq11 were supported and reiterated by this analysis **(Figs. S38, S41)**. The k-mer profile dissimilarity analysis indicated that the 3.1-kbp repeat showed higher similarity to the *DYZ18* repeat (91%), and the 2.7-kbp repeat to *DYZ1* (85%), suggesting that the Yq11/Yq12 transition zone repeats (3.1-kbp and 2.7-kbp) are possibly derived from *DYZ18* and *DYZ1* **(Fig. S41)**. Lastly, the virtual digest and HMMER analyses were combined where after digestion fragment classification, a targeted HMMER analysis was performed to partition restriction fragments into their individual repeat subunits **(Fig. S38)**.

While previous studies reported a ratio of *DYZ1* to *DYZ2* repeat units as 2 to 1[13,20,21], we observed a nearly equal repeat unit ratio (collective sample mean *DYZ1*:*DYZ2* ratio: 1.09) within the Yq12 **(Fig. 4b; Table S34)**. These findings align with our observation of a nearly 60:40 ratio of total nucleotides accounted for by *DYZ1* and *DYZ2* across all analyzed assemblies. Finally, the dissimilarity of *DYZ1* repeats versus the constructed *DYZ1* consensus sequence was computed and visualized (see **Methods**). This analysis mirrored findings of the *DYZ2* repeat divergence analysis, with *DYZ1* subunits located near the center of *DYZ1* arrays tending to be less dissimilar (i.e., less diverged) than those found near the boundaries of arrays **(Fig. S40)**.

Yq12 mobile element insertions (MEIs)

The Yq12 region was screened for the presence of mobile element insertions (MEIs) generated by the target-primed reverse transcription mechanism in both the *DYZ1* and *DYZ2* arrays. Four putative *Alu* insertions were identified across the seven samples with full Yq12 assemblies (**Fig. 4f**). While three of the insertions resided within the *DYZ2* repeat unit, the fourth insertion was located within the *DYZ1* repeat unit. Based on the divergence (3% or less), all four putative insertions appeared considerably younger than the *Alu* sequence of the composite *DYZ2* repeat unit. Furthermore, all *Alu* elements harbored hallmarks of classical MEIs such as target site duplications, termination in an adenosine-rich tail, and endonuclease cleavage site **(Table S31)**. Two of the insertions were identified as *Alu*Y and one each as *Alu*Ye5, and *Alu*Yb8. Both *Alu*Y insertions occurred within the AT-rich simple repeat region of the *DYZ2* repeat; though at different locations and not within the same repeat unit. The *Alu*Yb8 element inserted into a *DYZ1* repeat; while the *Alu*Ye5 element inserted immediately upstream of the 5' *Alu* sequence of one *DYZ2* repeat and in 'sense orientation' relative to *DYZ2*.

Alu elements are unique in that the ancestral state (i.e., absence of the MEI) is known and the precise removal of a MEI is exceedingly rare[22]. Based on this, the approximate age of the insertions, and presence in all Y chromosome lineages, it can be inferred that the two *Alu*Y insertions have occurred early in human Y chromosome evolution prior to the rise of the now known Y chromosome lineages. Only the T2T Y assembly lacked evidence for one of the two *Alu*Y insertions. Based on its phylogenetic placement, this likely results from a deletion or gene conversion of repeat units harboring the insertion **(Fig. 4f)**. The *Alu*Ye5 insertion is unique to HG01890, and the *Alu*Yb8 element to HG01952. Further analysis revealed that the *Alu*Yb8 element is shared with HG01928 (assembly of the Yq12 subregion is not contiguous), supporting insertion in a common ancestor of HG01952 and HG01928 **(Fig. 4f; Table S31)**.

While there is little evidence for post-insertion expansion of the *Alu*Yb8 element in the *DYZ1* repeat, the MEIs within a *DYZ2* repeat show varying degree of expansion with considerable inter-individual variation (**Fig. 4f**). For example, one *Alu*Y insertion was identified in six out of seven assemblies with a copy number range from one (in HG01106) to seventeen (in HG02666). This further highlights the enormous inter-individual variation of the human Yq12 region. Furthermore, from the MEI patterns it can be inferred that the insertions occurred into different repeat arrays and that the expansion/duplication occurred independently for each MEI. Interestingly, each MEI insertion and their extensions occupy distinct areas within the Yq12 region with no overlap between the different MEIs **(Fig. 4f)**.

These findings, in conjunction with the overall *DYZ1* and *DYZ2* array expansion/contraction dynamics, point toward random unequal crossing over between sister chromatids for the subsequent expansions of the *Alu* elements as well as the duplication or deletion of *DYZ1* and *DYZ2* arrays[23]. Unequal crossing over would also explain the expansion and contraction of repeats within these arrays without changing the repeat pattern[23], though gene conversion and replication slippage as contributing factors cannot be ruled out. The lower interindividual variation with regard to array number, array size, and *DYZ1/2* repeat units of the inversion regions and arrays distal to the inversions at the proximal and distal ends of the characterized repeat region is in agreement with the known recombination and crossing-over suppression of inversions[24]. Furthermore, a reduction in unequal crossing over near/within the Yq12 inversions could protect against deleterious effects outside the heterochromatin region such as gene-containing regions of the Y chromosome.

Functional analysis

DNA methylation calls on the ONT reads were derived from Nanopolish [25], after methylation calling and QC (**Methods**) we used pycoMeth[26] to *de novo* segment the methylation profiles of the 41 QCed samples (**Fig. S31**). This resulted in the identification of 2,861 independent segments (**Table S32**). To identify the global impact of the different haplogroups on the segmentation we used a permanova test. Specifically we grouped haplogroups into 6 meta groups based on sample size and genetic distance, haplogroup A, B and C ("ABC" 4 samples), G and H ("GH" 2 samples), N and O ("NO" 6 samples), and Q and R ("QR" 11 samples), E (19 samples), J (4 samples - including NA24149, the father of HG002/NA24385), Methods). These grouped haplogroups explain 21% of the global variation in DNAme levels profiles (Permanova, *P* 0.0029). On a segment level we found that 340 segments are differentially methylated (DM) (FDR 20%, **Table S32**, (**Methods**)). Interestingly 218 (64%) of the segments have decreased DNAme levels in the QR haplogroups. The 340 DM segments are enriched to overlap regulatory information (Fisher exact *P < 2.2e-16,* odds ratio: 6.72*),* but depleted in overlap to genes (Fisher exact *P 2.088e-05,* odds ratio: 0.52*,* methods, **Table S32**).

Next to the effects of haplogroups on DNAme we tested for local DNA methylation quantitative trait loci (meQTLs). We leveraged the limixQTL pipeline to test for effects of genetic variation with 100,000 bases around the DNAme segment as identified using pycoMeth. We controlled for population structure by controlling for population as a random effect, and leveraged permutations to determine significance of effects (supplementary methods). We identified 10 segments with significant meQTLs (FDR 20%) and found a total of 194 meQTL effects. The majority of the effects are linked to SNVs (109), with 1 variant being an INV, and 1 effect being from a 171 base-pair insertion (**Table S33**).

Given that expression data is available only on a subset of the HGSVC and HPRC samples (21/44) we focussed on the 210 males from the Geuvadis project[27] to assess the effects of haplogroups on gene expression level. We find 64 of the 205 genes on chromosome Y expressed in the Geuvadis LCL gene expression data (**Table S41**). As with DNAme we first tested for global expression variation, here we leveraged the first character of the haplogroup as grouping ("E":44, "G":4, "I":23, "J":18,"N":22,"R":96, "T":3 (group:nSamples)), and find that Y haplogroup explains 4.8% of the variation in gene expression (Permanova, *P 0.005*), and in total 22 genes are significantly differentially expressed (FDR 10%). Even though the samples and Y haplotype distribution is different between the DNAme samples and the Geuvadis data we find 5 genes (*BCFORP1*, *LINC00280*, *LOC100996911*, *PRKY*, *UTY*) that have both DNAme effects as well as gene expression effects. Specifically *BCORP1* is interesting as the effect directions on average match between the Geuvadis and HGSVC expression datasets and the expression effect is negatively correlated (r -0.3; p:0.1) between the overlapping HGSVC samples (**Fig. S32**).

To demonstrate the utility of these highly contiguous Y assemblies in representing the genic diversity of other individuals, we analyzed full-length cDNA sequences (PacBio Iso-Seq) of testis samples from seven anonymous donors (**Methods**). Of 30 Y-chromosomal genes expressed with at least five cDNA reads, 23 had improved transcript alignments compared to the T2T Y reference sequence, which provided only equal or inferior alignments (**Fig. S60; Table S42**). Most notably, *DAZ2* transcripts had alignments improved by 15.5% on average, due to the variable internal repeat structure. Across all genes, a full 19% of the improved alignments came from the Y assembly of a single sample, HG01596. We also generated Iso-seq data on eight matched samples corresponding to *de novo* Y assemblies (**Fig. S61; Table S43**). Aligning to a matched *de novo* Y assembly instead of the T2T Y reference improved between 14-51% of cDNA alignments.

Hi-C data has been widely utilized to characterize the 3D structure of the genome and identify chromatin structures, such as topologically associated domains (TADs) that play central roles in gene regulation. Previous research has primarily focused on Hi-C data analysis in autosomes, while here we investigate the variation of chromatin structures in diverse Y chromosomes. Using Hi-C data available from 40 samples, we identified TADs and TAD boundaries for Y chromosomes of these individuals by evaluating their insulation scores, which indicate the variations of the contact density of every Hi-C bin compared to adjacent bins (**Fig. S62-S63**; **Methods**)[28]. Regions with high insulation scores are more likely to be found inside TADs and regions among TADs intend to have low insulation scores. In total, 112 TAD boundaries at 10 kbp resolution were detected in our merged callset of 40 samples (**Table S44**). We illustrated the average and variance (maximum difference between any of the two samples) of insulation scores of each sample to indicate the changes of chromatin structures together with the corresponding methylation profiles and chrY assembly (**Figure S31b**). For the 340 DMRs which are detected in the aforementioned methylation analysis, we performed Kruskal-Wallis H tests (FDR 20%) with the same 6 meta haplogroups on the insulation scores (10 kbp resolution) in each DMR to detect regions that are differentially methylated as well as differentially insulated. Among the 26 DMRs that intersected with 21 differentially insulated regions (DIRs), we found one of such region (DMR: chrY-7289920-7290751, DIR: chrY-7290001-7300000) that harbors the PRKY gene which is both differentially DNA methylated and differentially expressed (**Table S45**).

Supplementary Figures

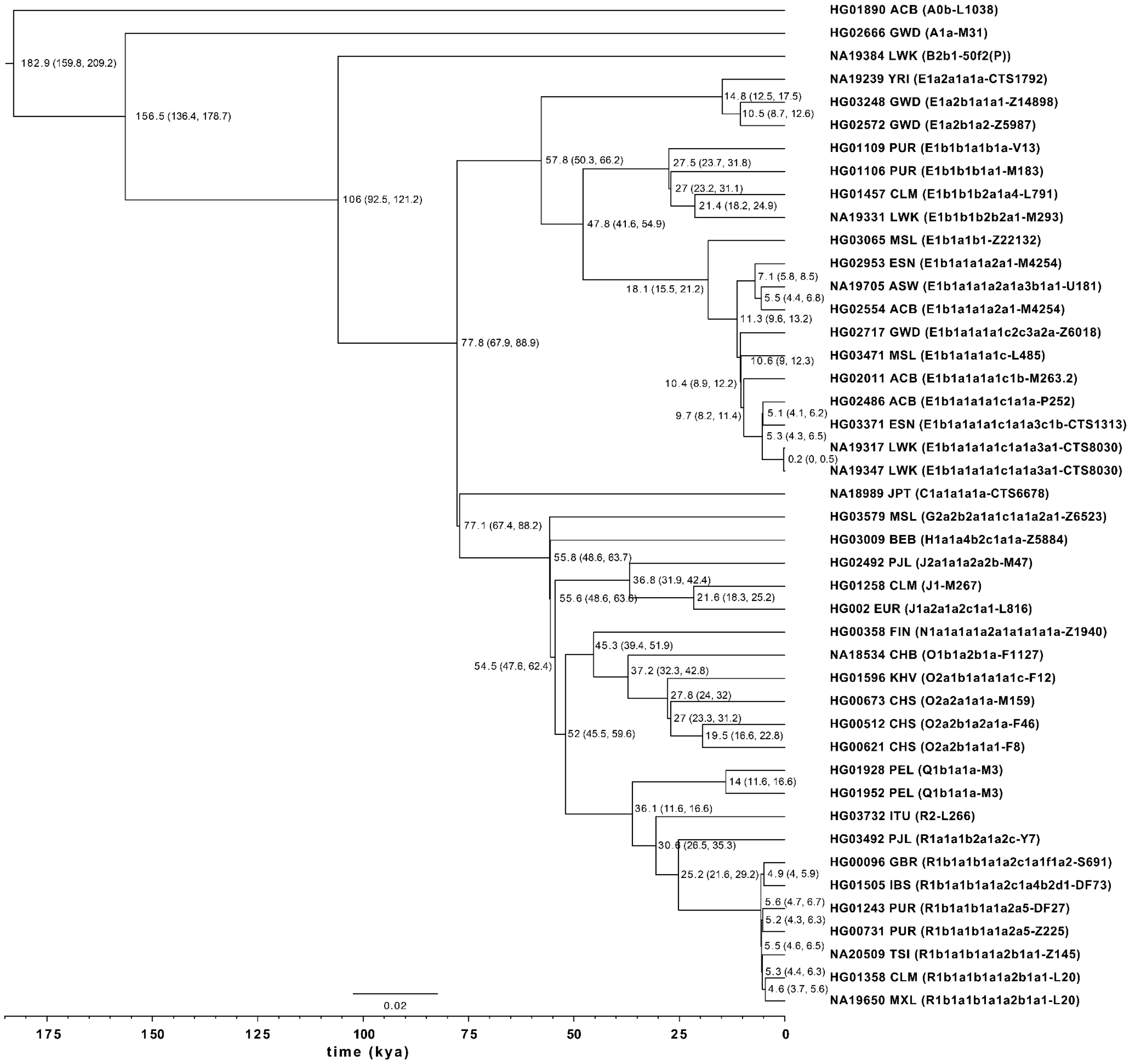

HG01890 ACB (A0b-L1038)
HG02666 GWD (A1a-M31)
NA19384 LWK (B2b1-50f2(P))
NA19239 YRI (E1a2a1a1a-CTS1792)
HG03248 GWD (E1a2b1a1a1-Z14898)
HG02572 GWD (E1a2b1a2-Z5987)
HG01109 PUR (E1b1b1a1b1a-V13)
HG01106 PUR (E1b1b1b1a1-M183)
HG01457 CLM (E1b1b1b2a1a4-L791)
NA19331 LWK (E1b1b1b2b2a1-M293)
HG03065 MSL (E1b1a1b1-Z22132)
HG02953 ESN (E1b1a1a1a2a1-M4254)
NA19705 ASW (E1b1a1a1a2a1a3b1a1-U181)
HG02554 ACB (E1b1a1a1a2a1-M4254)
HG02717 GWD (E1b1a1a1a1c2c3a2a-Z6018)
HG03471 MSL (E1b1a1a1a1c-L485)
HG02011 ACB (E1b1a1a1a1c1b-M263.2)
HG02486 ACB (E1b1a1a1a1c1a1a-P252)
HG03371 ESN (E1b1a1a1a1c1a1a3c1b-CTS1313)
NA19317 LWK (E1b1a1a1a1c1a1a3a1-CTS8030)
NA19347 LWK (E1b1a1a1a1c1a1a3a1-CTS8030)
NA18989 JPT (C1a1a1a1a-CTS6678)
HG03579 MSL (G2a2b2a1a1c1a1a2a1-Z6523)
HG03009 BEB (H1a1a4b2c1a1a-Z5884)
HG02492 PJL (J2a1a1a2a2b-M47)
HG01258 CLM (J1-M267)
HG002 EUR (J1a2a1a2c1a1-L816)
HG00358 FIN (N1a1a1a1a2a1a1a1a1a-Z1940)
NA18534 CHB (O1b1a2b1a-F1127)
HG01596 KHV (O2a1b1a1a1a1c-F12)
HG00673 CHS (O2a2a1a1a-M159)
HG00512 CHS (O2a2b1a2a1a-F46)
HG00621 CHS (O2a2b1a1a1-F8)
HG01928 PEL (Q1b1a1a-M3)
HG01952 PEL (Q1b1a1a-M3)
HG03732 ITU (R2-L266)
HG03492 PJL (R1a1a1b2a1a2c-Y7)
HG00096 GBR (R1b1a1b1a1a2c1a1f1a2-S691)
HG01505 IBS (R1b1a1b1a1a2c1a4b2d1-DF73)
HG01243 PUR (R1b1a1b1a1a2a5-DF27)
HG00731 PUR (R1b1a1b1a1a2a5-Z225)
NA20509 TSI (R1b1a1b1a1a2b1a1-Z145)
HG01358 CLM (R1b1a1b1a1a2b1a1-L20)
NA19650 MXL (R1b1a1b1a1a2b1a1-L20)
182.9 (159.8, 209.2)
156.5 (136.4, 178.7)
106 (92.5, 121.2)
14.8 (12.5, 17.5)
10.5 (8.7, 12.6)
57.8 (50.3, 66.2)
27.5 (23.7, 31.8)
27 (23.2, 31.1)
21.4 (18.2, 24.9)
47.8 (41.6, 54.9)
18.1 (15.5, 21.2)
7.1 (5.8, 8.5)
5.5 (4.4, 6.8)
11.3 (9.6, 13.2)
10.6 (9, 12.3)
10.4 (8.9, 12.2)
9.7 (8.2, 11.4)
5.1 (4.1, 6.2)
5.3 (4.3, 6.5)
0.2 (0, 0.5)
77.8 (67.9, 88.9)
77.1 (67.4, 88.2)
55.8 (48.6, 63.7)
36.8 (31.9, 42.4)
55.6 (48.6, 63.6)
21.6 (18.3, 25.2)
54.5 (47.6, 62.4)
45.3 (39.4, 51.9)
37.2 (32.3, 42.8)
27.8 (24, 32)
27 (23.3, 31.2)
19.5 (16.6, 22.8)
52 (45.5, 59.6)
14 (11.6, 16.6)
36.1 (11.6, 16.6)
30.6 (26.5, 35.3)
25.2 (21.6, 29.2)
4.9 (4, 5.9)
5.6 (4.7, 6.7)
5.2 (4.3, 6.3)
5.5 (4.6, 6.5)
5.3 (4.4, 6.3)
4.6 (3.7, 5.6)
0.02
175
150
125
100
75
50
25
0
time (kya)

Figure S1. Phylogenetic relationships of the analyzed Y chromosomes. Split times as estimated according to the BEAST analysis are shown with 95% HPD interval in brackets (kya - thousand years ago). Sample ID is followed by population designation, full Y haplogroup label according to ISOGG v15.73 and terminal marker ID. Population abbreviations: ACB - African Caribbean in Barbados; ASW - African Ancestry in SW USA; BEB - Bengali in Bangladesh; CHB - Han Chinese in Beijing, China; CHS - Han Chinese South; CLM - Colombian in Medellín, Colombia; ESN - Esan in Nigeria; FIN - Finnish in Finland; GBR - British From England and Scotland; GWD - Gambian in Western Division – Mandinka; IBS - Iberian Populations in Spain; ITU - Indian Telugu in the U.K.; JPT - Japanese in Tokyo, Japan; KHV - Kinh in Ho Chi Minh City, Vietnam; LWK - Luhya in Webuye, Kenya; MSL - Mende in Sierra Leone; MXL - Mexican Ancestry in Los Angeles CA USA; PEL - Peruvian in Lima Peru; PJL - Punjabi in Lahore, Pakistan; PUR - Puerto Rican in Puerto Rico; TSI - Toscani in Italia; YRI - Yoruba in Ibadan, Nigeria.

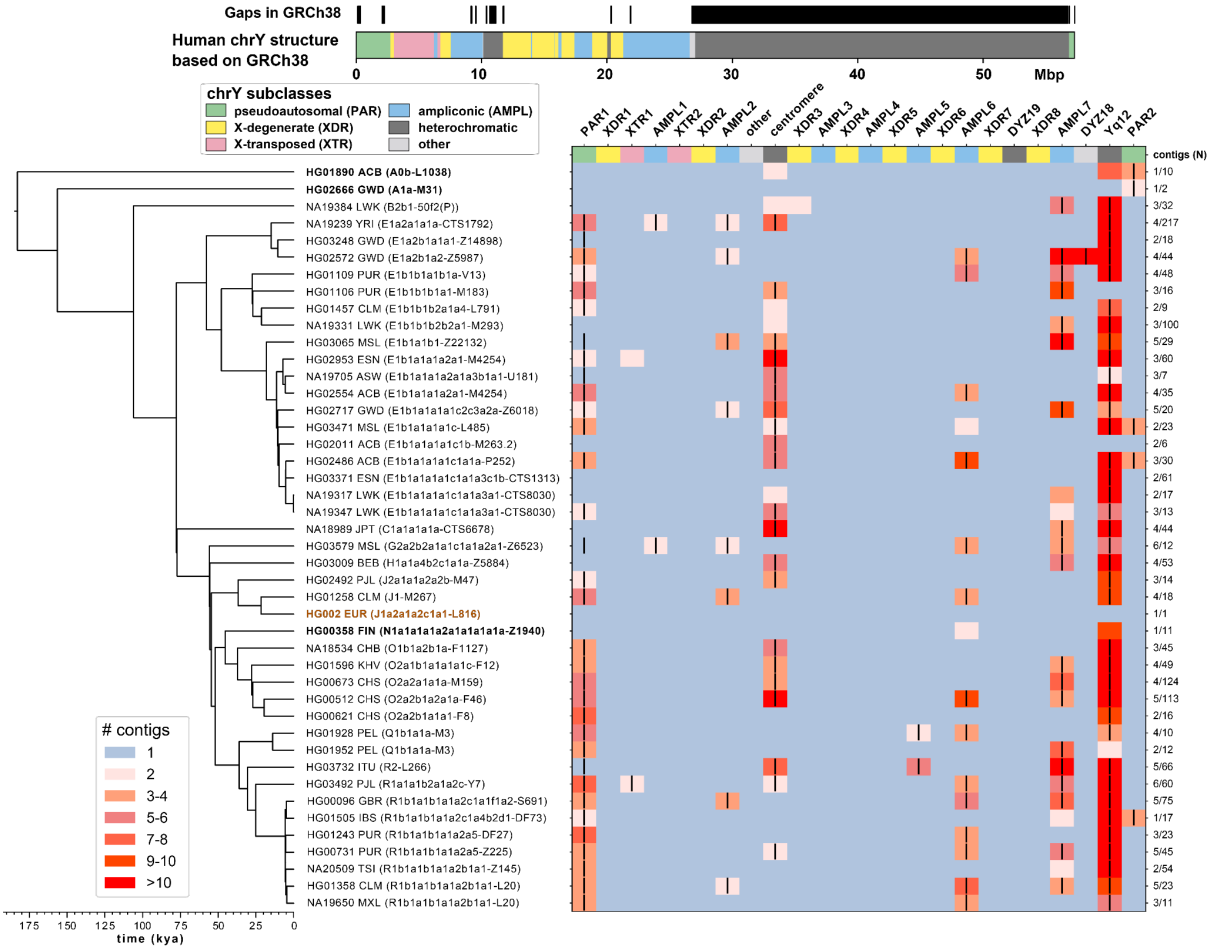


Figure S2. Phylogenetic relationships of the analyzed Y chromosomes and assembly completeness. Phylogenetic relationships of the analyzed Y chromosomes with branch lengths drawn proportional to the estimated times between successive splits according to BEAST analysis. Summary of Y assembly completeness with the number of contigs containing sequence from specific sequence class indicated with different colors (on the right - number of Y contigs needed to achieve the plotted assembly contiguity/total number of assembled Y contigs for each sample). Sample IDs include the population abbreviation, and the full Y lineage and terminal marker in brackets. See Figure S1 for population abbreviations.

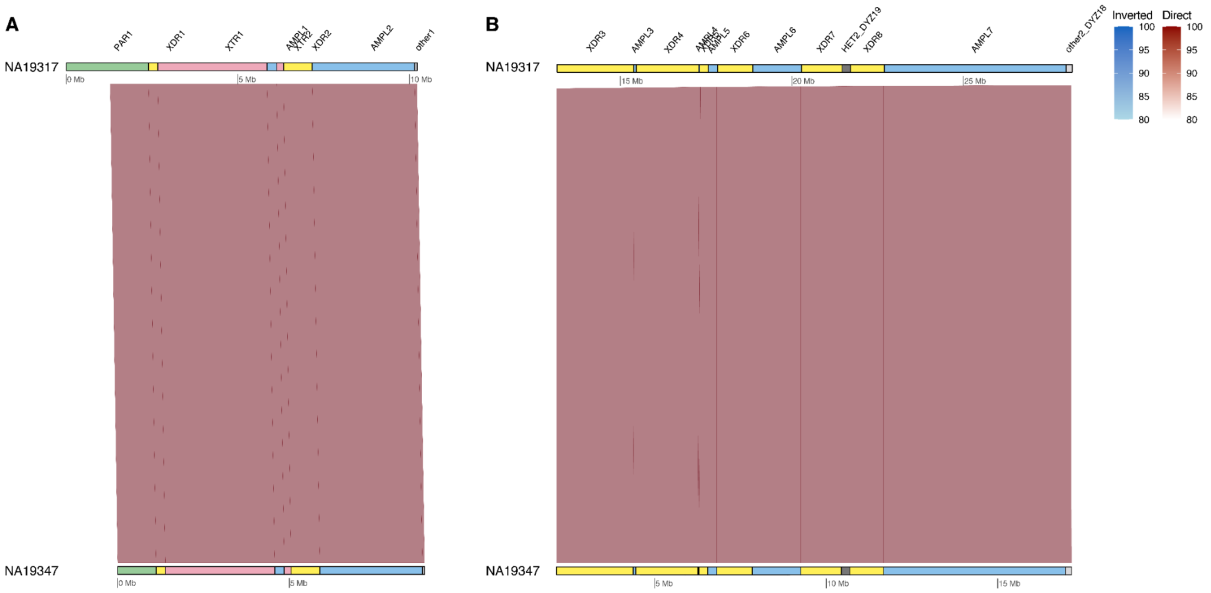


Figure S3. Comparison of the Y assemblies from closely related African Y chromosomes (NA19317 vs NA19347). Comparison of contiguously assembled regions spanning: **A.** from PAR1 until the end of other1, and **B.** from XDR3 to the end of *DYZ18*. Pairwise sequence alignments of 21/24 contiguously assembled Y-chromosomal subregions showed sequence identity ranging from 99.982% to 100% (**Table S6**), with 100% sequence identify in three subregions (other1, *DYZ19* and *DYZ18*). Eight subregions (XDR1, XTR2, XDR2, XDR3, AMPL3, AMPL4, XDR6, and XDR8) have no substitutions and the number of indels range from 2 to 26. XTR1 subregion shows the lowest sequence identity (99.982%) with 185 mismatches and 389 indels/gaps in the alignment.

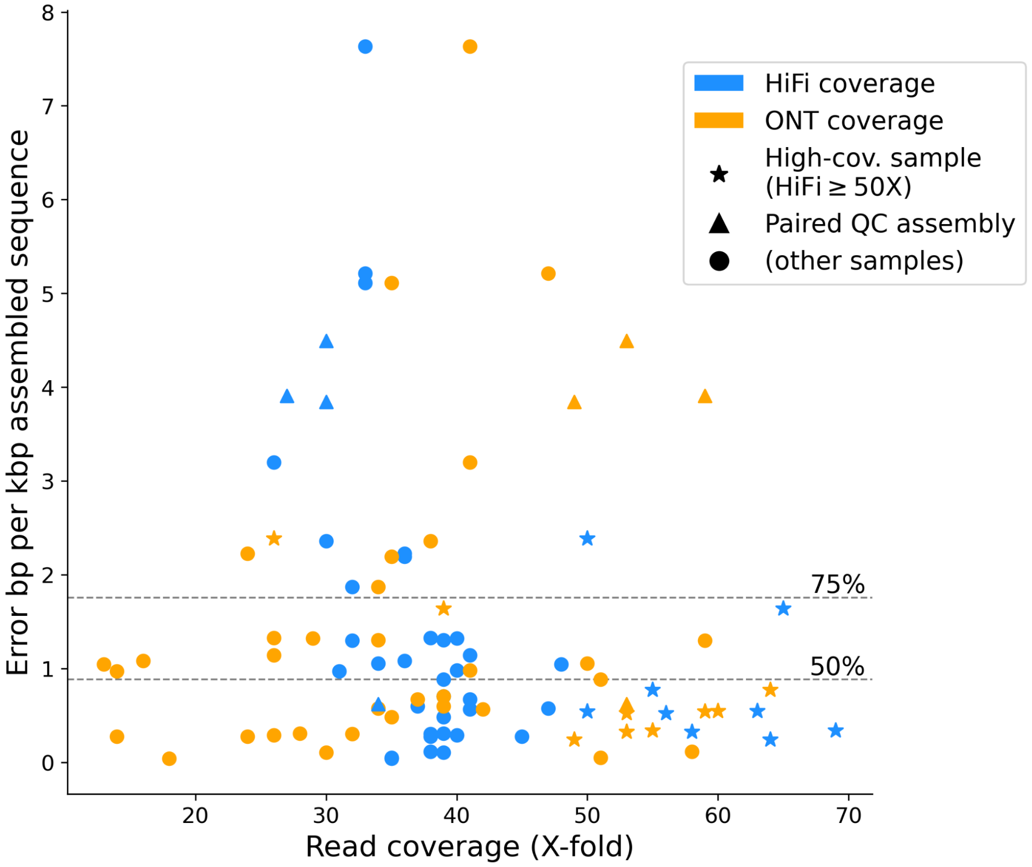


Figure S4. Scatter plot of input read coverage for both ONT (orange) and HiFi (blue) per sample (X-fold coverage relative to a ~3.1 Gbp genome size, x-axis) and putative assembly errors (flagged bp per kbp assembled sequence, y-axis). "Star" markers highlight high-coverage samples. "Triangle" markers indicate assemblies created for QC purposes using approximately half of the HiFi coverage of the respective high-coverage sample. Dashed horizontal lines indicate the second and third quartile of samples.

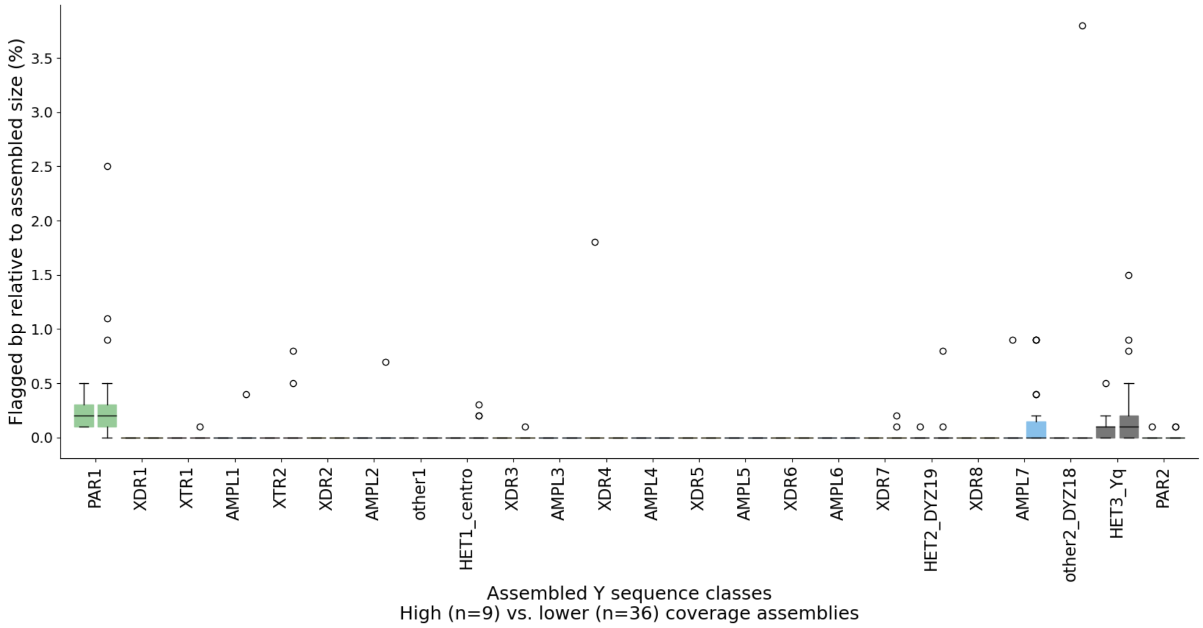


Figure S5. Comparison of putative assembly errors in high- and lower-coverage assemblies per Y sequence class. Errors are depicted as percent of bp flagged as potentially erroneous for high-coverage (n=9, left boxplots) and lower-coverage assemblies (n=36, right boxplots). Boxplots are colored according to the Y sequence class (**Fig. 1a**). Distributions of annotated errors were compared per each sequence class using a two-sided Mann-Whitney-U test. The differences are not statistically significant at α = 0.05 after multiple testing correction (Benjamini-Hochberg).

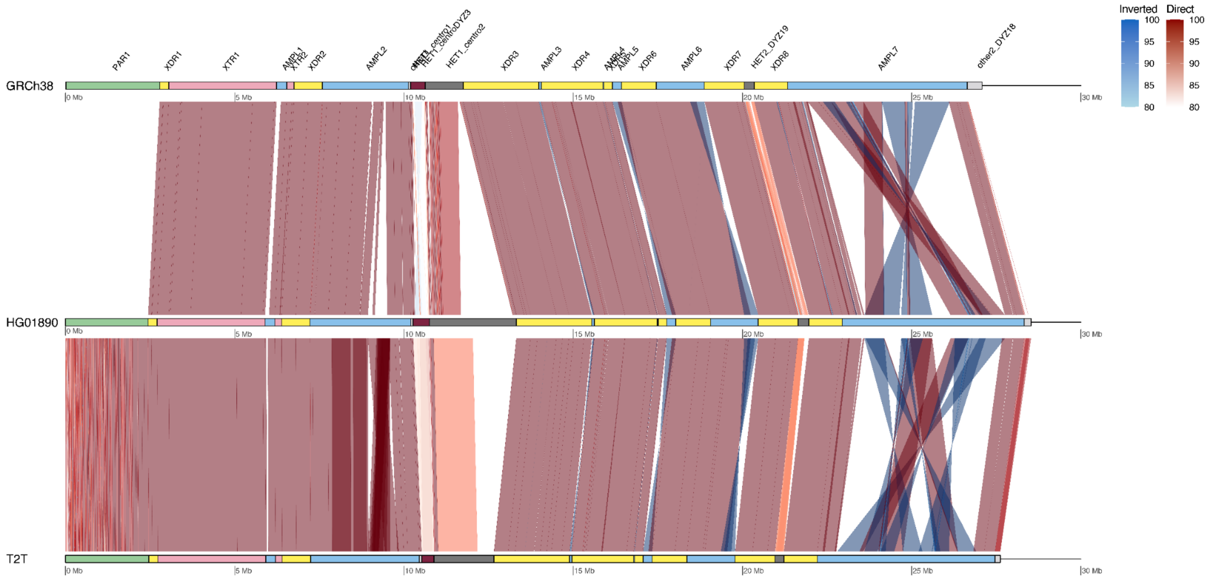


Figure S6: Comparison between GRCh38, HG01890 and T2T Y.

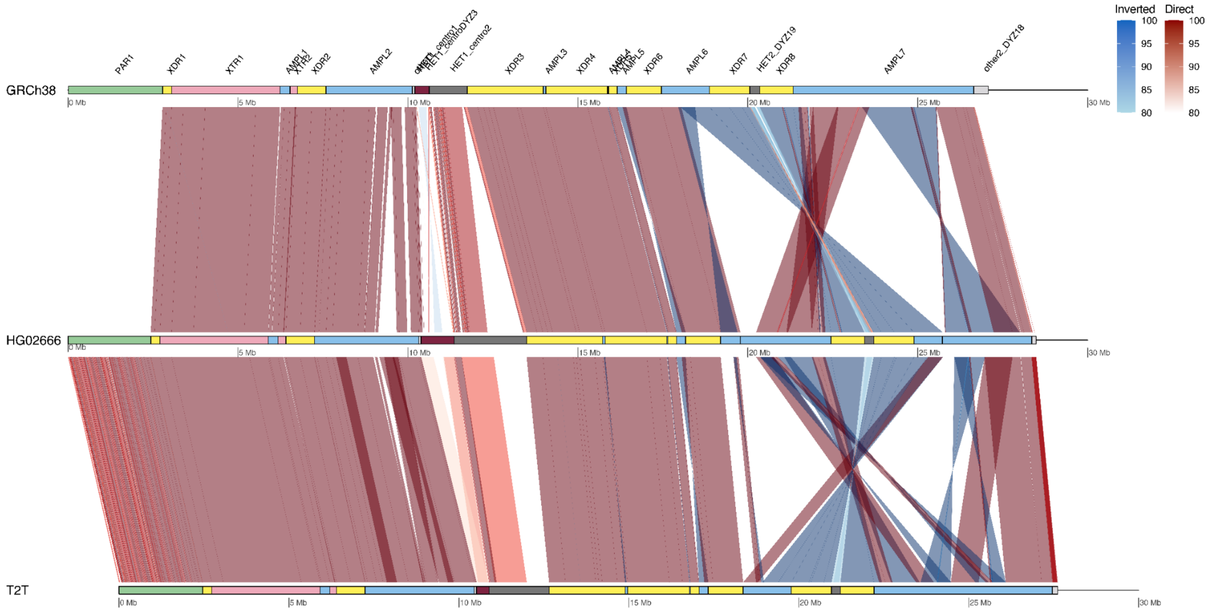


Figure S7: Comparison between GRCh38, HG02666 and T2T-T.

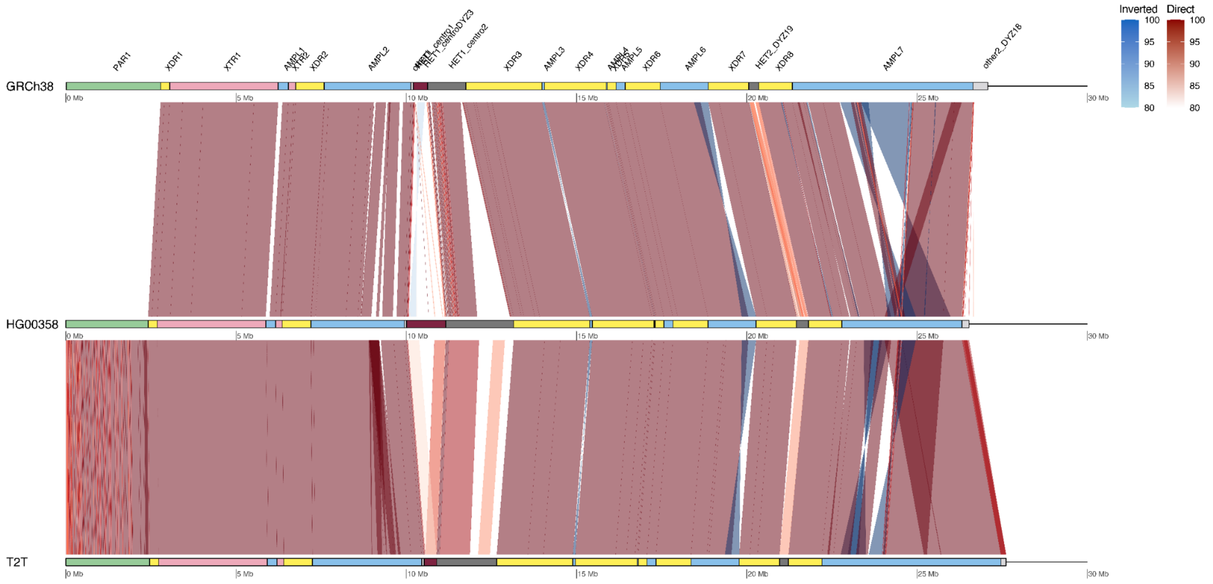


Figure S8: Comparison between GRCh38, HG00358 and T2T Y.

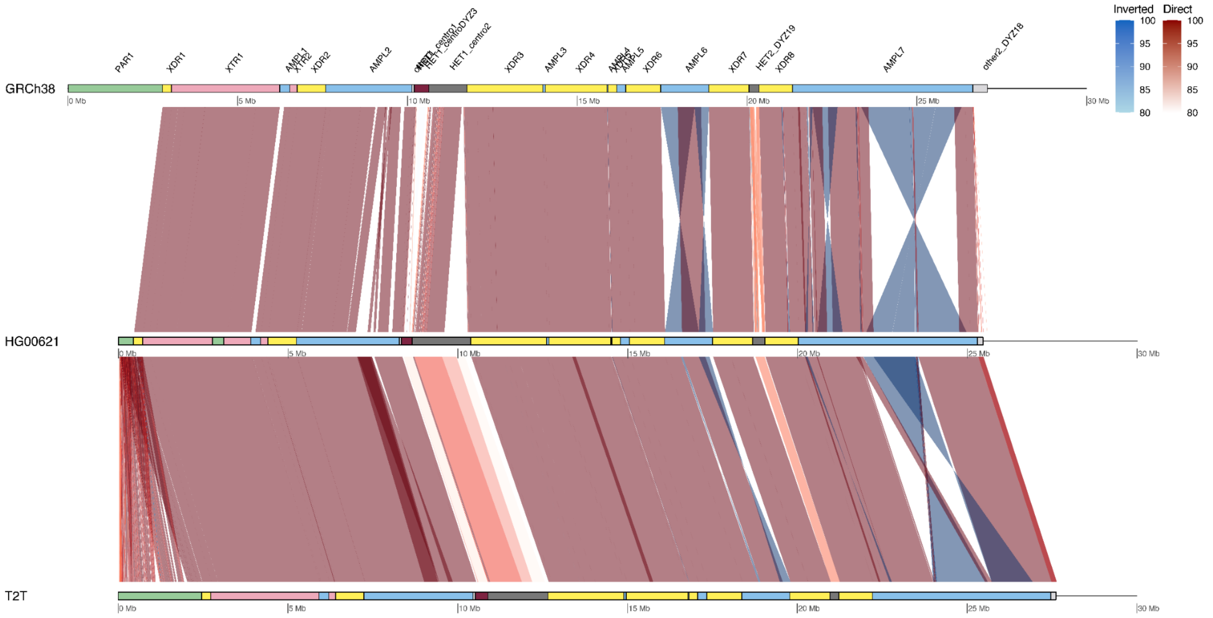


Figure S9: Comparison between GRCh38, HG00621 and T2T Y.

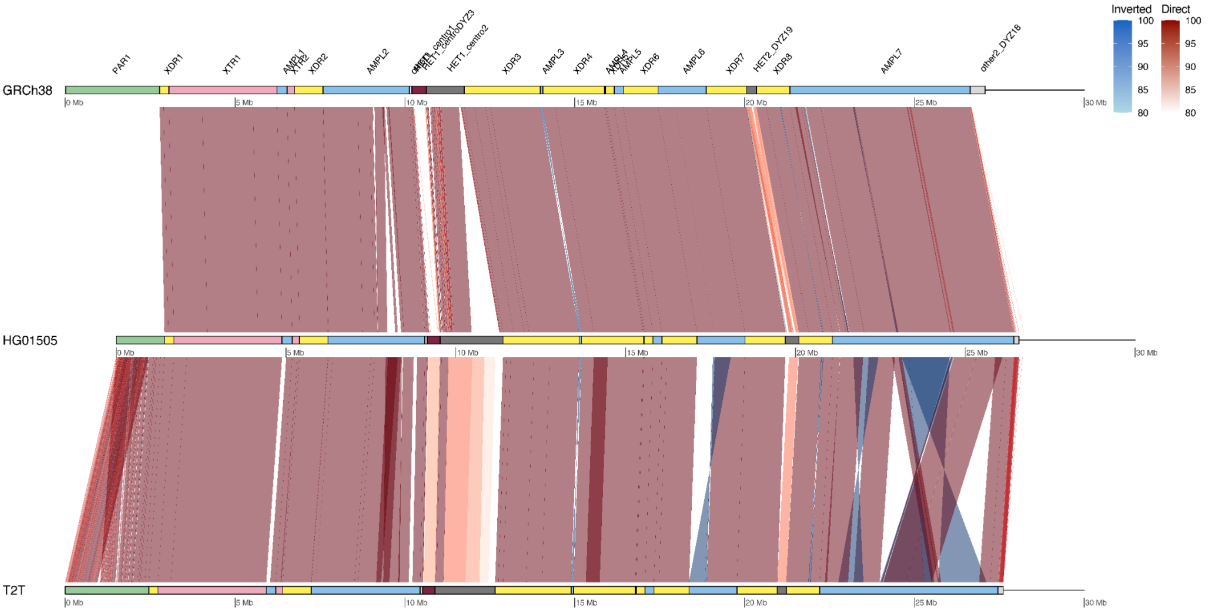


Figure S10: Comparison between GRCh38, HG01505 and T2T Y. Note - GRCh38 and HG01505 are phylogenetically closely related, both representing haplogroup R1b. Highly similar assembly and lack of large difference between GRCh38 and HG01505 supports accuracy of our de novo assemblies.

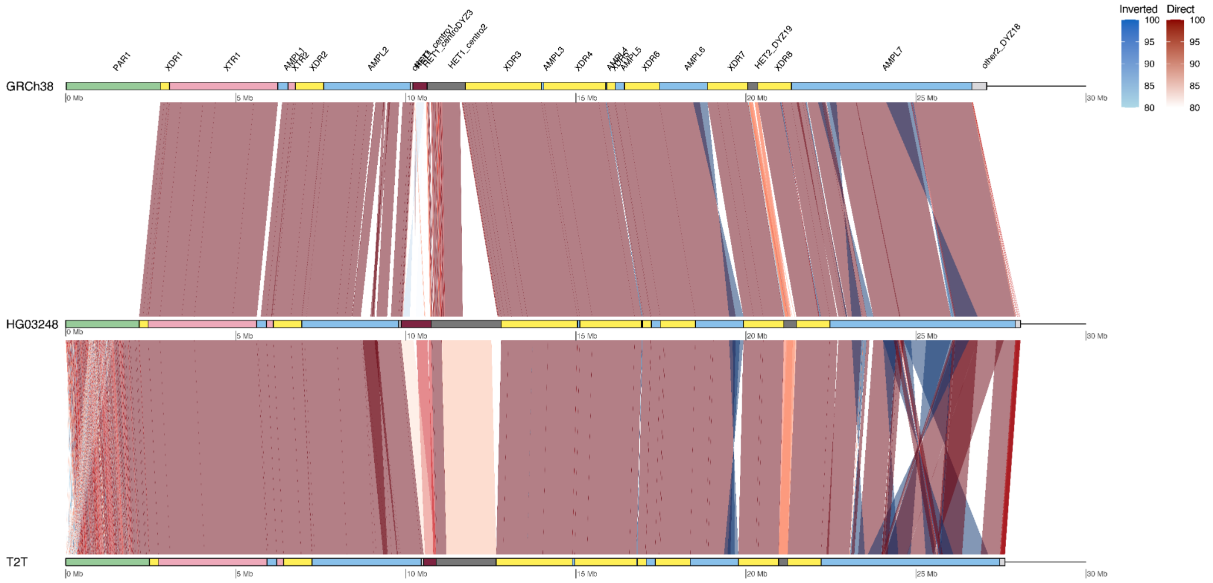


Figure S11: Comparison between GRCh38, HG03248 and T2T Y.

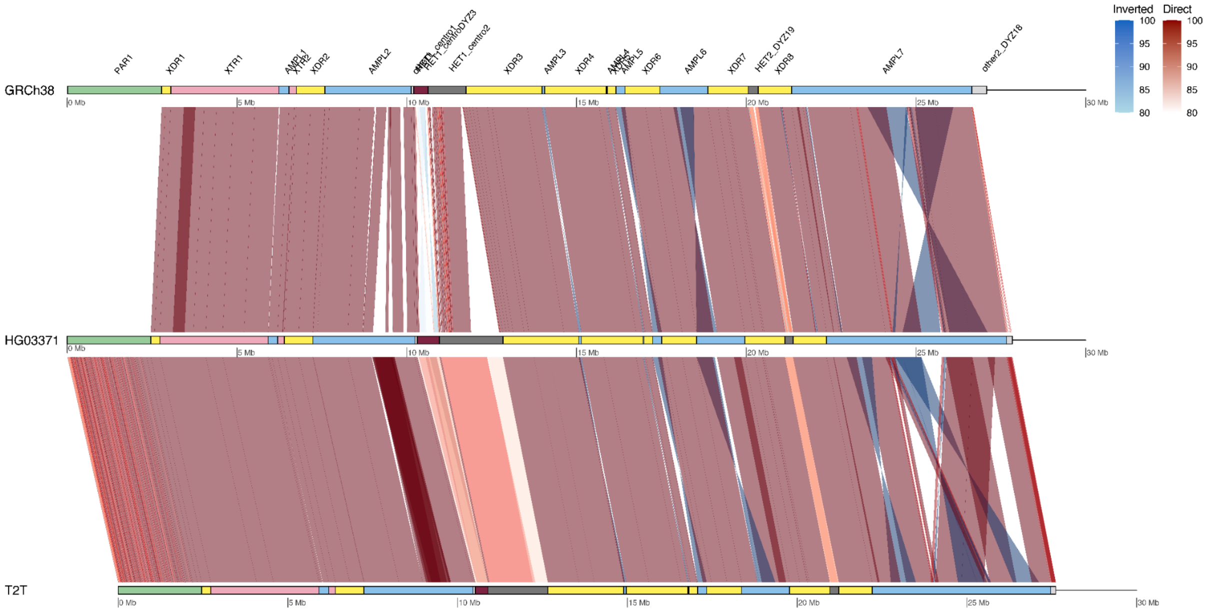


Figure S12: Comparison between GRCh38, HG03371 and T2T Y.

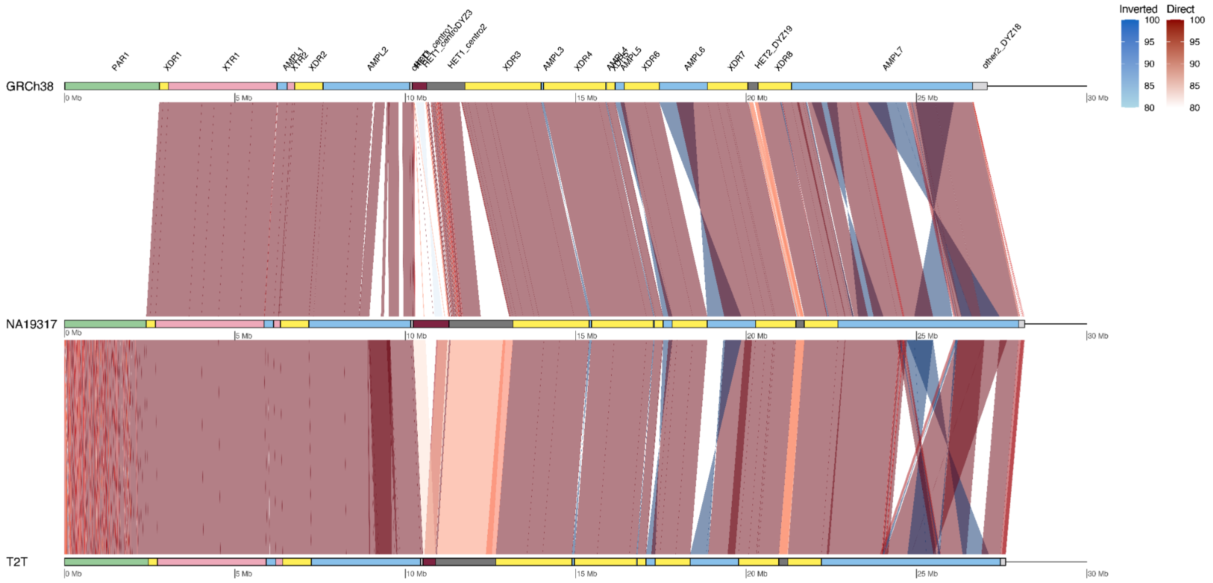


Figure S13: Comparison between GRCh38, NA19317 and T2T Y.

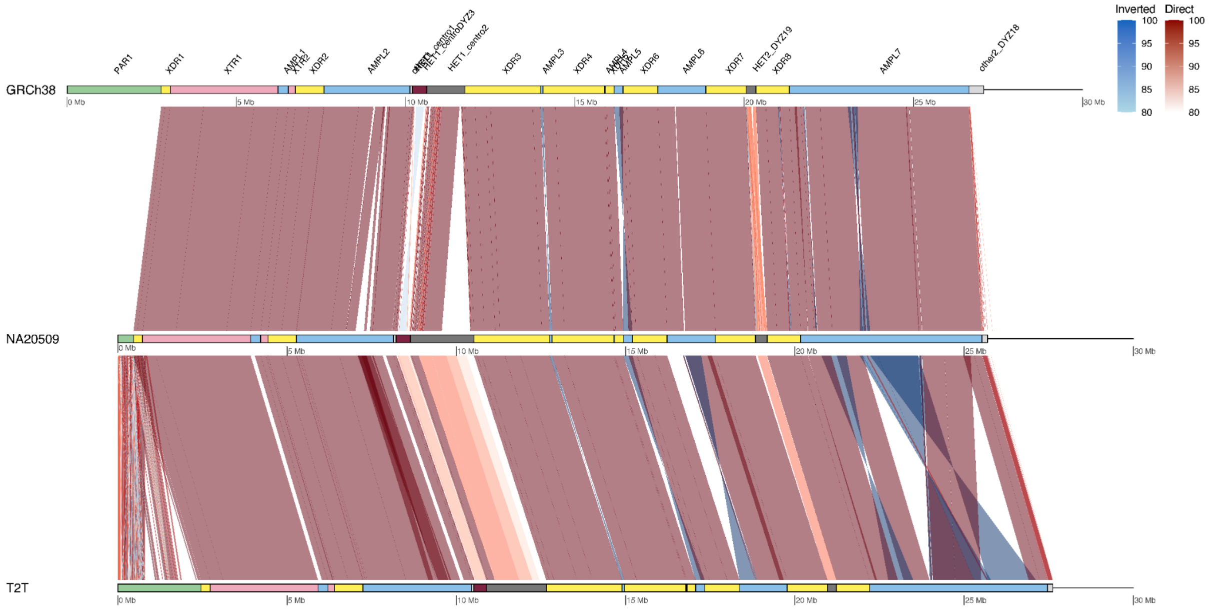


Figure S14: Comparison between GRCh38, NA20509 and T2T Y.

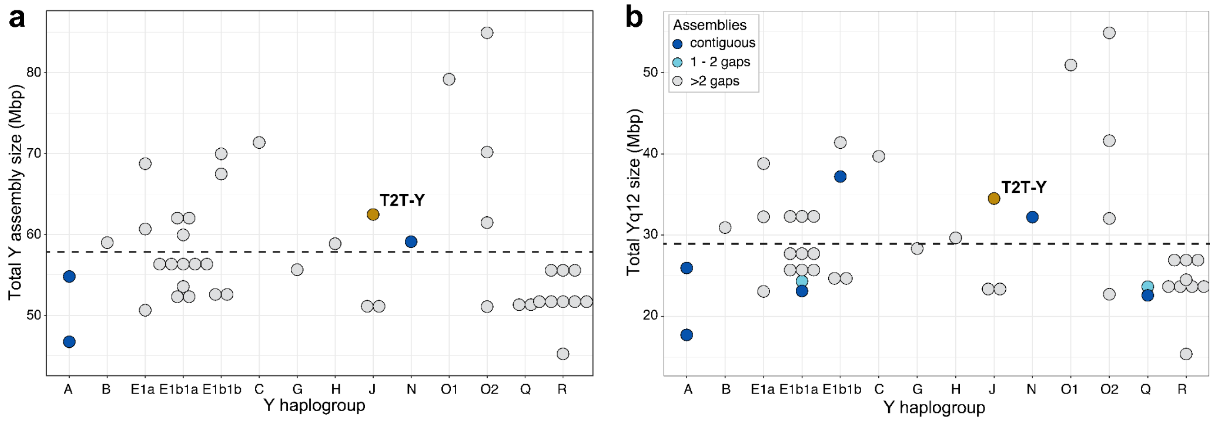


Figure S15. Y assembly sizes across Y haplogroups. **a**. The total combined Y assembly size. **b**. The total combined Yq12 subregion size. Samples with contiguous assembly, with 1-2 or more gaps and the T2T Y are indicated with different colours. Black dashed line indicated the mean (57.6 Mbp for total Y assembly and 29.0 Mbp for the Yq12 subregion).

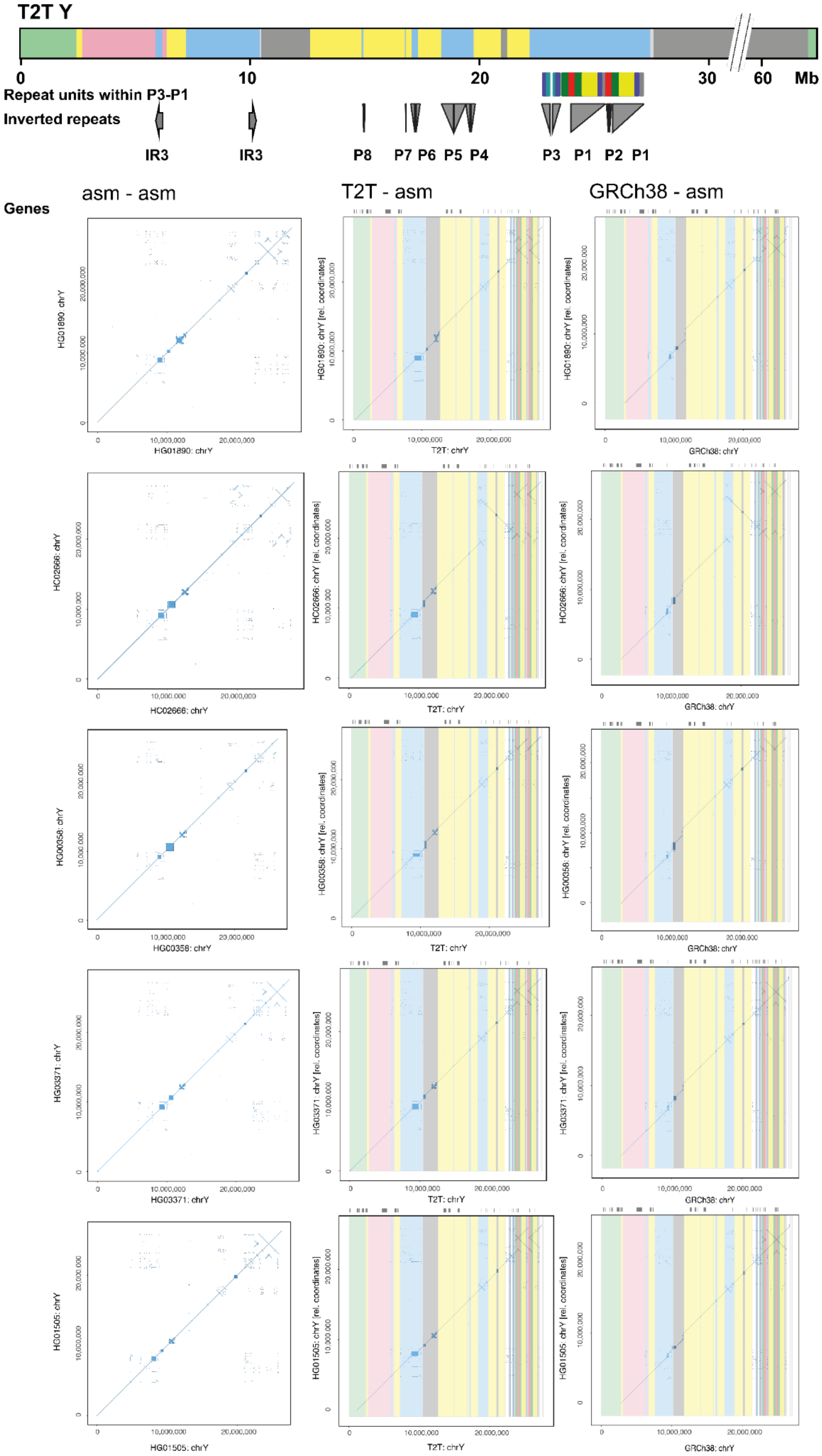


Figure S16. Dotplots of five samples contiguously assembled across the euchromatic regions (from PAR1 until Yq12 heterochromatic region) with self dotplot on the left, compared to T2T Y in center and to GRCh38 on the right, annotated with sequence classes and SD repeat units in ampliconic 7 region.

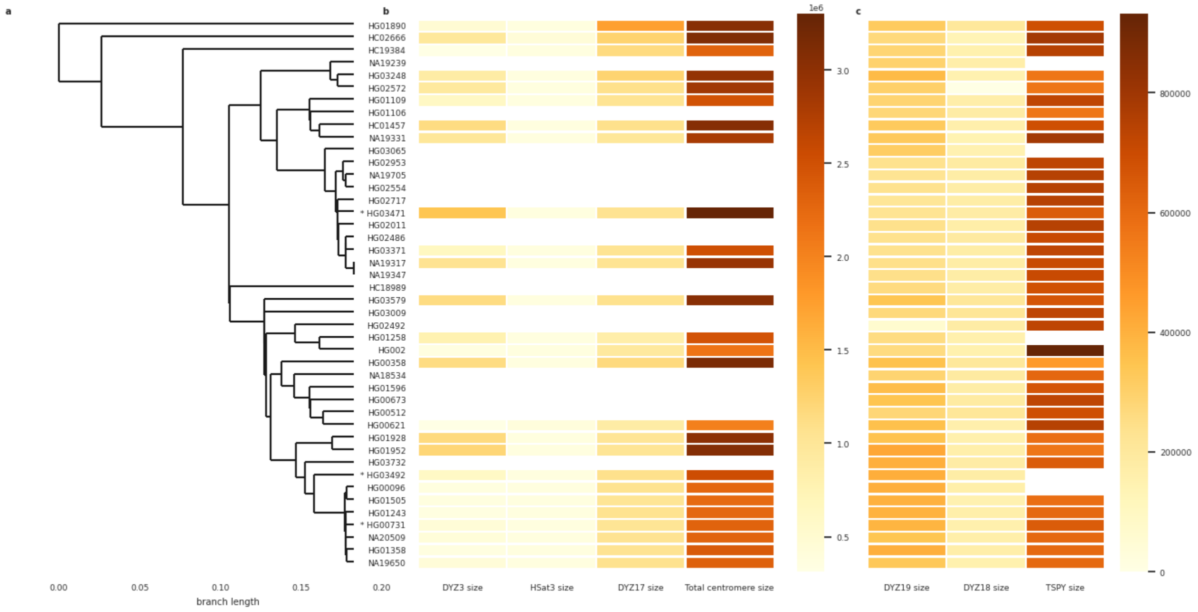


Figure S17. Size variation of the (peri-)centromeric region and repeat arrays (*DYZ3* alpha-satellite array, *Hsat3*, *DYZ17* array, and total (peri-)centromeric region) on the left and the *DYZ19*, *DYZ18*, and the TSPY copy-number variable repeat arrays on the right, with sizes shown as a heatmap. **a.** Phylogenetic clustering of the samples, as described in **Fig. S1**. **b.** Size variation heatmap for each pericentromeric region, and the total centromere size in millions of base pairs. White fill indicates that the size information of the region is not available due to non-contiguous assembly of the region. Asterisk to the left of the sample name indicates samples (HG00731, HG03471, and HG03492) with one assembly gap in the (peri-)centromeric region. **c.** Size variation heatmap for *DYZ19*, *DYZ18* and TSPY repeat arrays. The sizes of the (peri-)centromeric regions (*DYZ3* alpha-satellite array, *Hsat3*, and *DYZ17* array) were regressed against each other, but none achieved significant correlations.

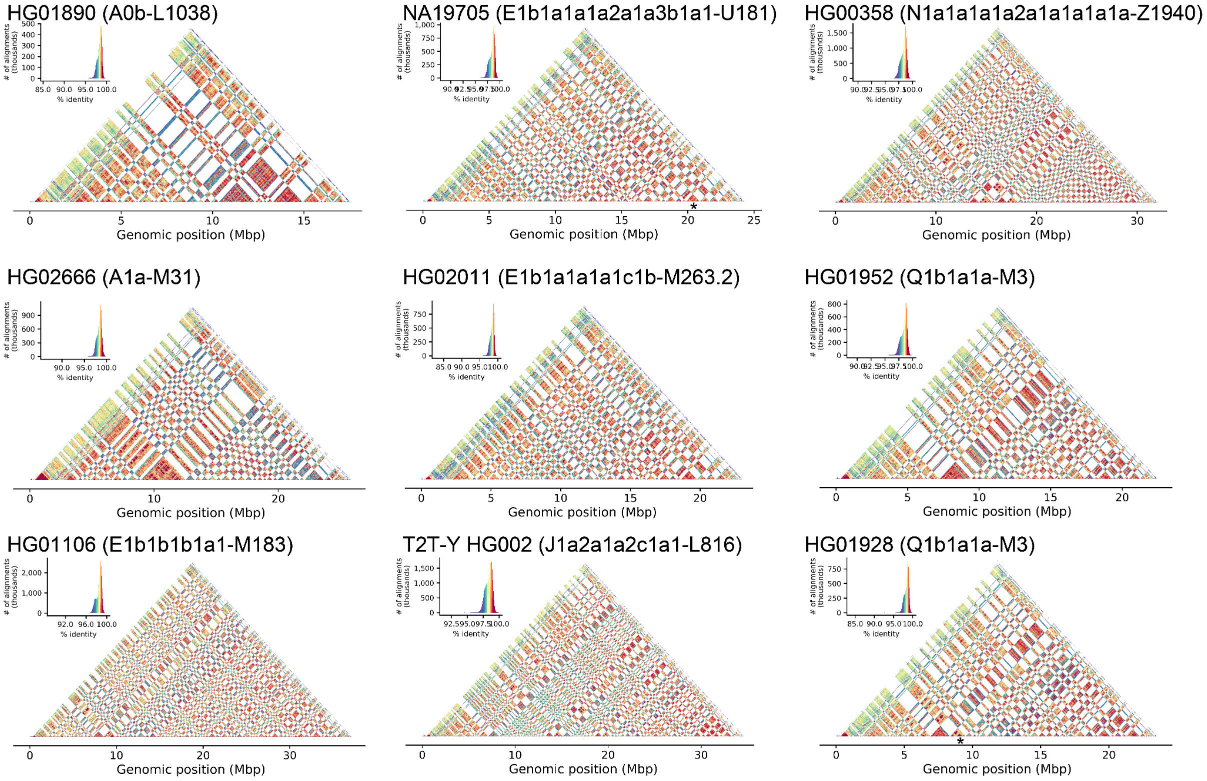


Figure S18. Sequence identity heatmaps of the Yq12 subregion for six contiguously assembled samples (HG01890, HG02666, HG01106, HG02011, HG00358 and HG01952), two samples (NA19705 and HG01928) with a single gap in the Yq12 subregion (gap location marked with asterisk) and the T2T Y from HG002 using 5kb window size.

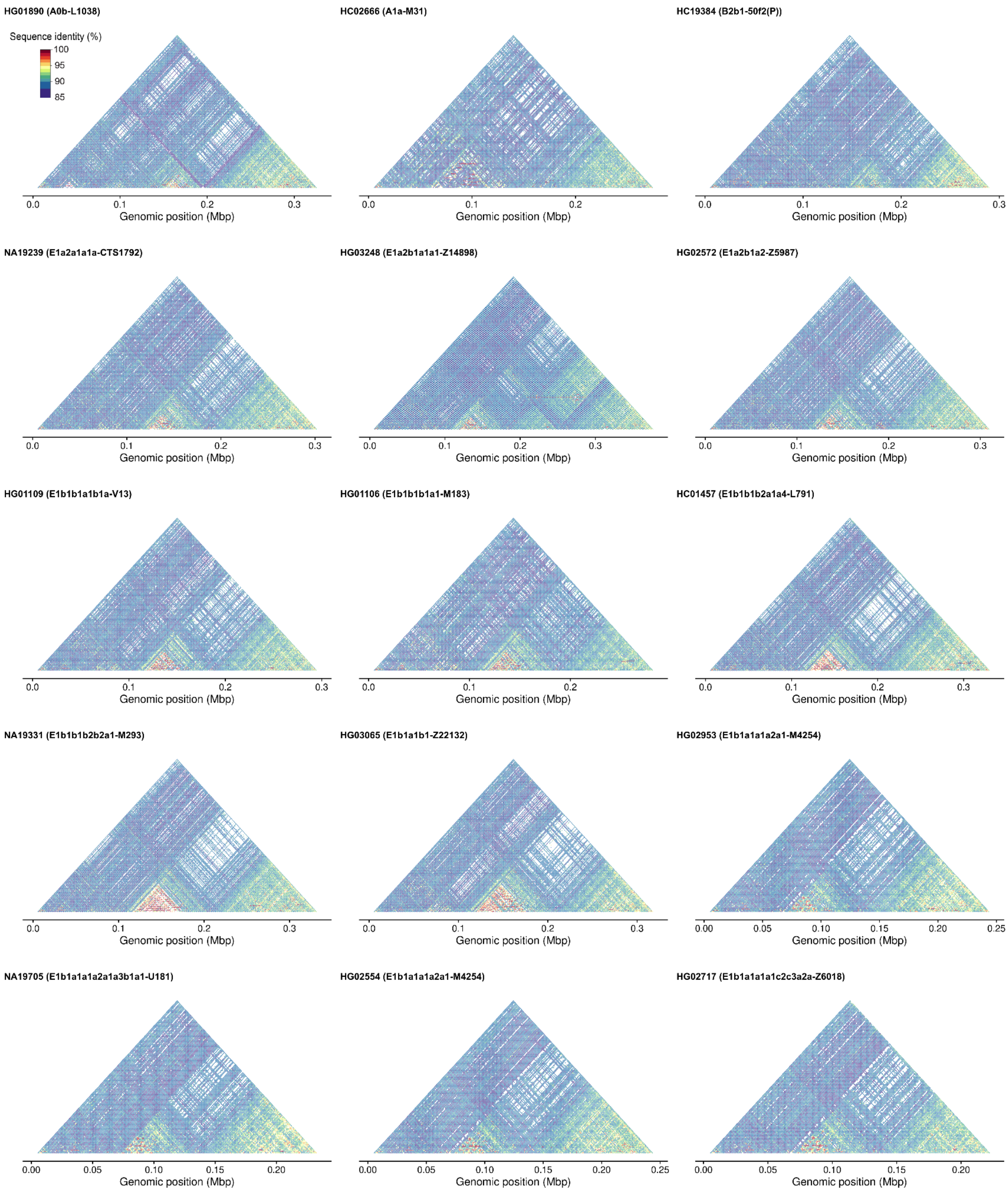


Figure S19. Sequence identity heatmaps of the *DYZ19* subregion across samples, including the T2T Y, in phylogenetic order. 5000 bp of flanking sequence was added to the *DYZ19* genomic interval and 1 kbp window size was used when running StainedGlass.

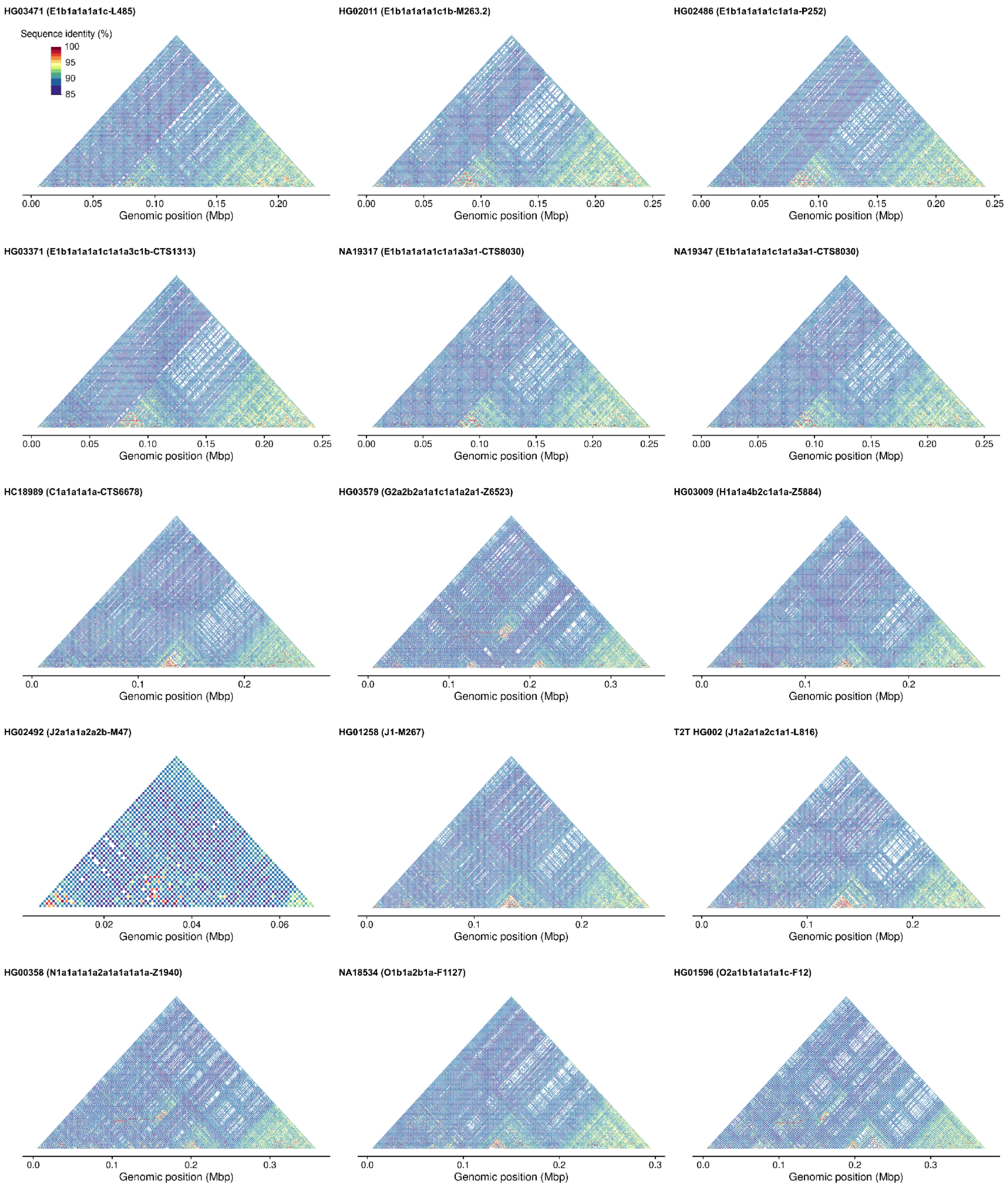


Figure S20. Sequence identity heatmaps of the *DYZ19* subregion across samples, including the T2T Y, in phylogenetic order. 5000 bp of flanking sequence was added to the *DYZ19* genomic interval and 1 kbp window size was used when running StainedGlass.

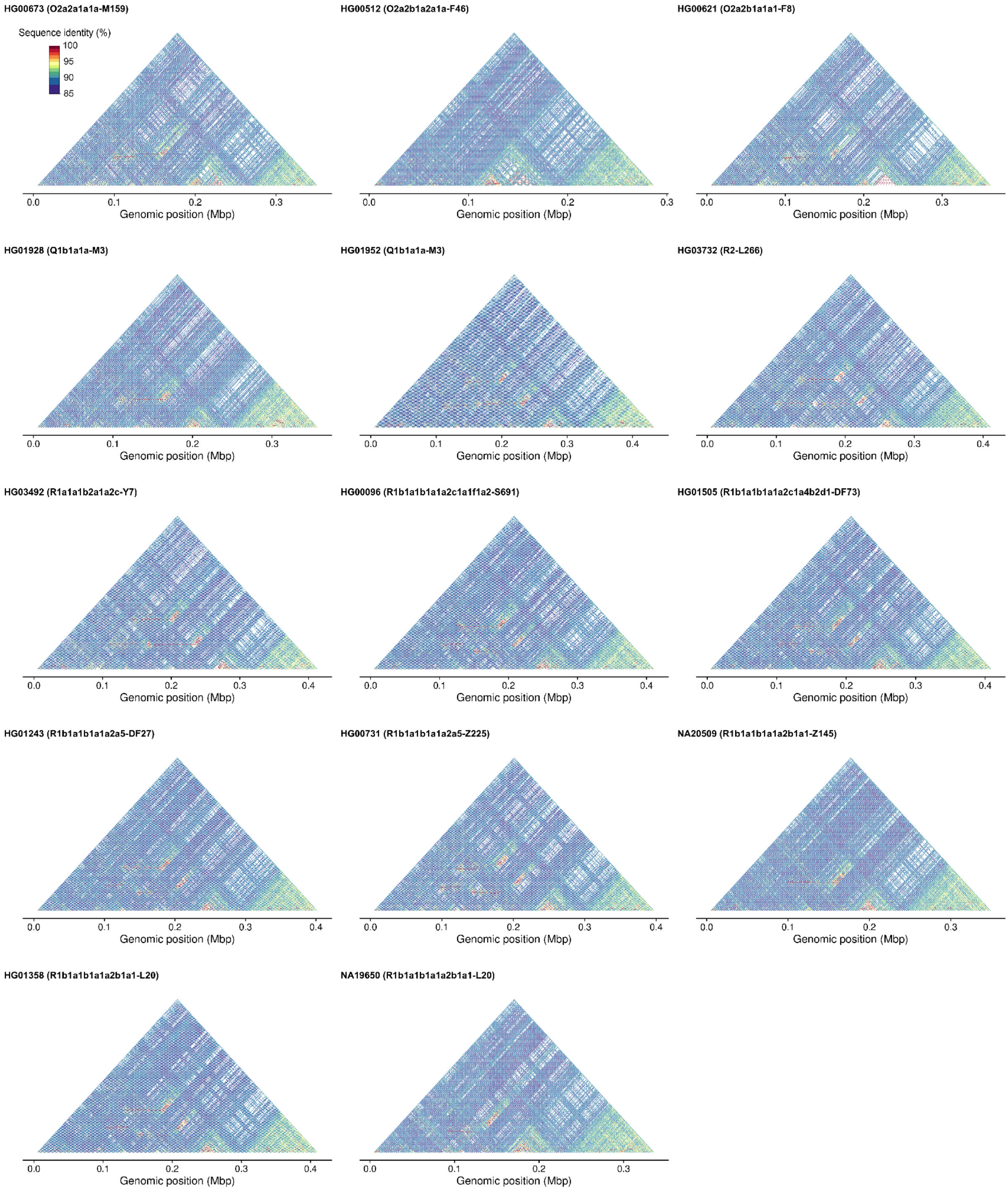


Figure S21. Sequence identity heatmaps of the *DYZ19* subregion across samples, including the T2T Y, in phylogenetic order. 5000 bp of flanking sequence was added to the *DYZ19* genomic interval and 1 kbp window size was used when running StainedGlass.

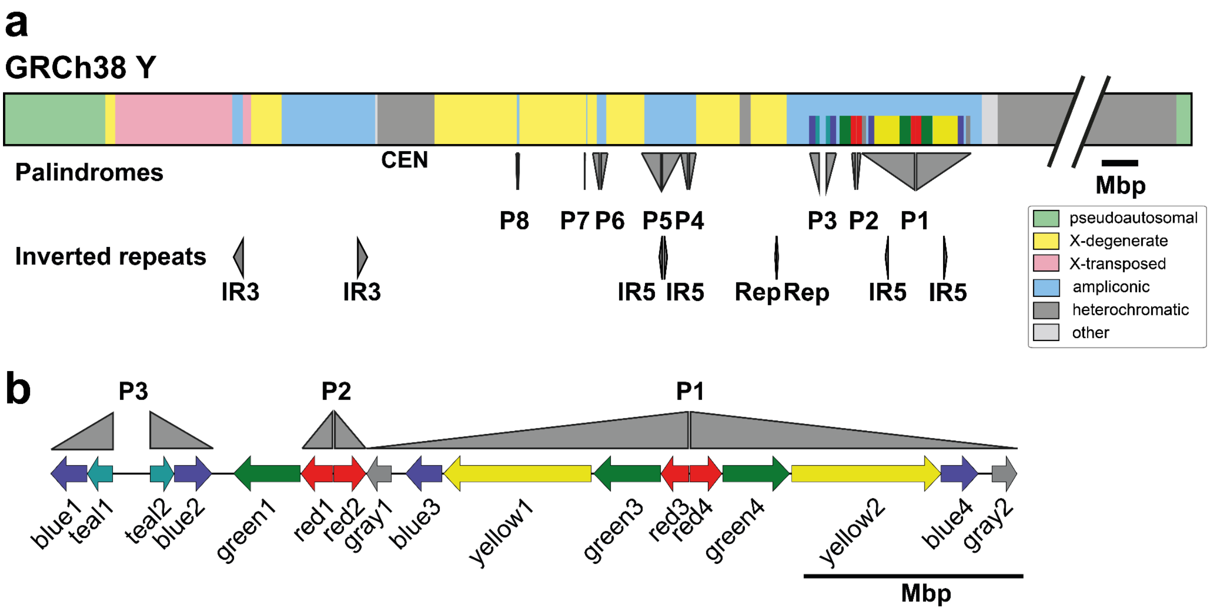


Figure S22. Schematic representation of inverted repeats involved in inversions. **a.** The GRCh38 Y reference structure with annotations of segmental duplications in *AZFc*/ampliconic subregion 7, with palindromes (P8-P1) and inverted repeats (IR) shown below. The repeat coordinates relative to GRCh38 Y reference sequence were obtained from Teitz et al [7]. **b.** Annotation of segmental duplications in *AZFc*/ampliconic subregion 7 following the naming originally proposed by Kuroda-Kawaguchi et al [29].

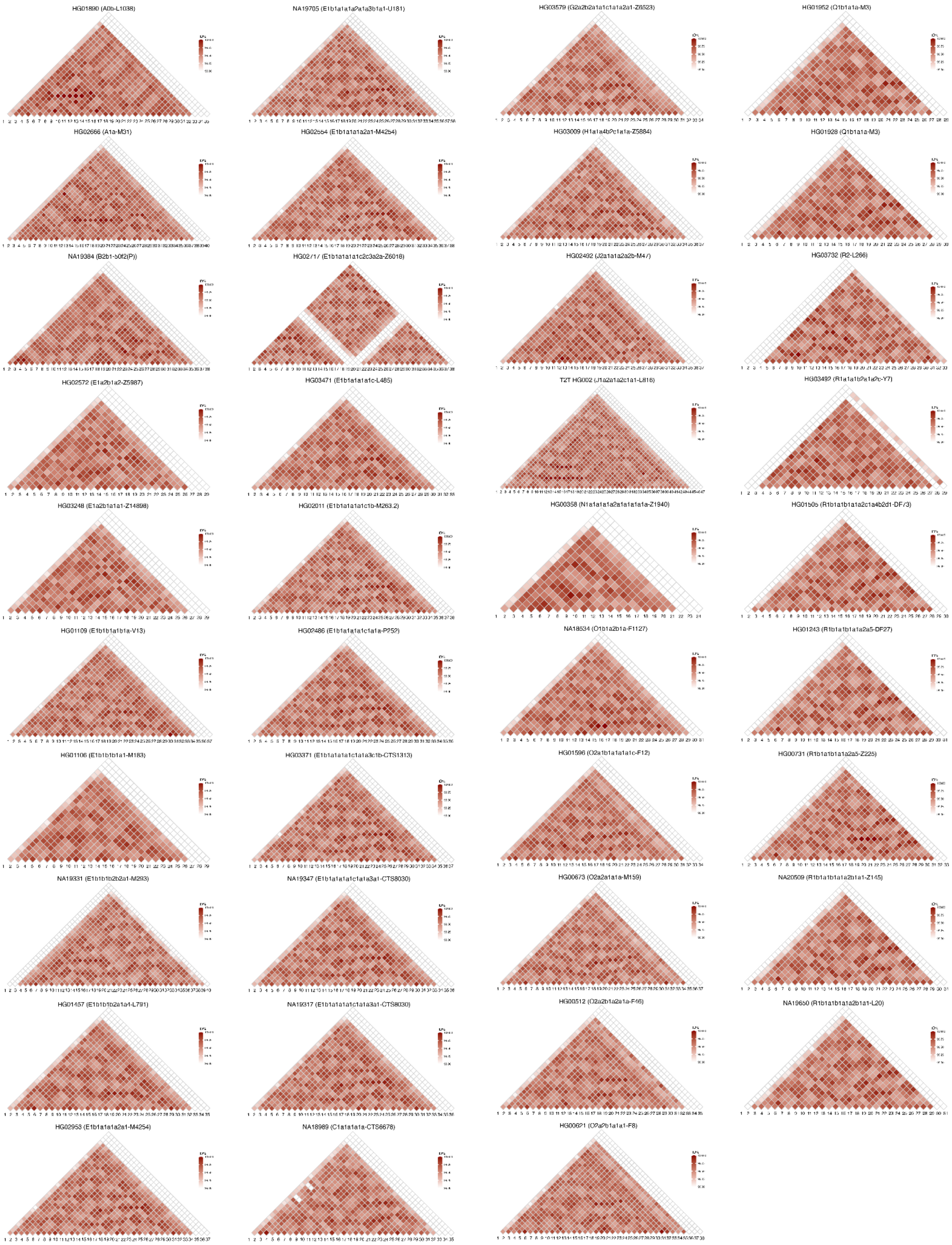


Figure S23. Sequence identity heatmaps of ~20.3-kbp long TSPY repeat units for 39 males in phylogenetic order (from top to down from the deepest-rooting sample). Red shades from lighter to darker indicate sequence identity from 99-100%, respectively, while white fill indicates sequence identity below 99%.

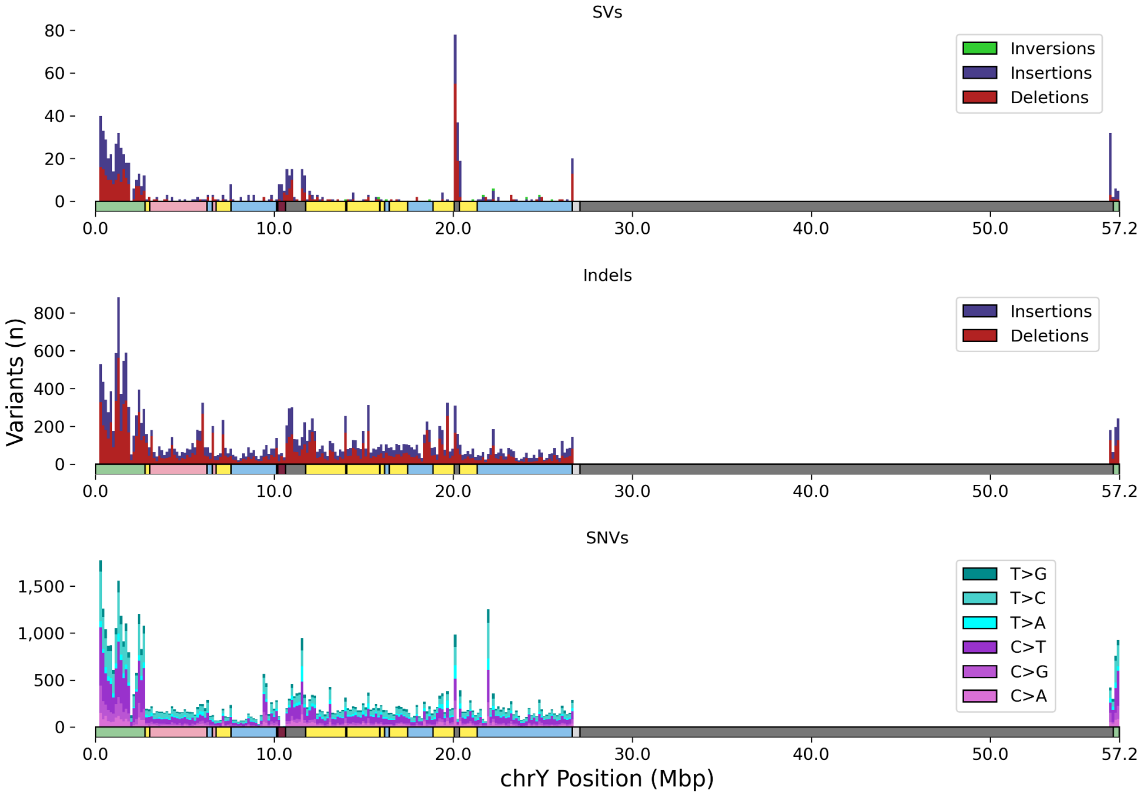


Figure S24. Distribution of variant sizes for SVs (≥ 50 bp, top), Indels (< 50 bp, middle), and SNV (bottom) across the Y chromosome (color by region) as identified using PAV. High peaks in heterochromatin are apparent for SVs, but not SNVs and indels.

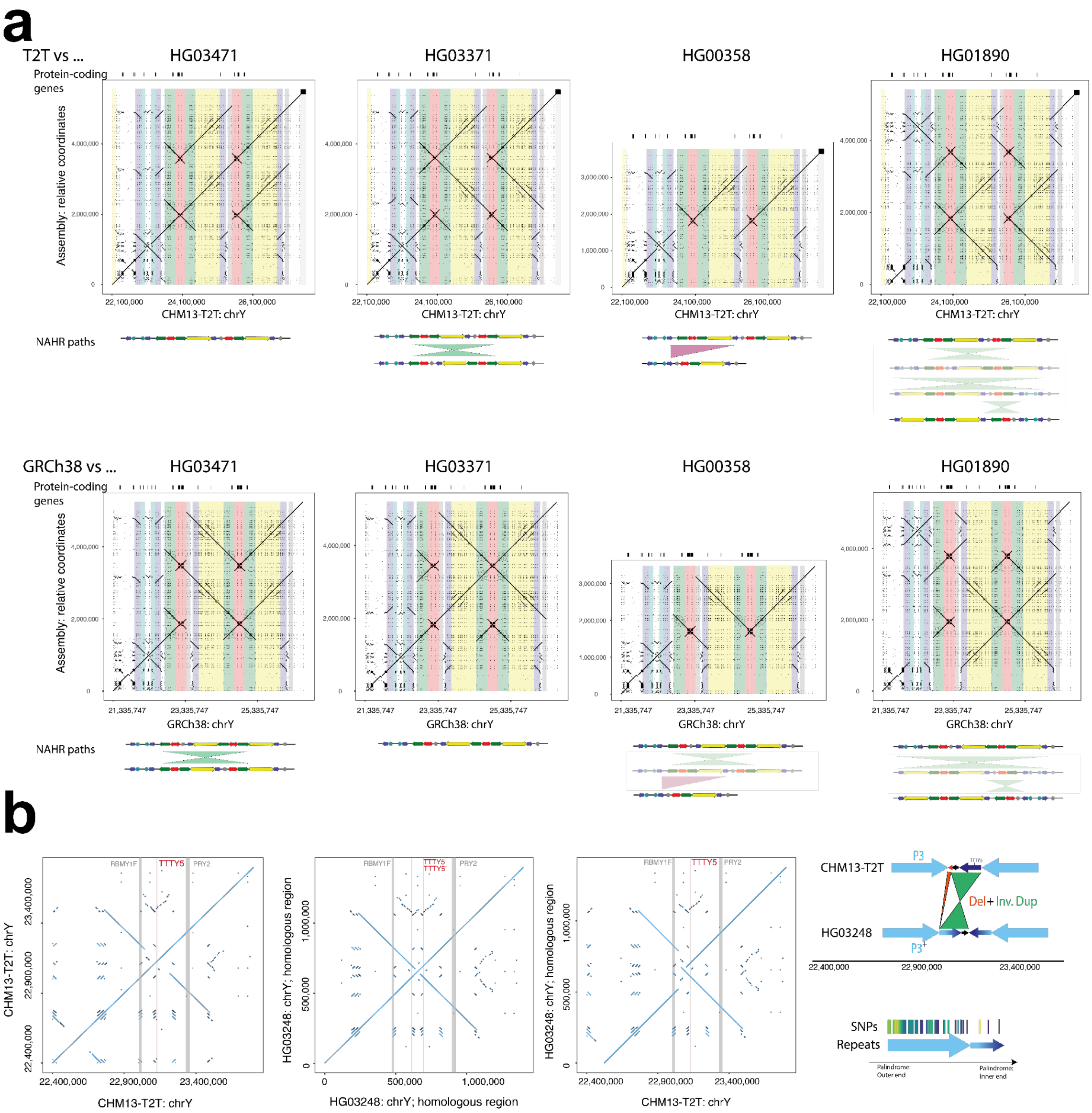


Figure S25. Examples of structural variation identified in the *de novo* assembled Y chromosomes. **a.** Inversions identified in the *AZFc*/ampliconic 7 subregion. Top - comparison between the T2T Y and the *de novo* assemblies, bottom - GRCh38 Y and the *de novo* assemblies. Potential NAHR path is shown below the dotplot. **b.** Inverted duplication affecting roughly two thirds of the 161 kbp unique 'spacer' sequence in the P3 palindrome, spawning a second copy of the TTTY5 gene and elongating the LCRs in this region. A detailed sequence view reveals a high sequence similarity between the duplication and its template and its placement in Y phylogeny supports emergence of this variant in the common ancestor of haplogroup E1a2 carried by NA19239, HG03248 and HG02572 (**Fig. 3a**).

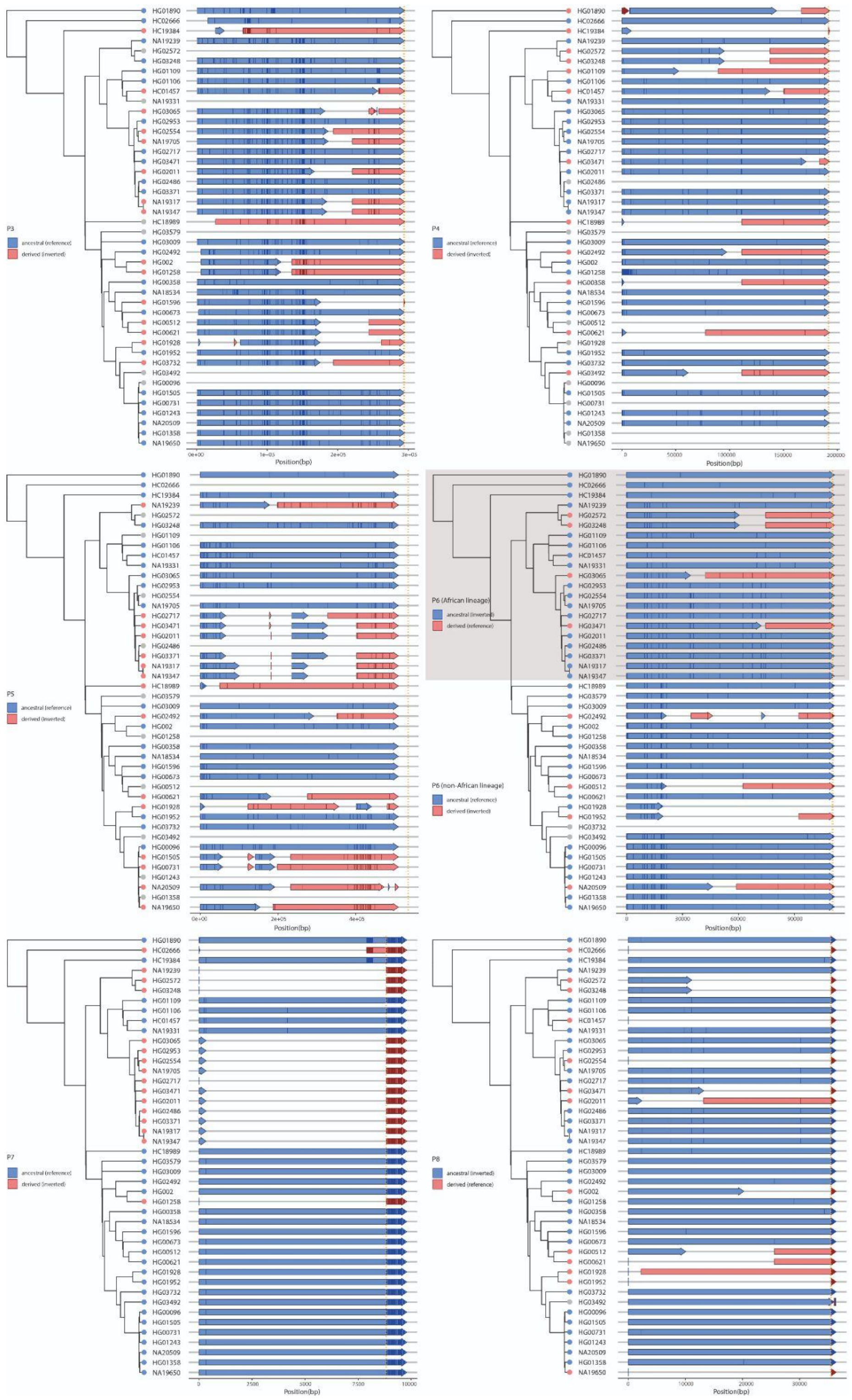


Figure S26. Breakpoint locations identified for 6 euchromatic inversions in palindromes P3, P4, P5, P6, P7 and P8. The red tip colors (derived state) in the phylogenetic tree indicate samples which have undergone an inversion and therefore carry the 'spacer' region in inverted orientation compared to samples with blue tip (ancestral state). 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. In P6, breakpoint locations were determined separately for African Y lineages (haplogroups A, B and E, gray shaded area) and non-African Y lineages, using two different sets of ancestral and derived states.

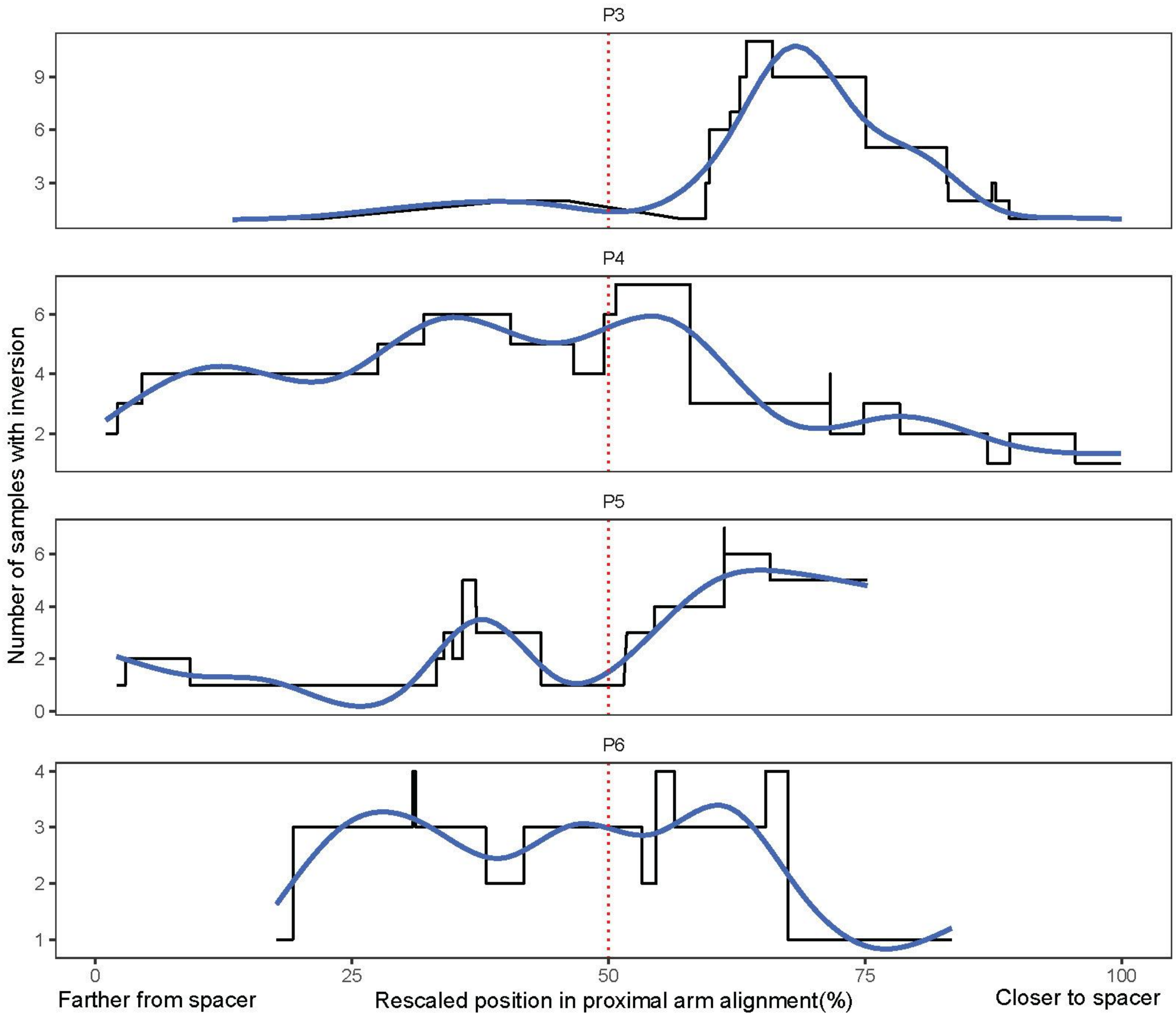
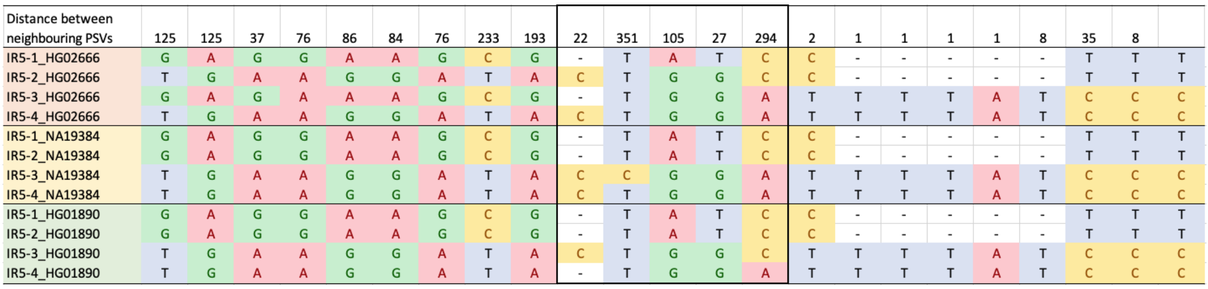


Figure S27. Rescaled breakpoint locations identified for 4 euchromatic inversions in palindromes P3, P4, P5, and P6. The start and end positions of each breakpoint range were rescaled to have the same start (0%) and end position (100%) across 4 palindromes. The y-axis indicates the number of samples that have inversion breakpoints at the corresponding position in the x-axis. The trend line indicated in blue is displayed by a smoothing function implemented in ggplot2 (geom_smooth, method =”gam”). P7 and P8 were excluded due to the small number of informative PSVs and therefore, wide breakpoint ranges.



Distance between neighbouring PSVs	125	125	37	76	86	84	76	233	193	22	351	105	27	294	2	1	1	1	1	8	35	8	
IR5-1_HG02666	G	A	G	G	A	A	G	C	G	-	T	A	T	C	C	-	-	-	-	-	T	T	T
IR5-2_HG02666	T	G	A	A	G	G	A	T	A	C	T	G	G	C	C	-	-	-	-	-	T	T	T
IR5-3_HG02666	G	A	G	A	A	A	G	C	G	-	T	G	G	A	T	T	T	T	A	T	C	C	C
IR5-4_HG02666	T	G	A	A	G	G	A	T	A	C	T	G	G	A	T	T	T	T	A	T	C	C	C
IR5-1_NA19384	G	A	G	G	A	A	G	C	G	-	T	A	T	C	C	-	-	-	-	-	T	T	T
IR5-2_NA19384	G	A	G	G	A	A	G	C	G	-	T	A	T	C	C	-	-	-	-	-	T	T	T
IR5-3_NA19384	T	G	A	A	G	G	A	T	A	C	C	G	G	A	T	T	T	T	A	T	C	C	C
IR5-4_NA19384	T	G	A	A	G	G	A	T	A	C	T	G	G	A	T	T	T	T	A	T	C	C	C
IR5-1_HG01890	G	A	G	G	A	A	G	C	G	-	T	A	T	C	C	-	-	-	-	-	T	T	T
IR5-2_HG01890	G	A	G	G	A	A	G	C	G	-	T	A	T	C	C	-	-	-	-	-	T	T	T
IR5-3_HG01890	T	G	A	A	G	G	A	T	A	C	T	G	G	C	T	T	T	T	A	T	C	C	C
IR5-4_HG01890	T	G	A	A	G	G	A	T	A	-	T	G	G	A	T	T	T	T	A	T	C	C	C

Figure S28. Inversion breakpoint identification for the *IR5/IR5* inversion in HG02666. The alignment shows all 4 IR5 repeats from three samples (HG02666 - inverted, NA19384 and HG01890 - reference orientation), with only informative PSV positions and genotypes shown (i.e., sites identical between the IR5 repeats and across individuals have been removed for visualization purposes). In NA19384 and HG01890 the IR5 repeats located within the P5 palindrome (IR5-1 and IR5-2) show a distinct PSV pattern from the IR5 copies located within the P1 palindrome. HG02666 which carries an inversion, the change of this pattern indicates the location of the inversion breakpoints and is highlighted by a black box. Inversion breakpoints relative to GRCh38 are: chrY:18,036,429-18,036,932 and chrY:24,036,893-24,037,396 for proximal and distal breakpoints, respectively.

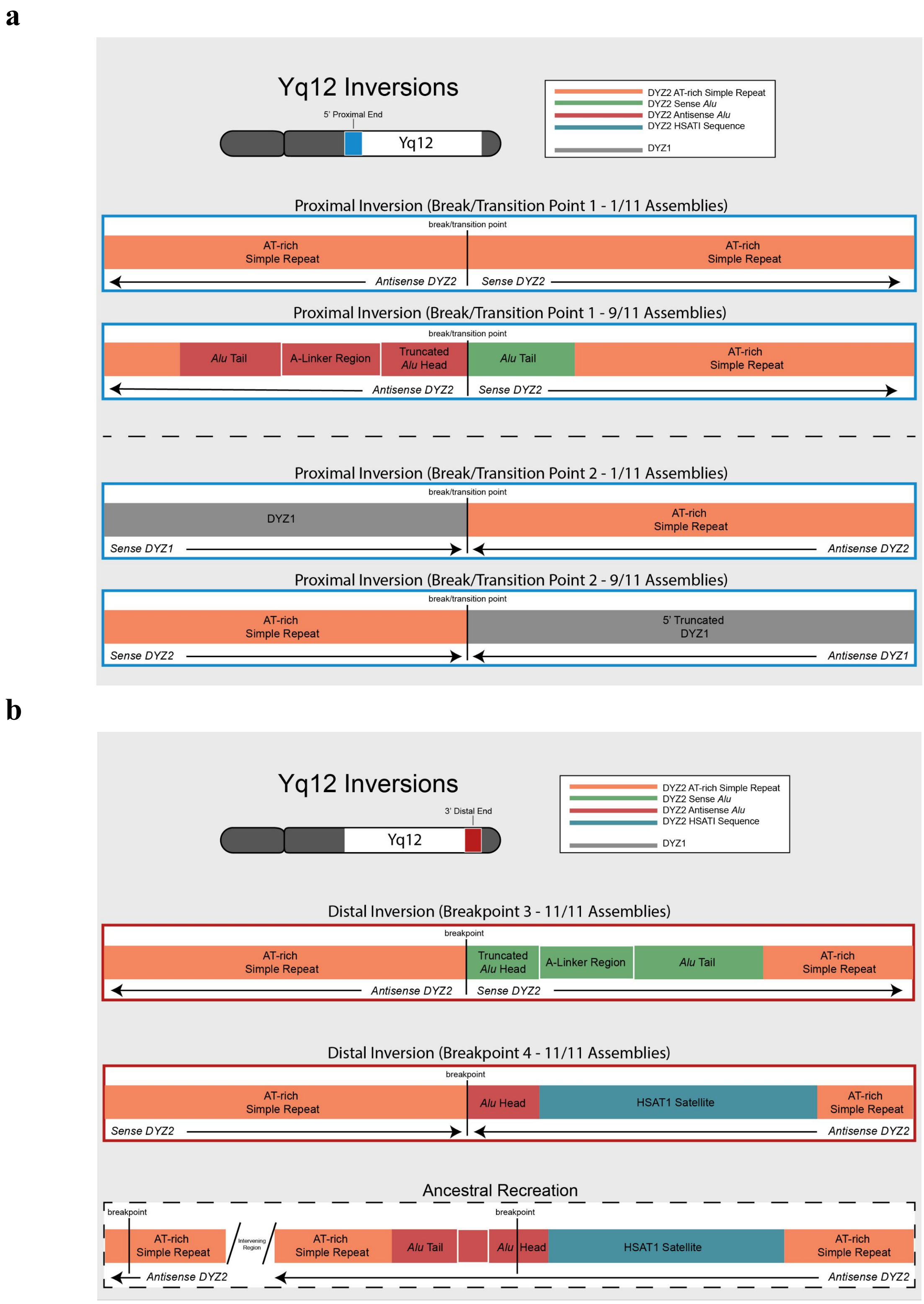


Figure S29. Inversions identified in the proximal and distal regions of the Yq12 subregion. **a.** Shows the proximal inversion break/transition regions. The top two rows show the inversion found in HG01890 and the bottom two rows the nine other genomes. The proximal inversion region is deleted in HG01106. The inversion break/transition points are described as such since the coordinates for where region changes orientation is located (see **Table S28**), but the exact breakpoint has not been elucidated. **b.** Shows the distal inversion breakpoints (top row) shared in all genomes, as well as an ancestral recreation of the region before the inversion.

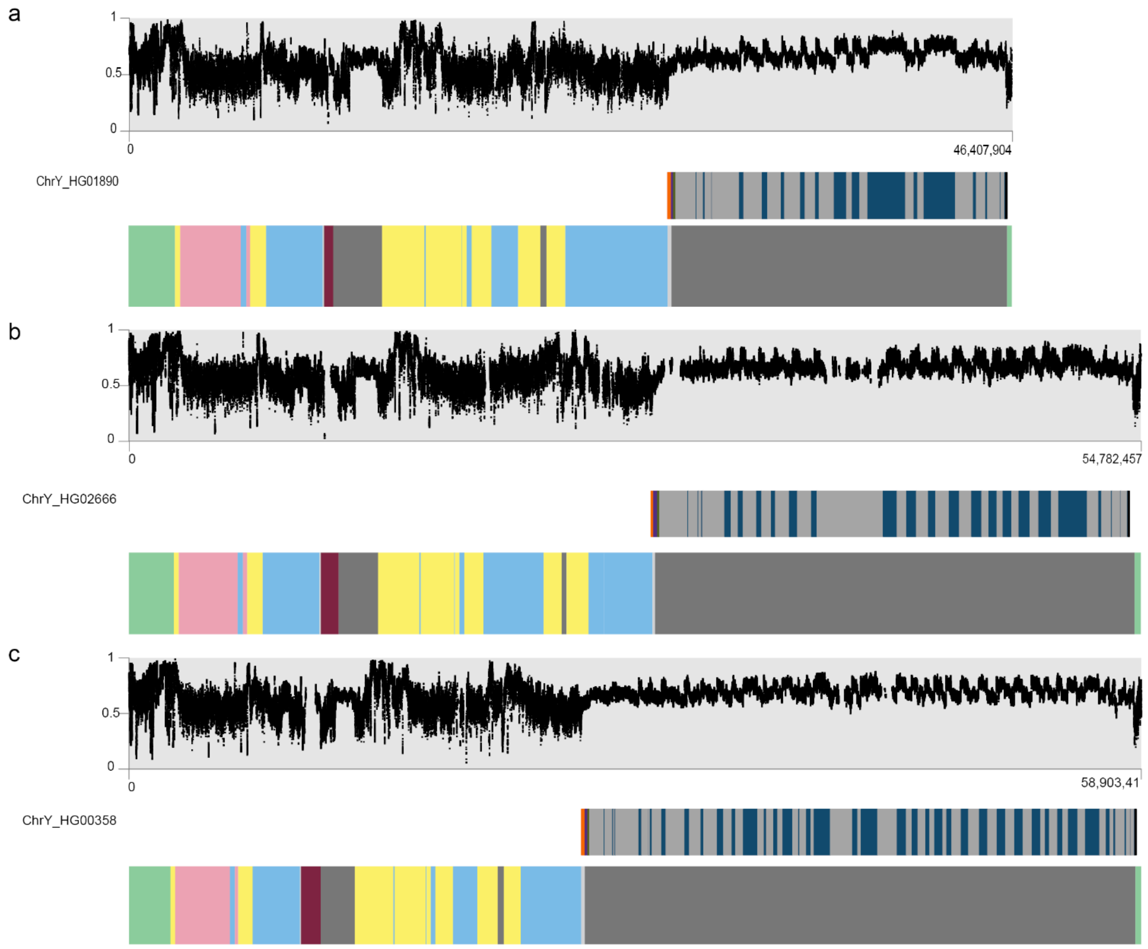


Figure S30. Methylation patterns as determined from the ONT data across the three contiguously assembled Y chromosomes, with repeat array locations (orange - *DYZ18*, purple - 2.7kb-repeat, green - 3.1kb-repeat, gray - *DYZ1*, blue - *DYZ2*) and Y-chromosomal subregion locations (see **Fig. 1a** for details) shown below as bar plots.

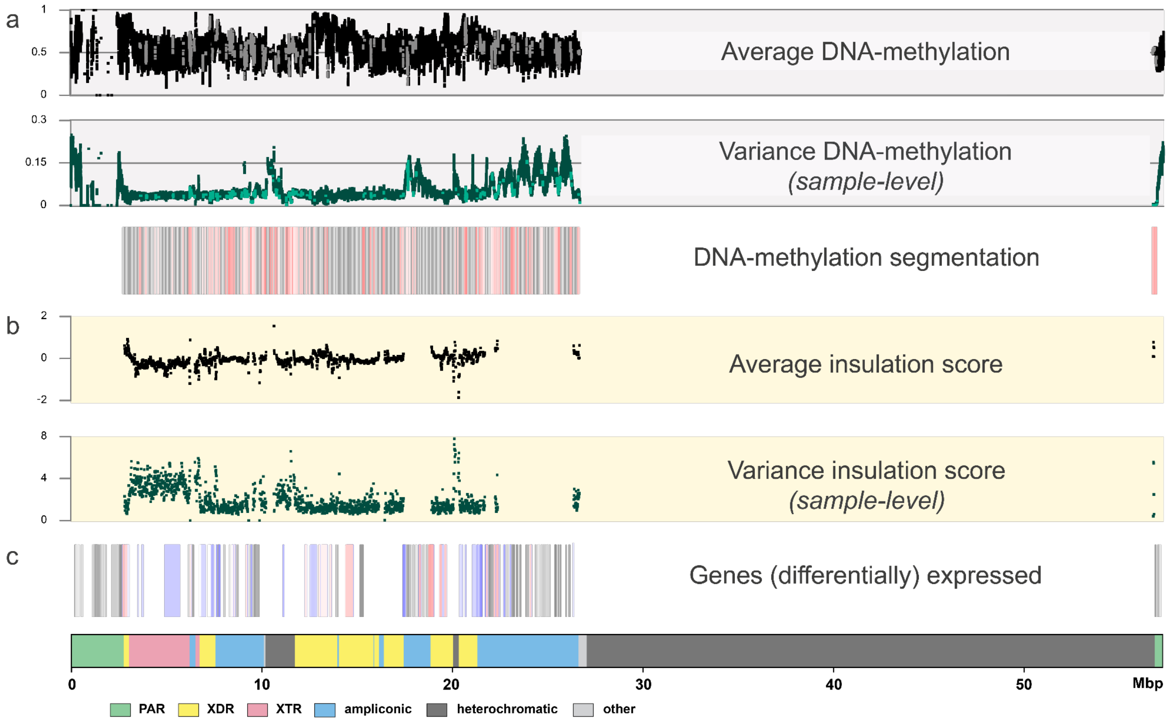


Figure S31. Functional analyses on the Y chromosome with DNA-methylation, RNA expression and HiC information as anchored to GRCh38. **a.** The top three panels show DNA-methylation levels and variation over the studied chromosomes. In black (top line) the average methylation is shown, in green (middle) the variation in DNA-methylation levels across the studied genomes. The bottom boxes represent the DNA methylation segmentation using PycoMeth-seg. In gray shades 2,861 methylation segments, and in red shades the 340 significantly differentially methylated segments (DMS). The CpG sites that fall in a DMS are colored in a lighter shade in the top two panels. **b.** Two panels showing average insulation scores (top) and variance of insulations scores between any two samples (bottom) across 40 samples with Hi-C data with 10 kbp resolution. Regions with lower insulation scores are more insulated and more likely to be TAD boundaries, while regions with higher scores are most likely to stay inside TADs. **c.** The Geuvadis based gene-expression analysis, shown are the 205 genes on Y chromosome (gray shades), the 64 genes expressed in the Geuvadis LCLs (blue shades), of which 22 are differentially expressed (red shades).

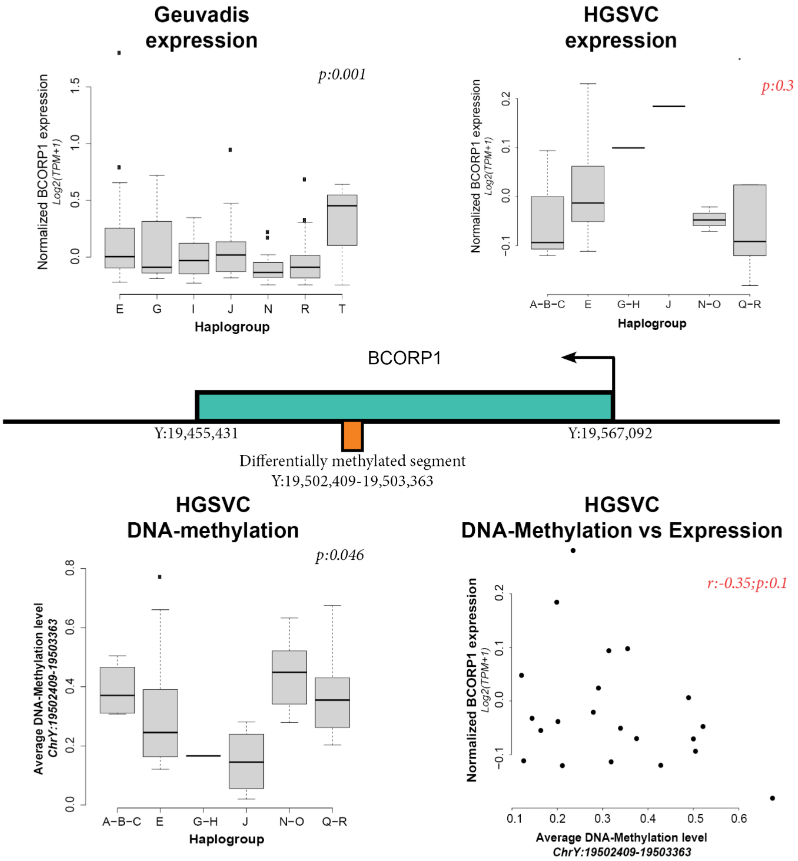


Figure S32. Schematic representation of the *BCORP1* gene and the effects of the haplogroup on gene expression and DNA-methylation (DNAme) levels. In the center an illustration of the *BCORP1* (in green), in the center of the gene a differentially DNAme segment is identified (in orange). The differential DNAme effect identified in the HGSVC samples is plotted in the bottom left boxplot. The *BCORP1* is found to be differentially expressed in the Geuvadis samples (top left boxplot). The expression effect is suggestive in the 21 HGSVC samples, expression of haplogroup E is on average higher than haplogroups G,H,J,N,O,Q,R. The expression effect of the haplogroup is inversely related to the DNA-methylation effect in this segment (Pearson R -0.35), with a suggestive P value of 0.1 indicating the relation in this small sample set.

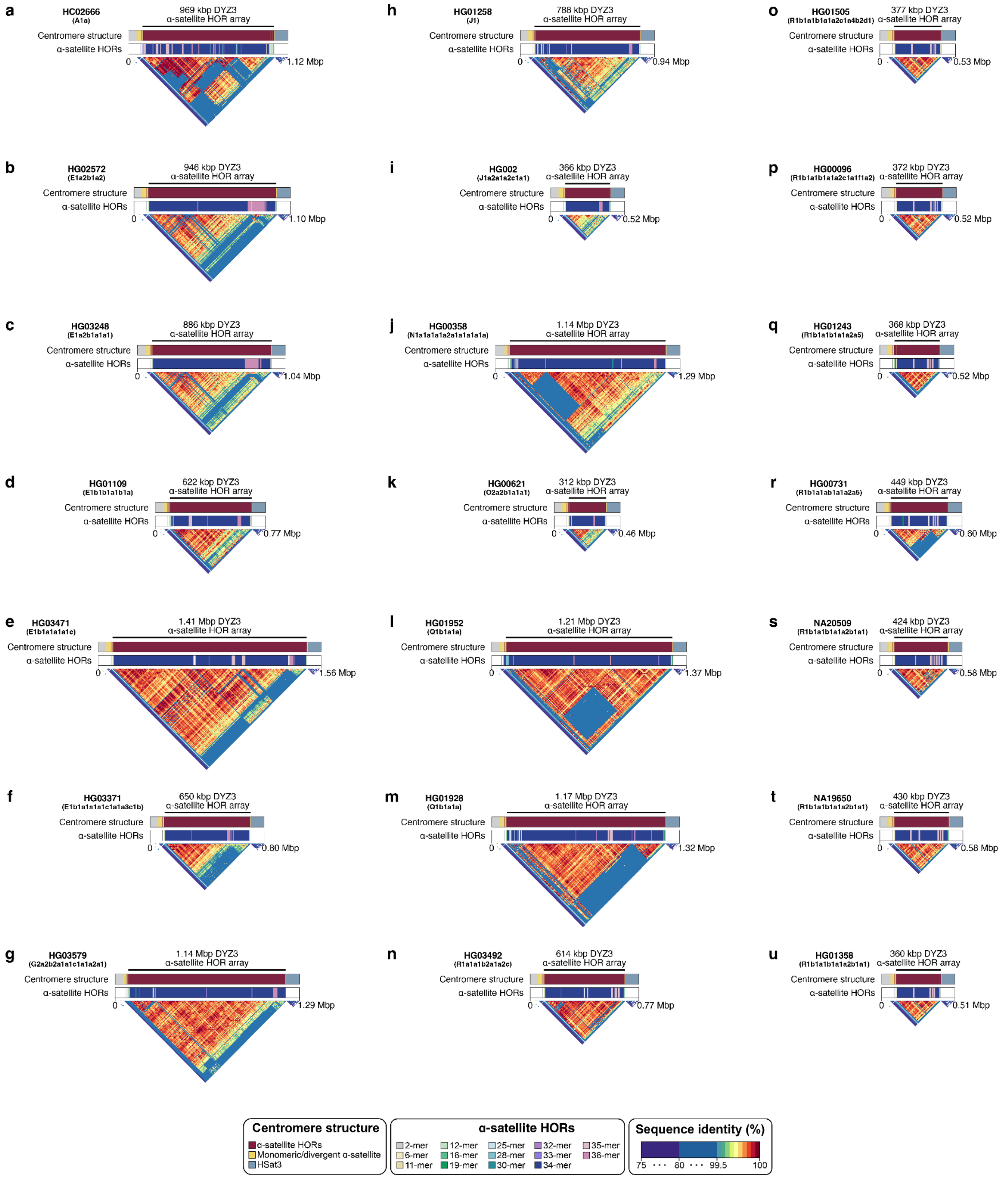


Figure S33. Organization of the chromosome Y centromeric region from 21 genomes representing all major superpopulations. The structure (top), α-satellite HOR organization (middle), and sequence identity heat map (bottom) for each centromere is shown and reveals the presence of novel HORs in over half of the centromeres. Note - the sizes of the *DYZ3* α-satellite array are shown on top as determined using RepeatMasker (**Methods**). The centromeres are ordered phylogenetically from the deepest-rooting sample, panel **a** to **u**.

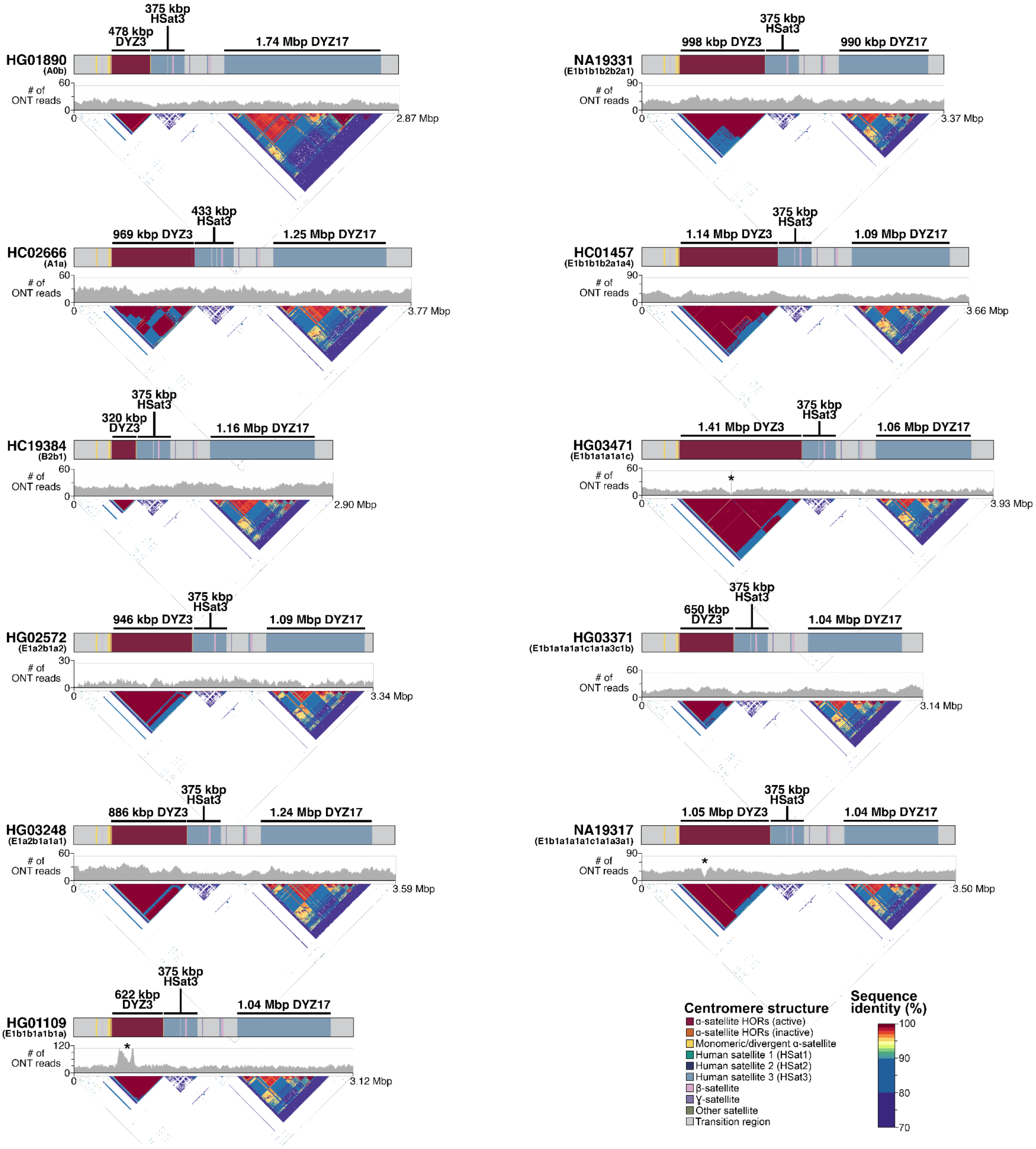


Figure S34. Genetic landscape of the Y-chromosomal pericentromeric region from samples carrying African Y lineages. The top panel shows locations and composition of the pericentromeric region with repeat array sizes shown for each Y chromosome (the *DYZ3* α-satellite array size as determined using RepeatMasker, **Methods**). The middle panel shows (UL-)ONT read depth and bottom sequence sequence identity head maps generated using StainedGlass pipeline (using 5-kb window size). Asterisks indicate two samples (HG01109 and NA19317) with a possible assembly collapse/error, and one sample (HG03471) with a single gap in the *DYZ3* array.

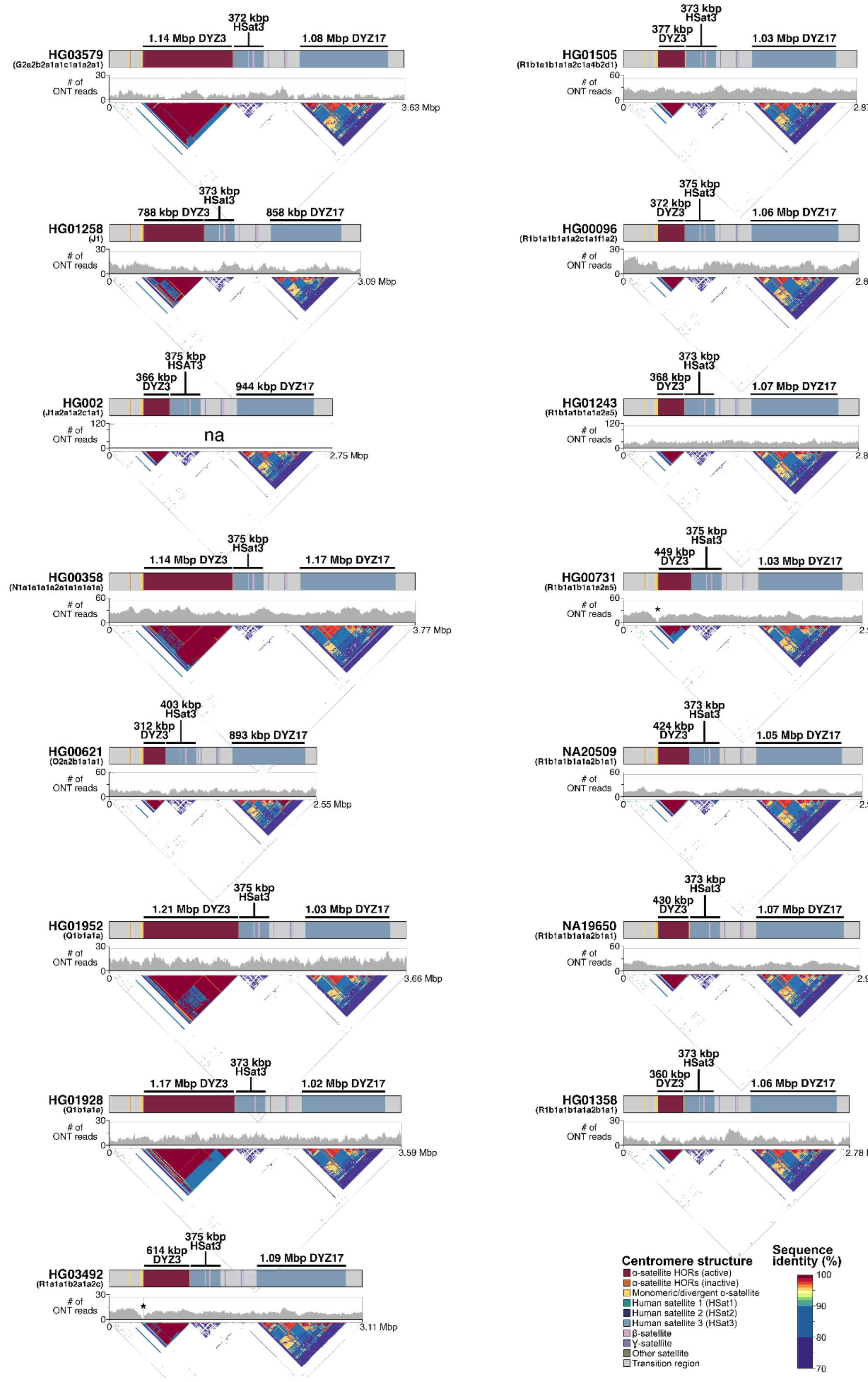


Figure S35. Genetic landscape of the Y-chromosomal pericentromeric region from samples carrying non-African Y lineages. The top panel shows locations and composition of the pericentromeric region with repeat array sizes shown for each Y chromosome (the *DYZ3* α-satellite array size as determined using RepeatMasker, **Methods**). The middle panel shows (UL-)ONT read depth and bottom sequence sequence identity head maps generated using StainedGlass pipeline (using 5-kb window size). Asterisks indicate two samples (HG03492 and HG00731) with a single gap in the *DYZ3* array. na - not available.

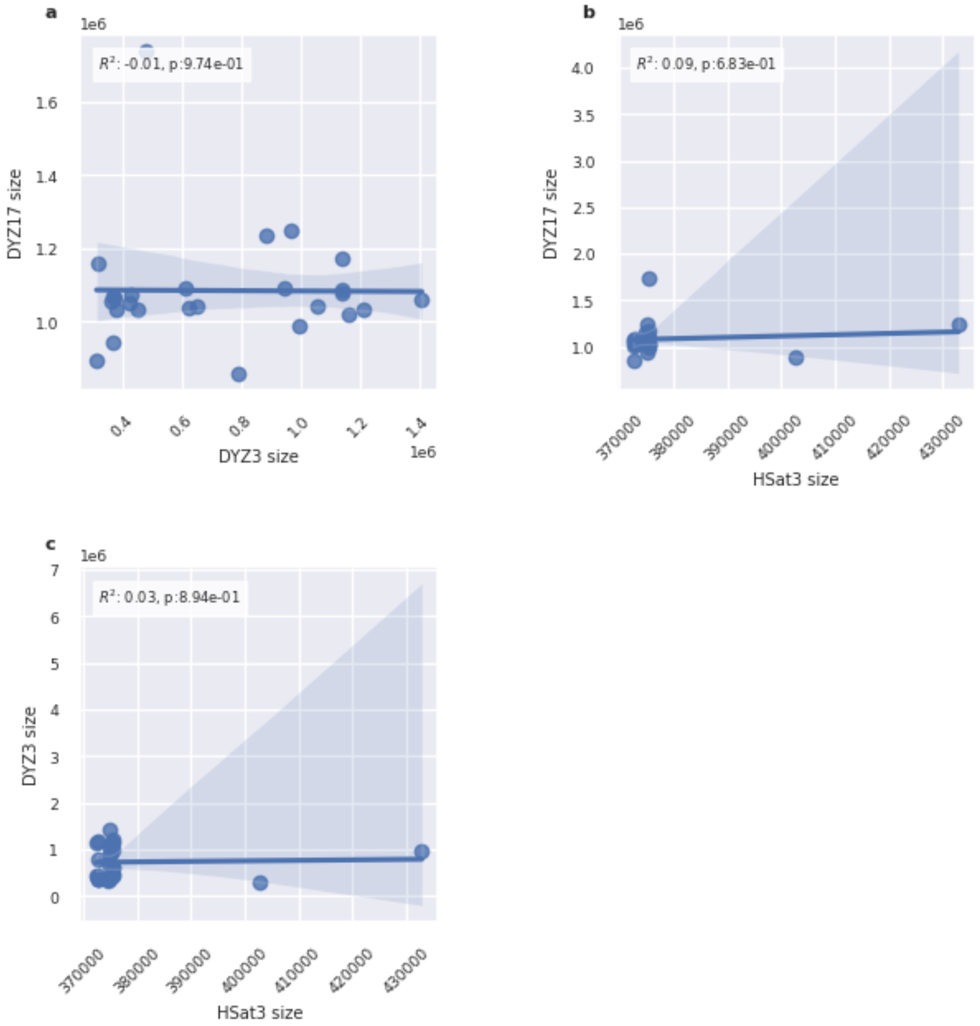


Figure S36. Regression plots between the sizes of (peri-)centromeric repeat arrays: *DYZ3* alpha-satellite array, Hsat3, and the *DYZ17* array. We report the correlation coefficient and the p-value on the upper-left corner box. No correlations attained a significant p-value.

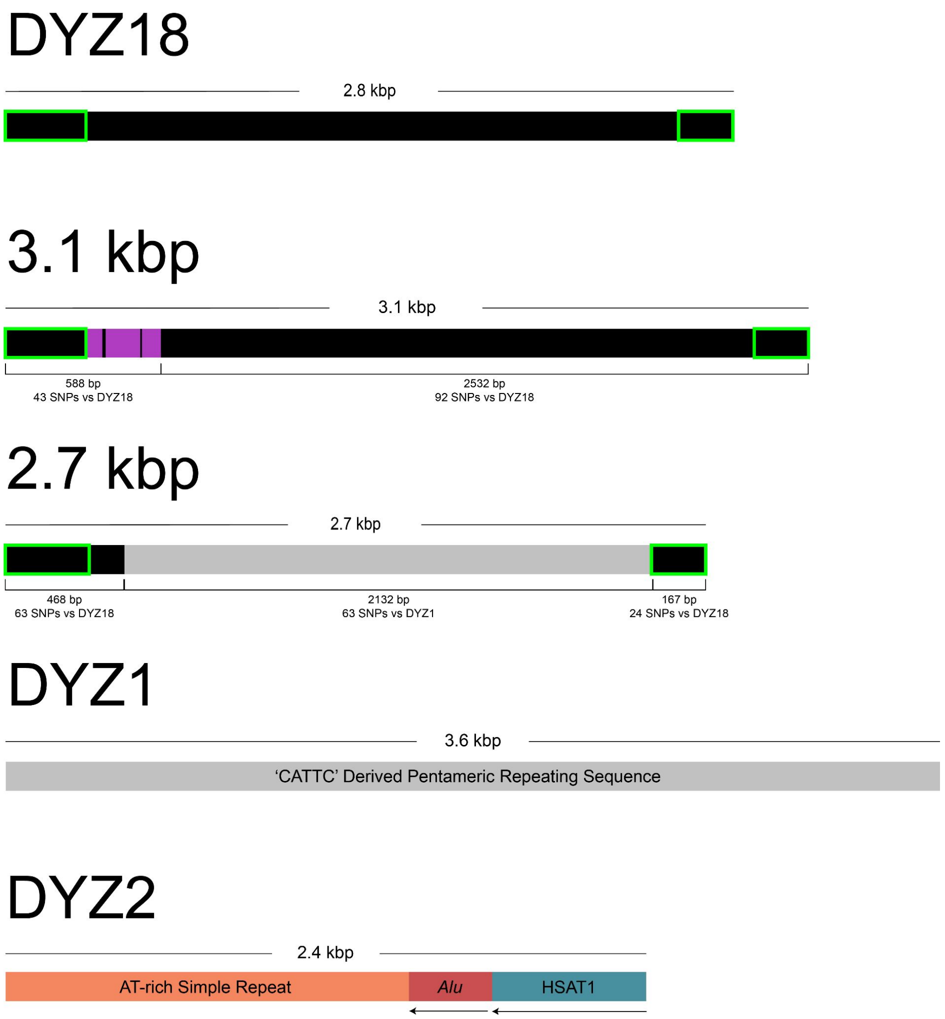


Figure S37. Composition and similarities of Yq12 heterochromatic repeat units. Green highlight - indicates regions with high sequence similarity to the *DYZ18* repeat unit. Light gray region in 2.7-kb repeat indicates a region of high sequence similarity to the *DYZ1* repeat unit. The purple region is a span of ~200-300 bases unique to the 3.1-kbp repeat.

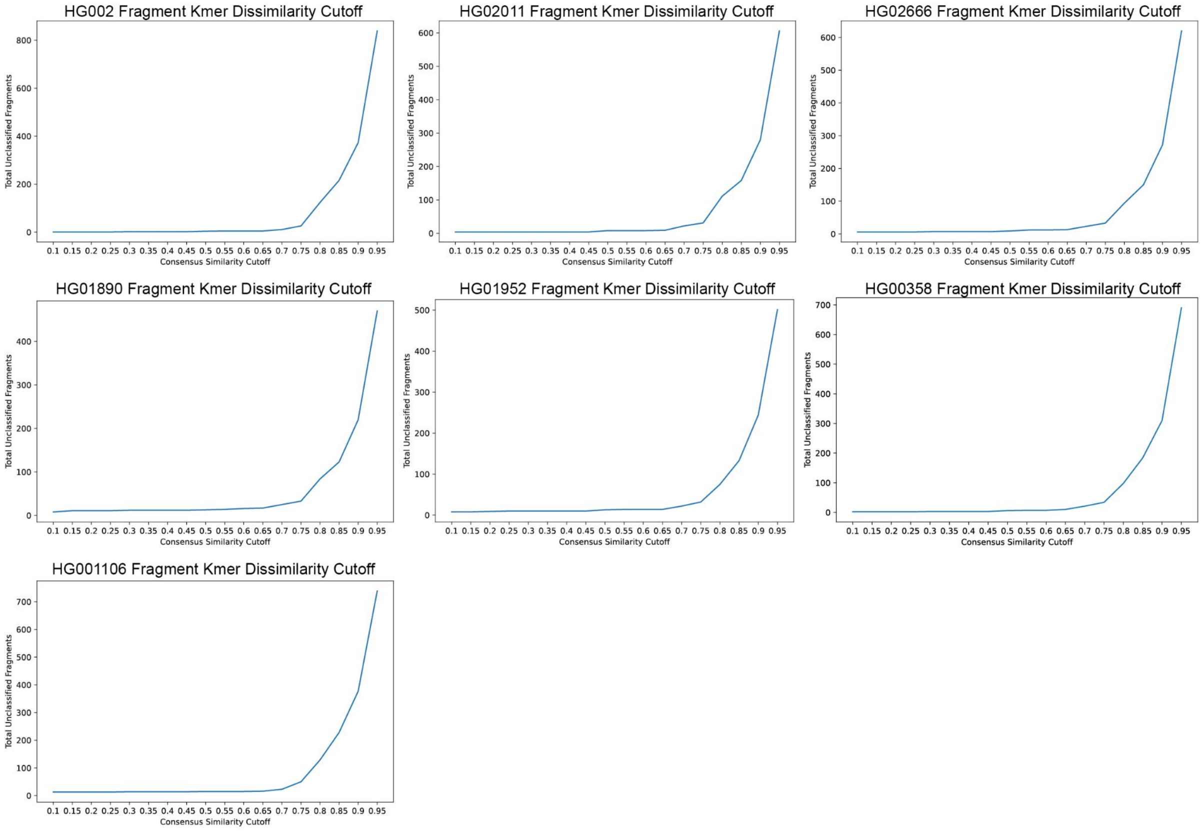


Figure S38. The line plots show the total HaeIII fragments (y-axis) that are unclassified at each k-mer abundance profile similarity cutoff (x-axis). Fragments were classified as either *DYZ18*, 3.1-kbp repeat, 2.7-kbp repeat, or *DYZ1* if their k-mer abundance profile was 75% or more similar. Each genome's plot exhibits an exponential growth in unclassified fragments above the 75% similarity cutoff.

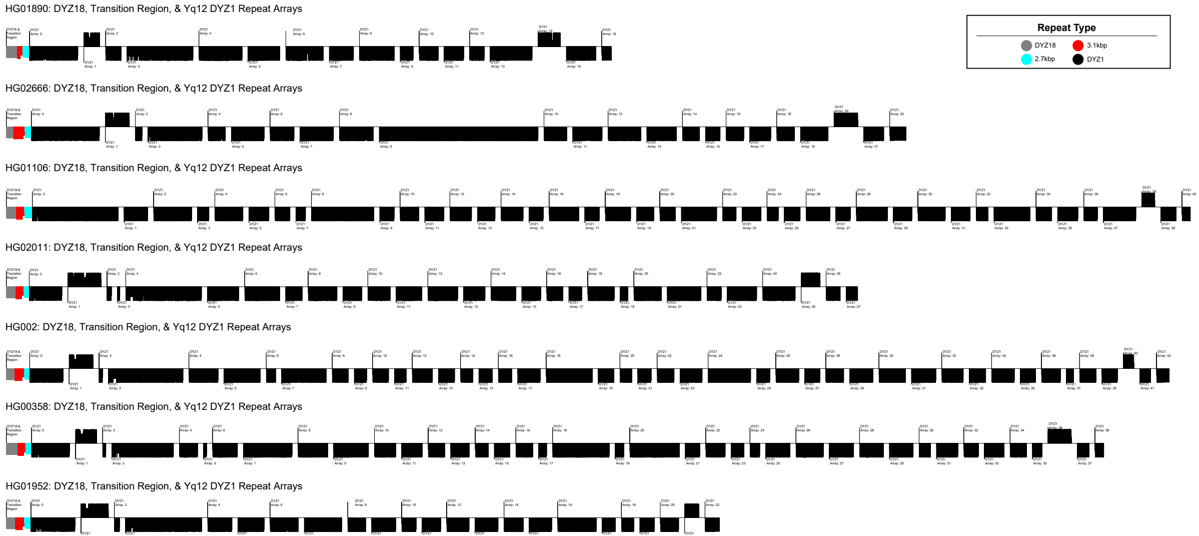

HG01890: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG02666: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG01106: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG02011: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG002: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG00358: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
HG01952: DYZ18, Transition Region, & Yq12 DYZ1 Repeat Arrays
Repeat Type
● DYZ18
● 3.1kbp
● 2.7kbp
● DYZ1

Figure S39. An overview of the *DYZ18* (gray), 3.1-kbp (red), 2.7-kbp (blue) and *DYZ1* (black) repeat arrays in the Yq11/transition region/Yq12 subregion within each of the seven samples with completely assembled Yq12 subregion. The length of individual lines is a function of the size of the repeat. The orientation (up = sense, down = antisense) was determined based on RepeatMasker annotations of satellite sequences within repeats.

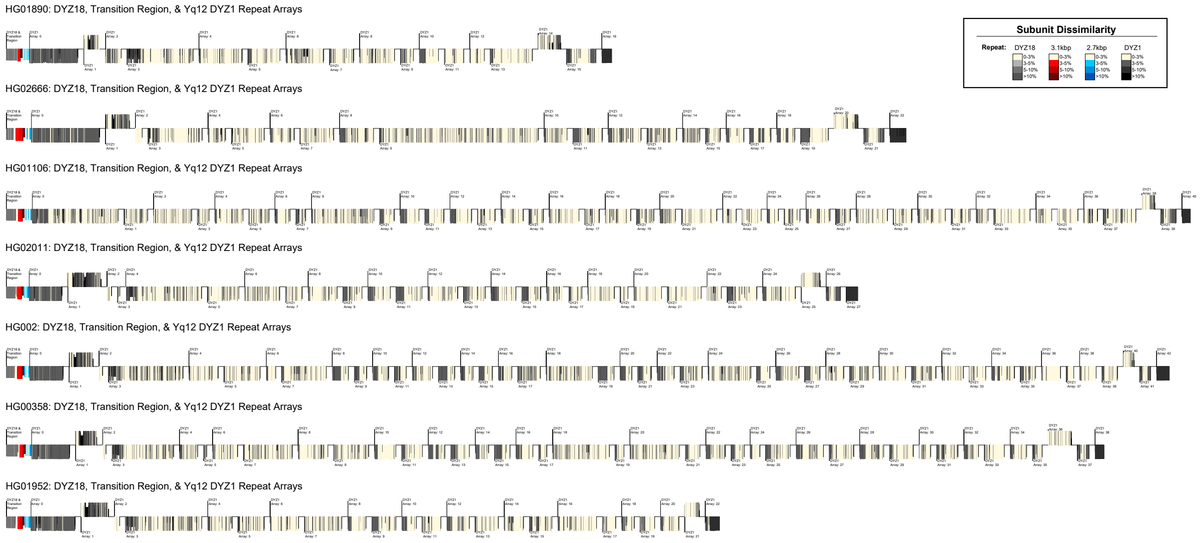


Figure S40. An overview of the Bray-Curtis distance/dissimilarity of k-mer abundance profiles for individual *DYZ18* (gray), 3.1-kbp (red), 2.7-kbp (blue) and *DYZ1* (black) repeats versus their consensus sequence. The Yq11/transition region/Yq12 subregion are shown for each of the seven samples with a completely assembled Yq12 subregion. Lighter colors indicate less distance/dissimilarity (more similar) k-mer abundance profiles compared to their consensus sequence. Results indicate arrays located on the proximal and distal boundaries of the Yq12 region contain repeats with k-mer abundance compositions less similar to their consensus sequence (i.e., more diverged). The length of individual lines is a function of the size of the repeat. The orientation (up = sense, down = antisense) was determined based on RepeatMasker annotations of satellite sequences within repeats.

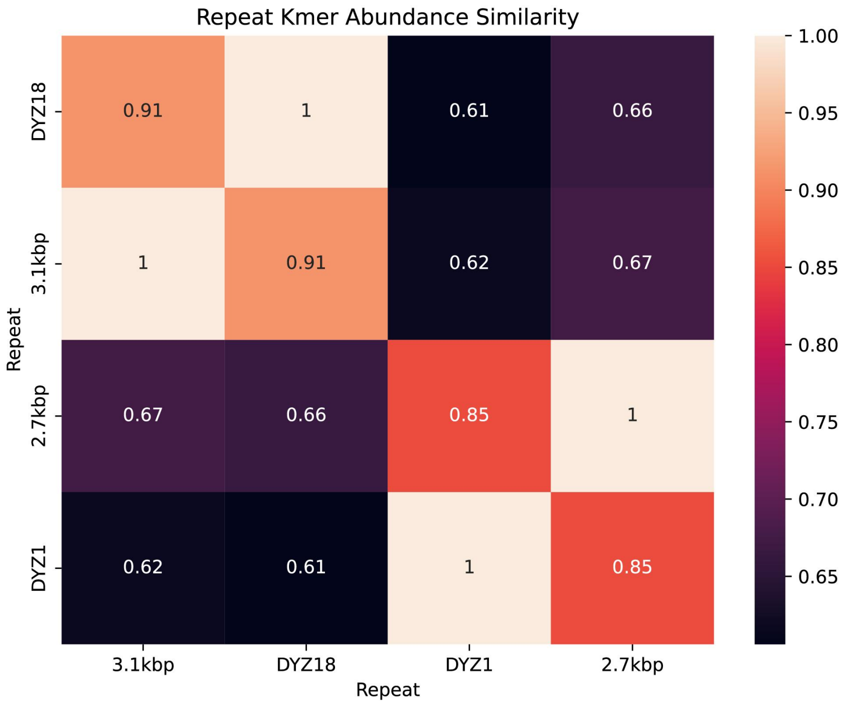


Figure S41. Heatmap of the complement of the Bray-Curtis distance/dissimilarity (i.e., 1-BC) between k-mer abundance profiles of *DYZ18*, 3.1-kbp, 2.7-kbp, and *DYZ1* consensus sequences is shown. The k-mer abundance profile of *DYZ1* was most similar to the 2.7-kbp repeat (85%), whereas the *DYZ18* and 3.1-kbp repeat sequences were more similar (91%) to each other.

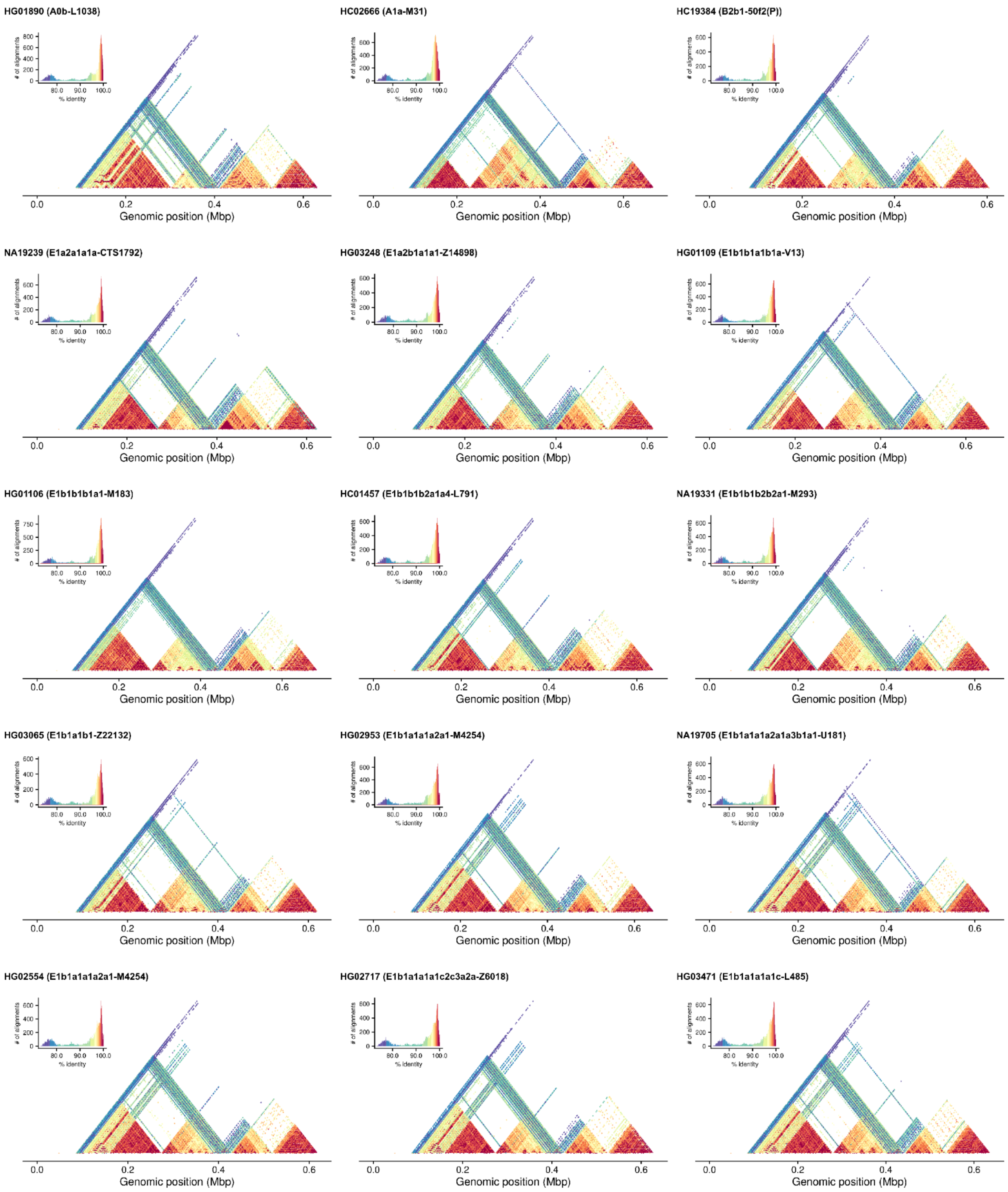


Figure S42. Sequence identity heat map of the Yq11/Yq12 transition region, including the *DYZ18*, 3.1-kbp, 2.7-kbp repeat arrays and 100 kbp of the first *DYZ1* repeat array generated using StainedGlass with 2 kbp window size. 100 kbp proximal to the *DYZ18* repeat array has also been included. Samples are ordered phylogenetically from the deepest-rooting sample (from left to right). The plot highlights higher sequence similarity between the *DYZ18* and 3.1-kbp repeat arrays, and between the 2.7-kbp and *DYZ1* repeat arrays, respectively.

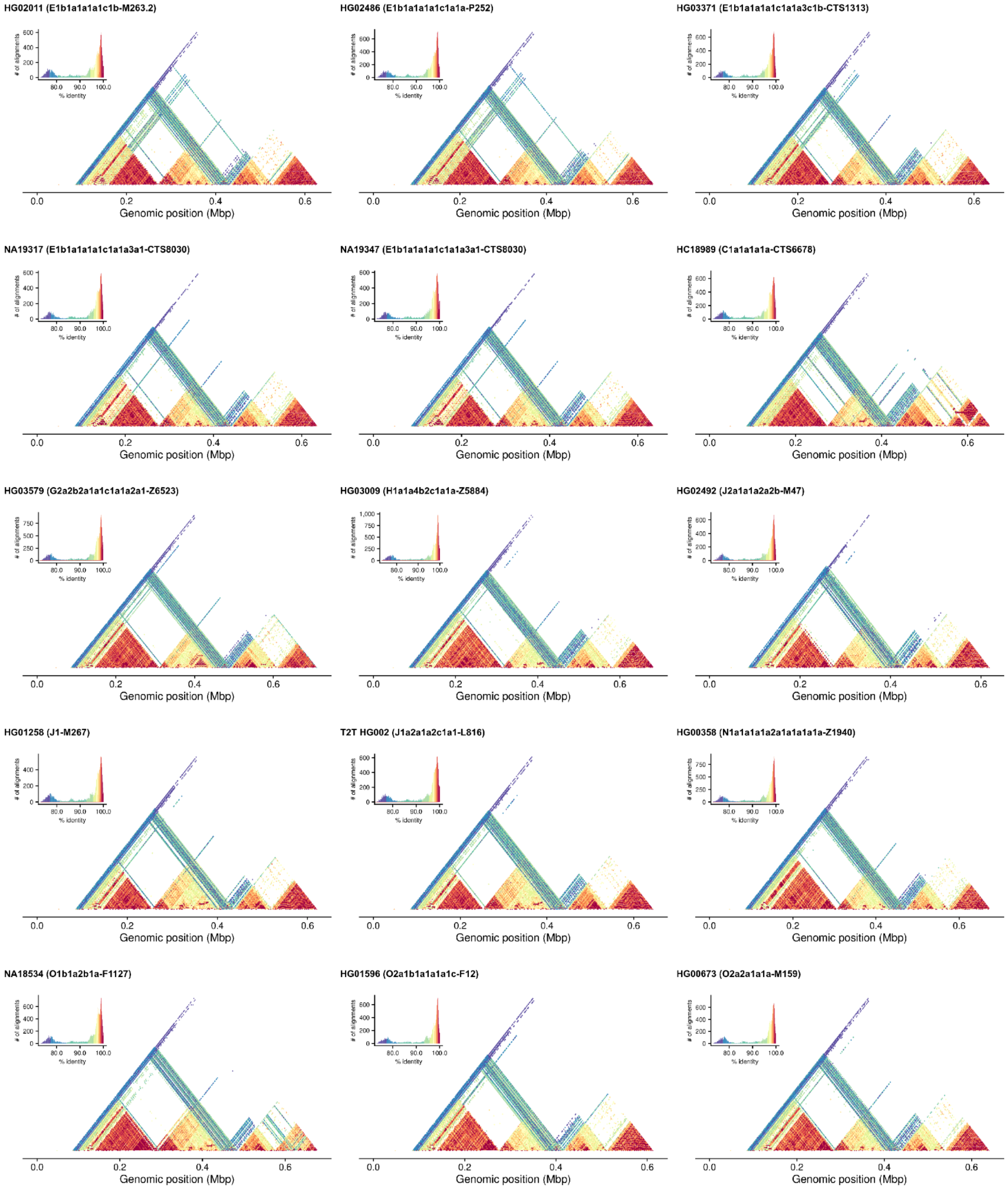
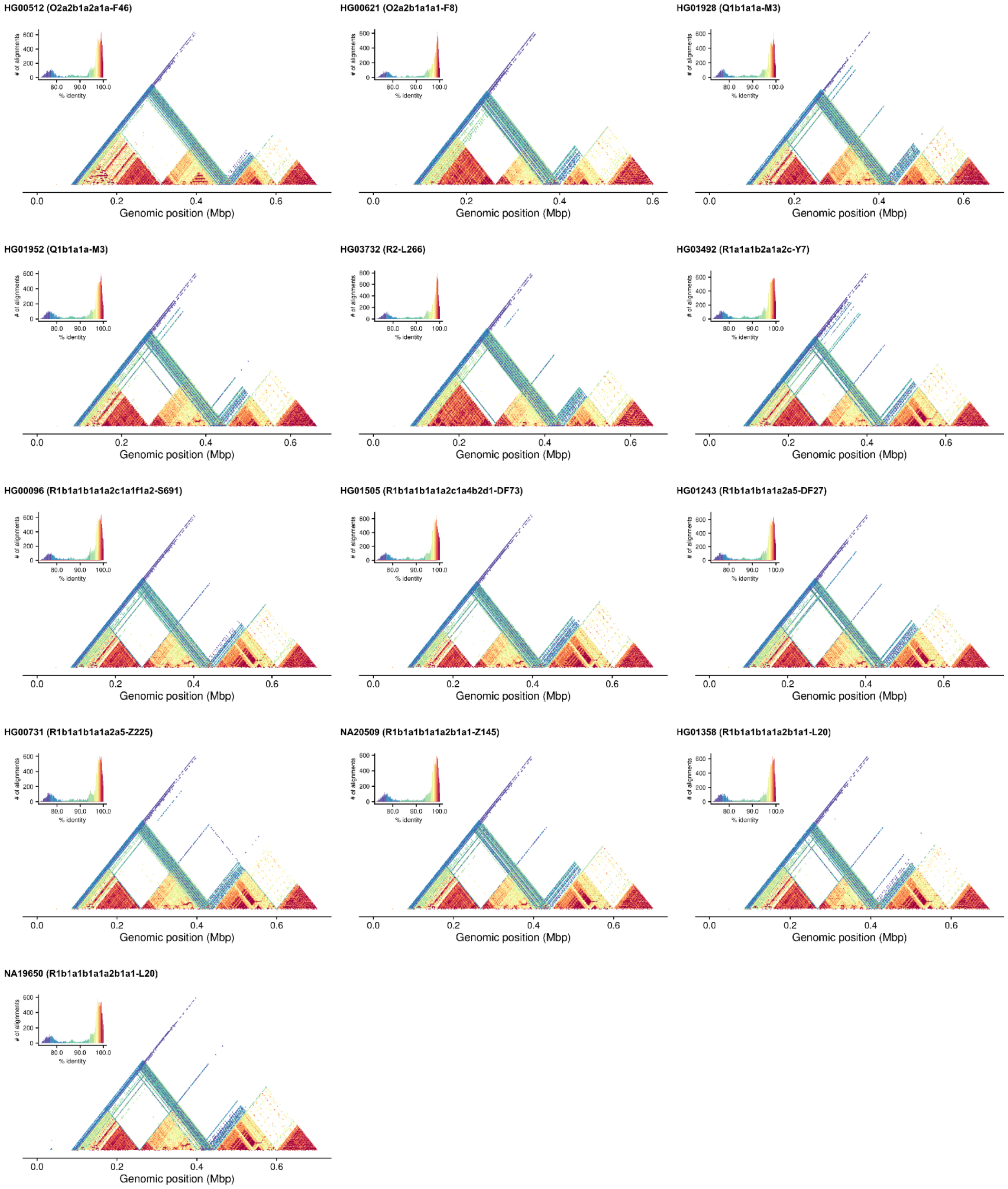


Figure S43. Sequence identity heat map of the Yq11/Yq12 transition region, including the *DYZ18*, 3.1-kbp, 2.7-kbp repeat arrays and 100 kbp of the first *DYZ1* repeat array generated using StainedGlass with 2 kbp window size. 100 kbp proximal to the *DYZ18* repeat array has also been included. Samples are ordered phylogenetically from the deepest-rooting sample (from left to right). The plot highlights higher sequence similarity between the *DYZ18* and 3.1-kbp repeat arrays, and between the 2.7-kbp and *DYZ1* repeat arrays, respectively.



HG00512 (O2a2b1a2a1a-F46)

HG00621 (O2a2b1a1a1-F8)

HG01928 (Q1b1a1a-M3)

HG01952 (Q1b1a1a-M3)

HG03732 (R2-L266)

HG03492 (R1a1a1b2a1a2c-Y7)

HG00096 (R1b1a1b1a1a2c1a1f1a2-S691)

HG01505 (R1b1a1b1a1a2c1a4b2d1-DF73)

HG01243 (R1b1a1b1a1a2a5-DF27)

HG00731 (R1b1a1b1a1a2a5-Z225)

NA20509 (R1b1a1b1a1a2b1a1-Z145)

HG01358 (R1b1a1b1a1a2b1a1-L20)

NA19650 (R1b1a1b1a1a2b1a1-L20)

Figure S44. Sequence identity heat map of the Yq11/Yq12 transition region, including the *DYZ18*, 3.1-kbp, 2.7-kbp repeat arrays and 100 kbp of the first *DYZ1* repeat array generated using StainedGlass with 2 kbp window size. 100 kbp proximal to the *DYZ18* repeat array has also been included. Samples are ordered phylogenetically from the deepest-rooting sample (from left to right). The plot highlights higher sequence similarity between the *DYZ18* and 3.1-kbp repeat arrays, and between the 2.7-kbp and *DYZ1* repeat arrays, respectively.

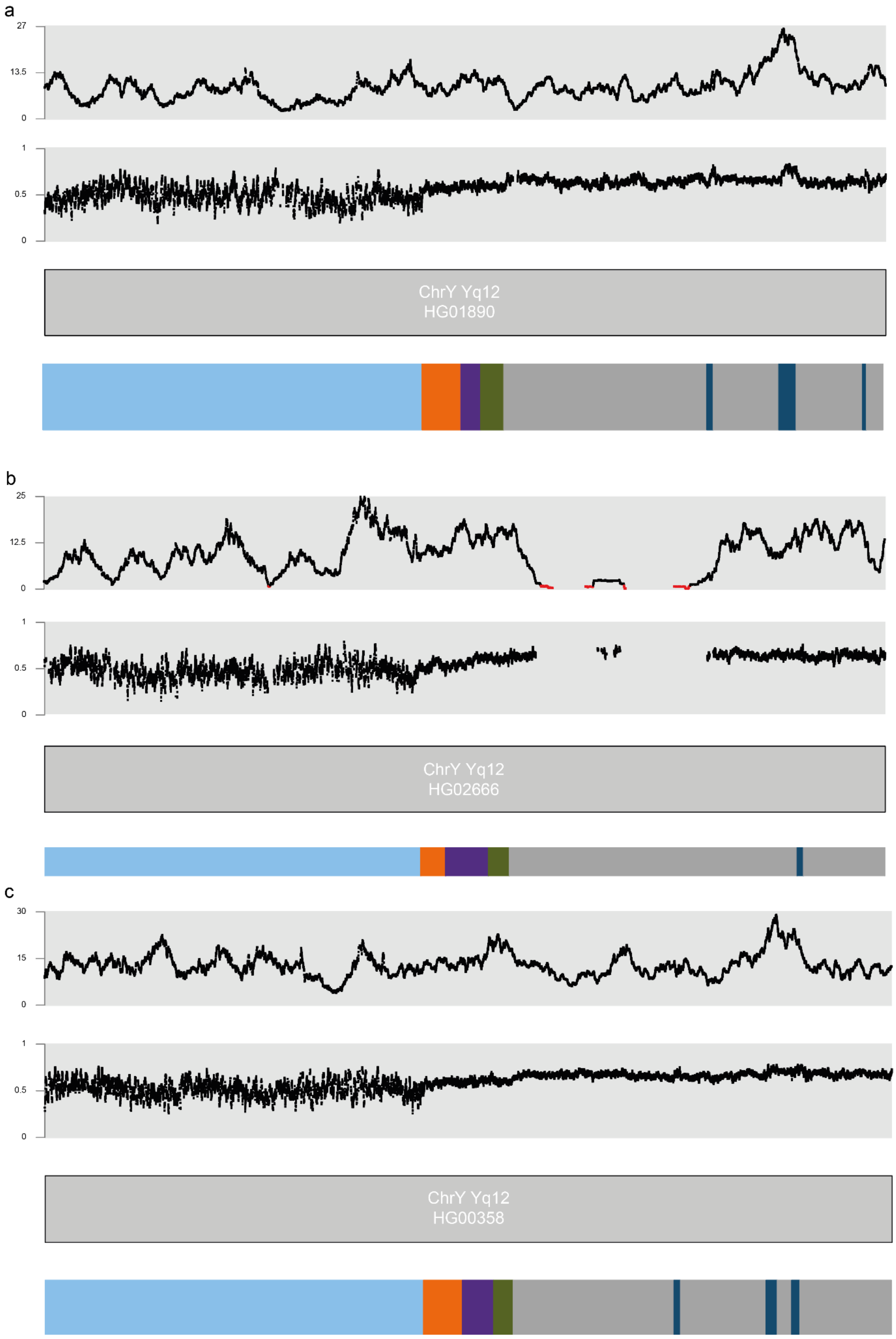


Figure S45. ONT read depth (top) and methylation patterns (below) around the boundary of Yq11 euchromatin and the Yq12 heterochromatic subregion across the three contiguously assembled Y chromosomes, with the sequence annotations shown below (light blue - ampliconic 7 subregion, orange - *DYZ18*, purple - 2.7-kb repeat, green - 3.1-kb repeat, gray - *DYZ1*, dark blue - *DYZ2*).

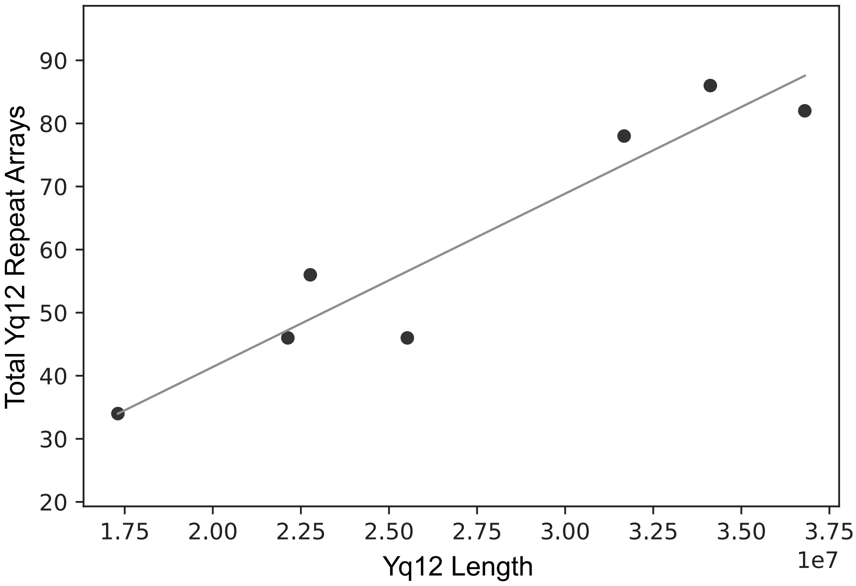


Figure S46. A scatter plot showing the total number of *DYZ1* and *DYZ2* arrays within the Yq12 subregions of each sample (n=7, samples with complete assembly plus the T2T Y) (y-axis) versus the total length of the Yq12 region (x-axis) is illustrated. This relationship was found to be significantly positively correlated (two-sided Spearman correlation=0.90; p-value < 0.05).

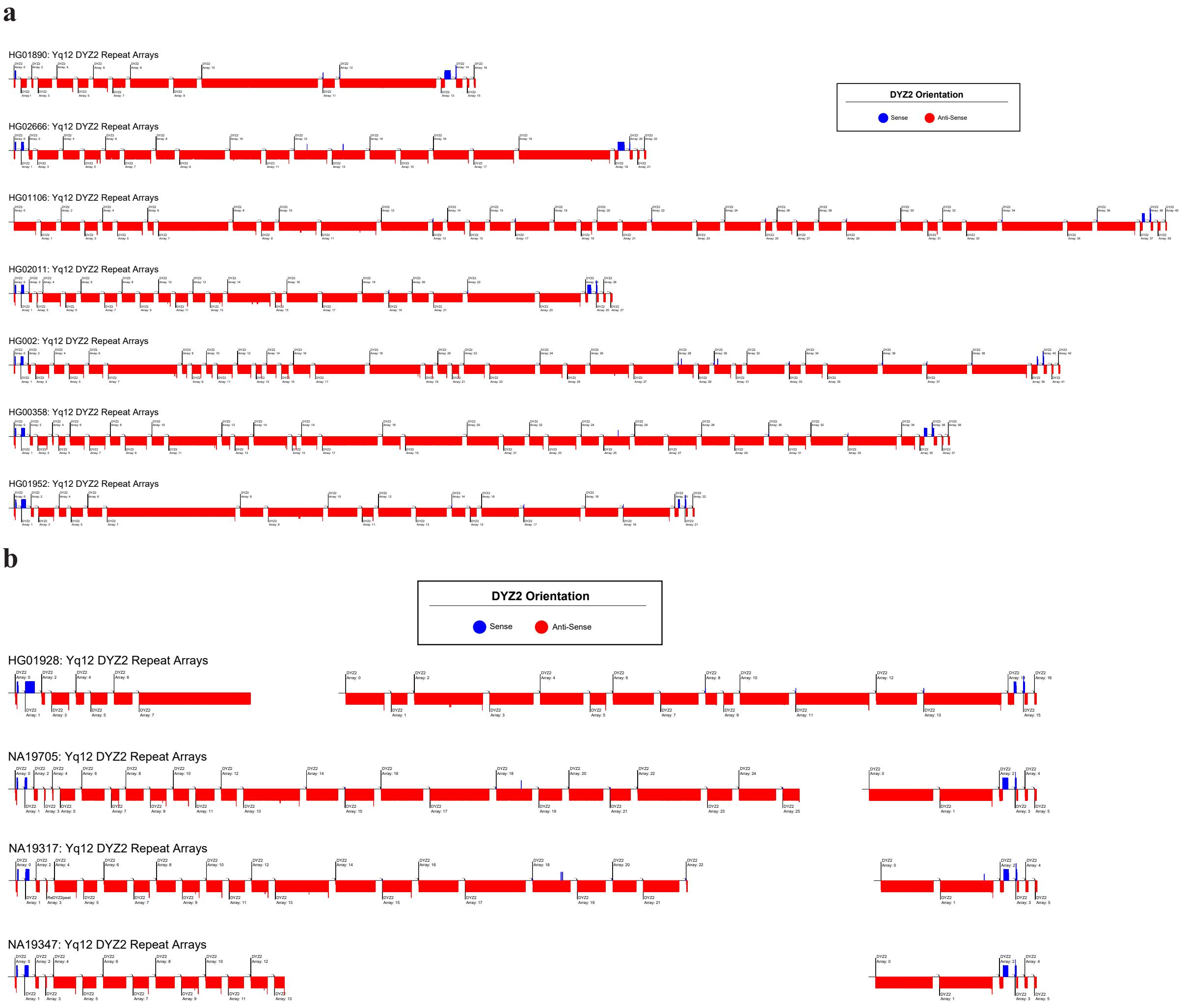


Figure S47. Overview of the *DYZ2* repeat array orientation and structure within the Yq12 subregion of each **a.** sample with completely assembled Yq12 subregion, and **b.** the four additional genomes (HG01928, NA19705, NA19317, NA19347) with incompletely assembled Yq12 regions. Red lines indicate individual *DYZ2* repeats in antisense orientation, blue lines indicate individual *DYZ2* repeats in sense orientation relative to the *DYZ2* consensus sequence. The length of each line is a function of the length of the repeat.

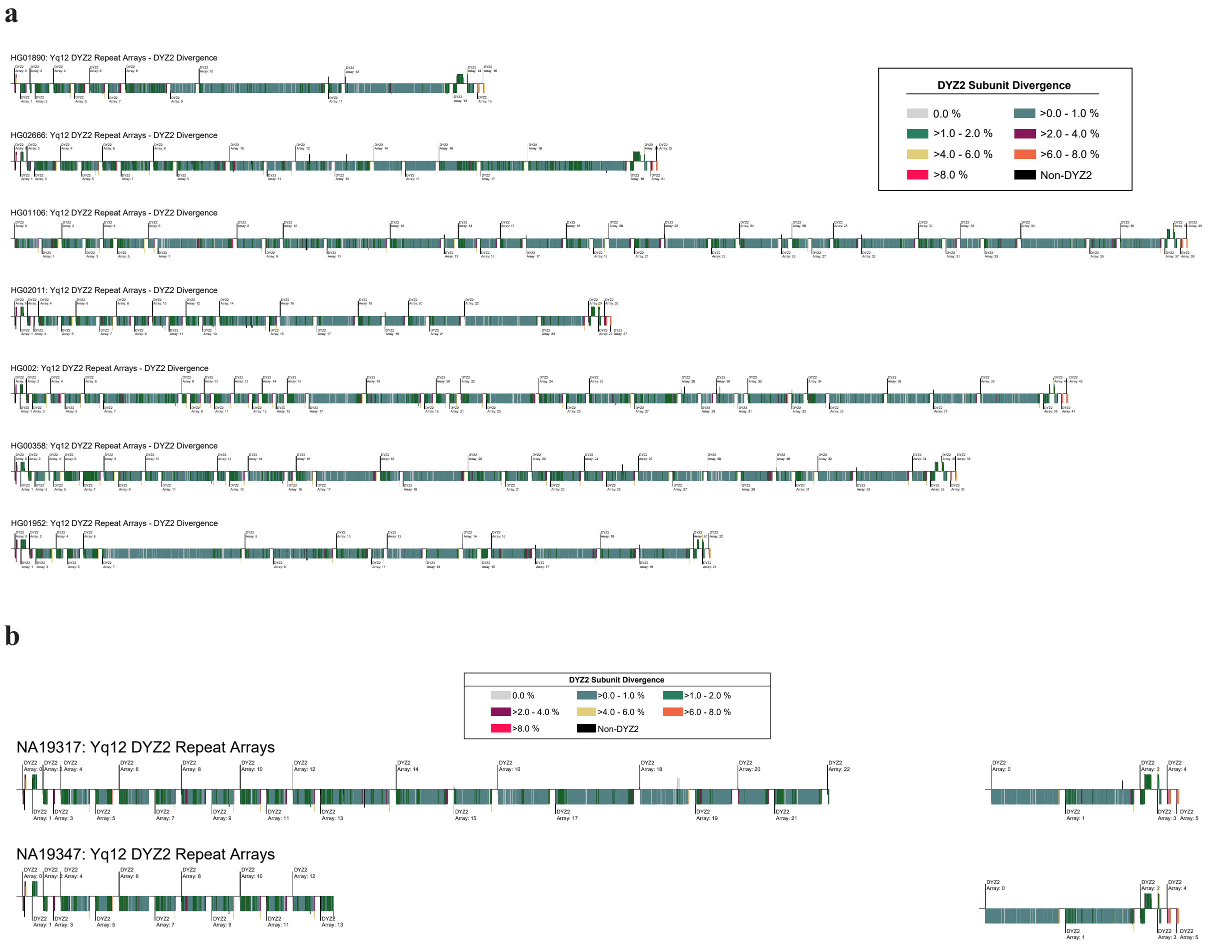


Figure S48. An overview of the divergence of individual *DYZ2* subunits for **a.** samples with completely assembled Yq12 subregion (HG01890, HG02666, HG01106, HG02011, T2T Y, HG00358, HG01952), and **b.** the two most closely related genomes (NA19317 and NA19347) with incompletely assembled Yq12 sunregions. A higher divergence was observed within the subunits located in arrays at the proximal and distal ends of the Yq12 subregion. Additionally, *DYZ2* subunits located near the boundaries of individual arrays tend to be more diverged than those located centrally. Between the closely related genomes, the divergence of *DYZ2* repeats within the shared *DYZ2* arrays are extremely similar.

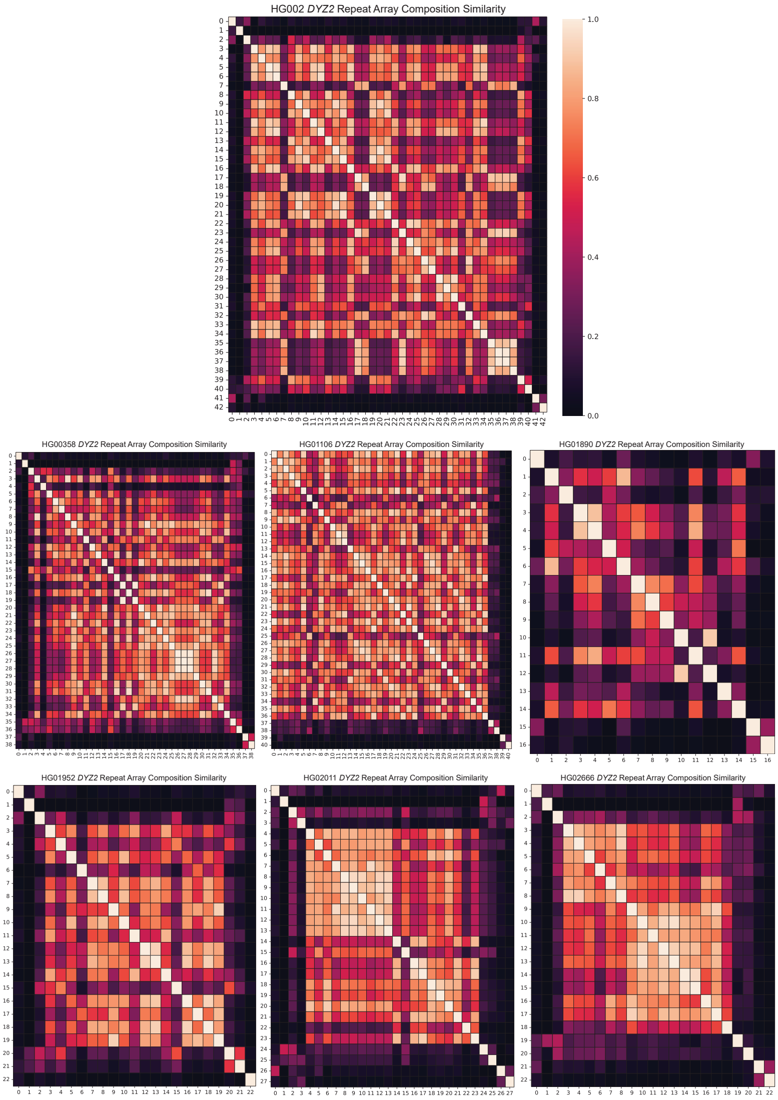


Figure S49. Heatmaps showing the complement of the Bray-Curtis (BC) distance/dissimilarity (i.e., 1-BC) for *DYZ2* repeat arrays within each genome with a completely assembled Yq12 subregion. Higher values (i.e., 1.00) indicate *DYZ2* arrays that contain exactly the same subunit composition whereas lower values (i.e., 0.0) suggest the opposite. Results show that the composition of arrays closer to one another tend to be more similar, except for the arrays located in the proximal and distal inversions, which tend to be more similar to each other than to surrounding arrays.

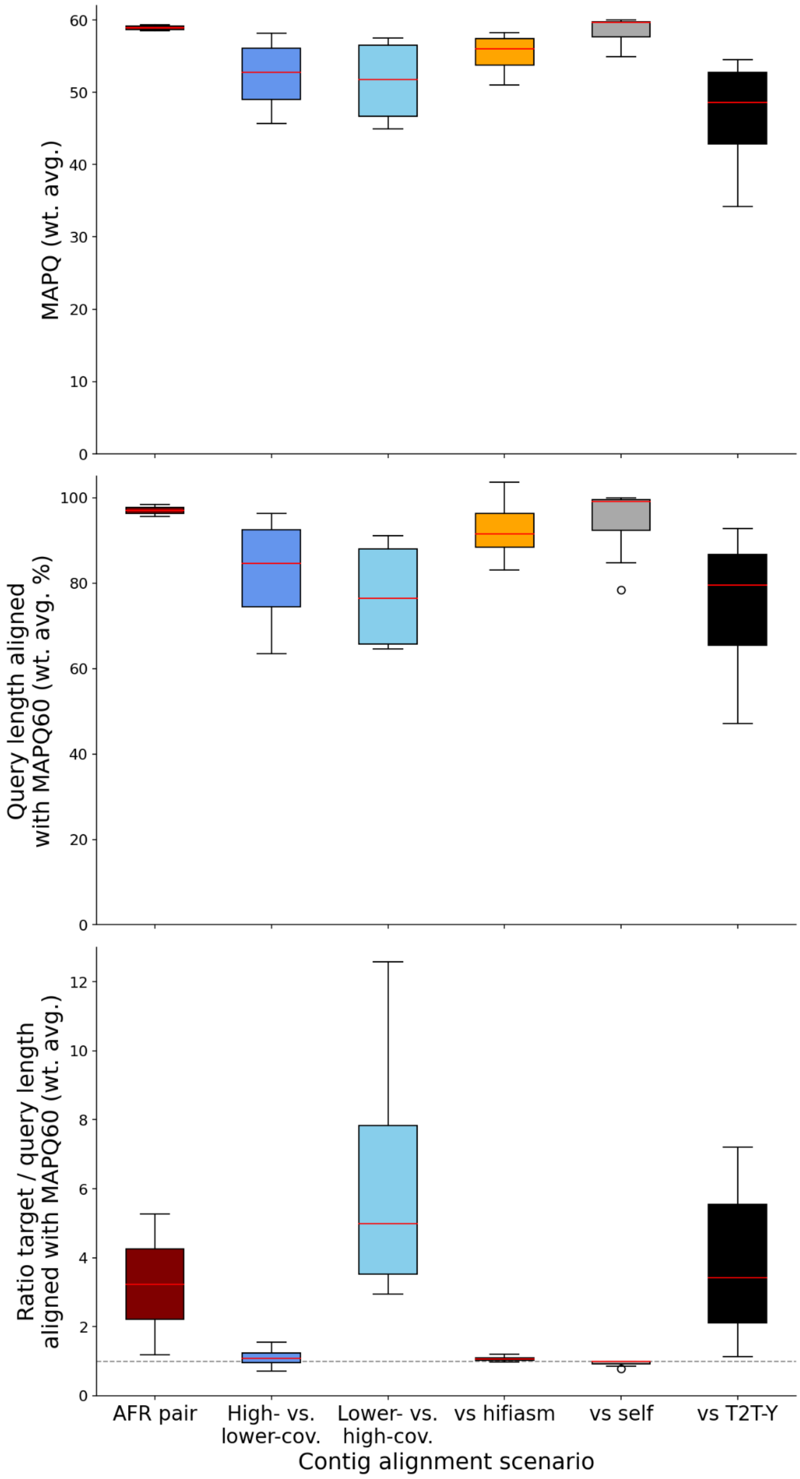


Figure S50. QC contig alignments for high-coverage samples in various scenarios: box plots depicting contig alignment statistics for (from left to right) the pair of closely related AFR samples (NA19317 and NA19347, dark red); the four selected high-coverage samples assembled with lower coverage for QC purposes, using the lower-coverage assembly as alignment target (dark blue) and vice versa the high-coverage assembly (light blue); the sample-matched alignment of Verkko- to hifiasm-assembled contigs (orange); the self-alignment of Verkko-assembled contigs (gray); contig-to-reference alignment using the T2T Y sequence as alignment target (black). Computed statistics per sample pair are (from top to bottom) average mapping quality (MAPQ) of the alignments weighted by alignment length (in bp); percent of the query sequence aligning with MAPQ 60, averaged over all contigs weighted by the contig length; ratio of target-to-query sequence lengths aligning with MAPQ 60, averaged over all contigs weighted by contig length.

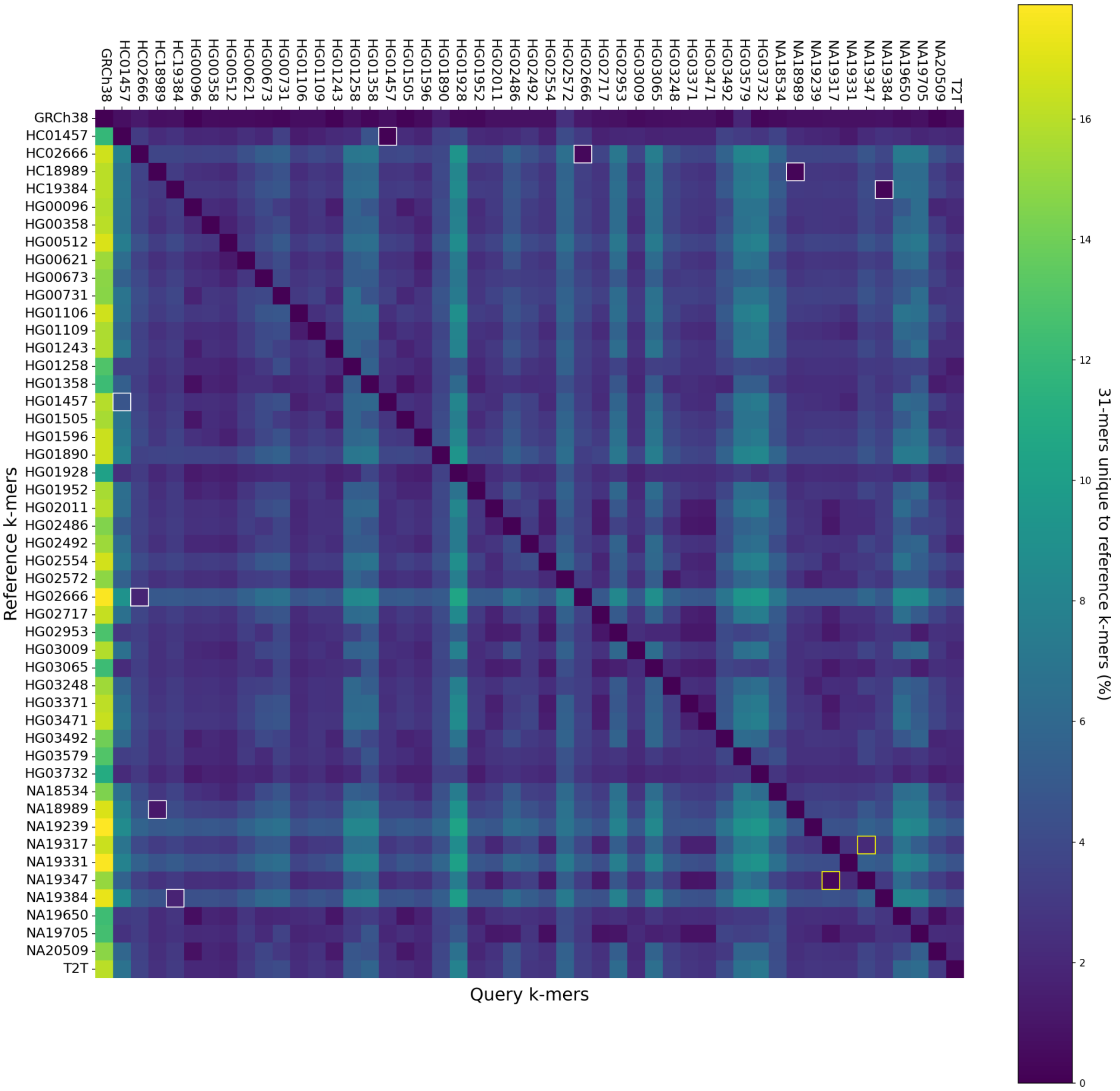


Figure S51. Unique sequence content in all assemblies expressed as percent of unique 31-mers relative to the respective query assembly. Comparisons of the high-/low-coverage pairs (HC02666/HG02666, HC01457/HG01457 etc.) are singled out by gray rectangles.

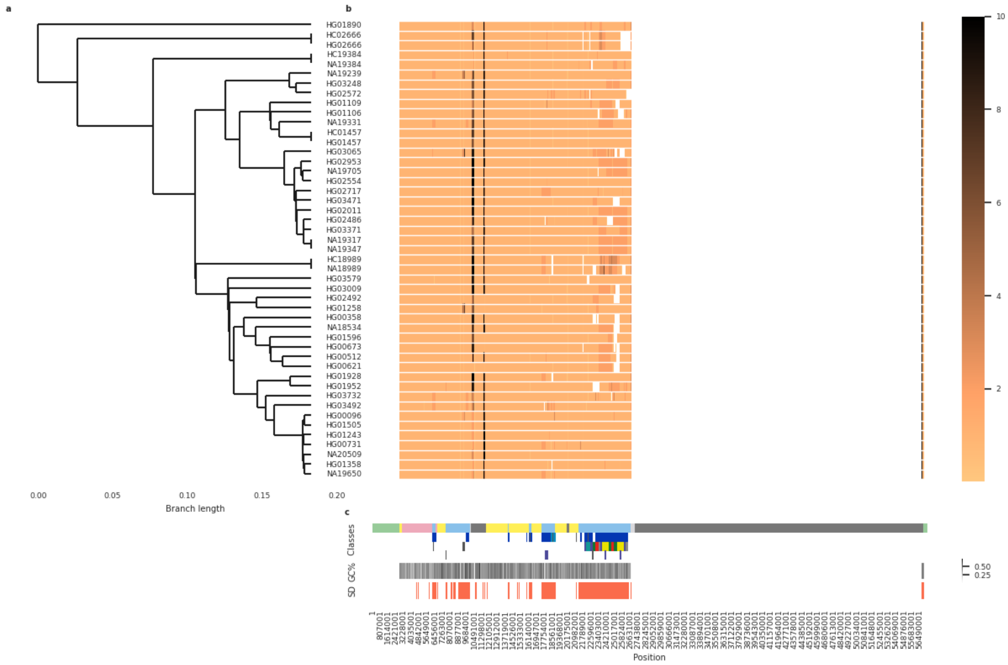


Figure S52. Composite plot depicting the Y contig alignments to the GRCh38 Y reference sequence across the whole Y chromosome span. **a.** Phylogenetic relationships of the samples (see **Fig. S1** for details). Note that two assemblies are visualized for 4 samples for which both high- and lower-coverage assemblies were generated (HG02666, NA19384, HG01457 and NA18989; HC - refers to the high-coverage assembly; see Methods). **b.** The coverage from Y contig alignments to reference sequence, with coverage=1 (light orange) in well-aligning regions. Darker shades indicate regions with multiple contig alignments, potentially indicating assembly errors or poor alignments, e.g., due to structural differences between the sample and reference or difficult sequence contexts such as high repeat content; white denotes regions with no coverage i.e., no contig to reference alignments (note - majority of the Yq12 subregion is not resolved in GRCh38, i.e., composed of 'Ns'). **c.** Y-chromosomal subregion locations as described in **Fig. 1a**, locations of inverted repeats (in dark blue) and *AZFc*/ampliconic subregion 7 segmental duplications as shown in **Fig. S22**, followed by GC% and segmental duplication locations (**Methods**).

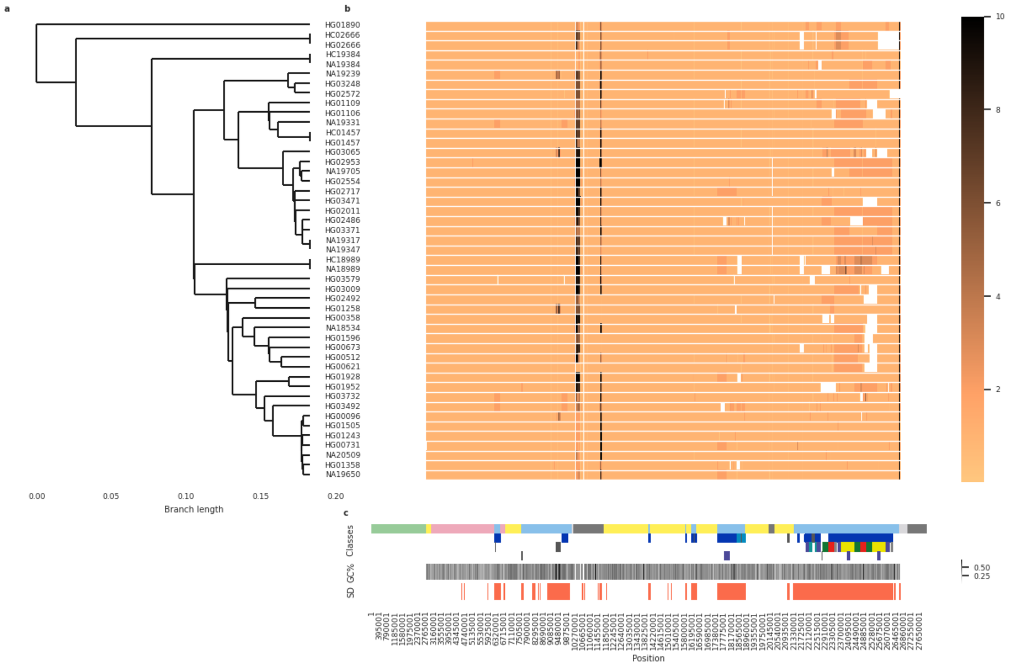


Figure S53. Composite plot depicting the Y contig alignments to the GRCh38 Y reference sequence excluding Yq12 and PAR2 subregions. **a.** Phylogenetic relationships of the samples (see **Fig. S1** for details). Note that two assemblies are visualized for 4 samples for which both high- and lower-coverage assemblies were generated (HG02666, NA19384, HG01457 and NA18989; HC - refers to the high-coverage assembly; see Methods). **b.** The coverage from Y contig alignments to reference sequence, with coverage=1 (light orange) in well-aligning regions. Darker shades indicate regions with multiple contig alignments, potentially indicating assembly errors or poor alignments, e.g., due to structural differences between the sample and reference or difficult sequence contexts such as high repeat content; white denotes regions with no coverage i.e., no contig to reference alignments (note - majority of the Yq12 subregion is not resolved in GRCh38, i.e., composed of 'Ns'). **c.** Y-chromosomal subregion locations as described in **Fig. 1a**, locations of inverted repeats (in dark blue) and *AZFc*/ampliconic subregion 7 segmental duplications as shown in **Fig. S22**, followed by GC% and segmental duplication locations (**Methods**).

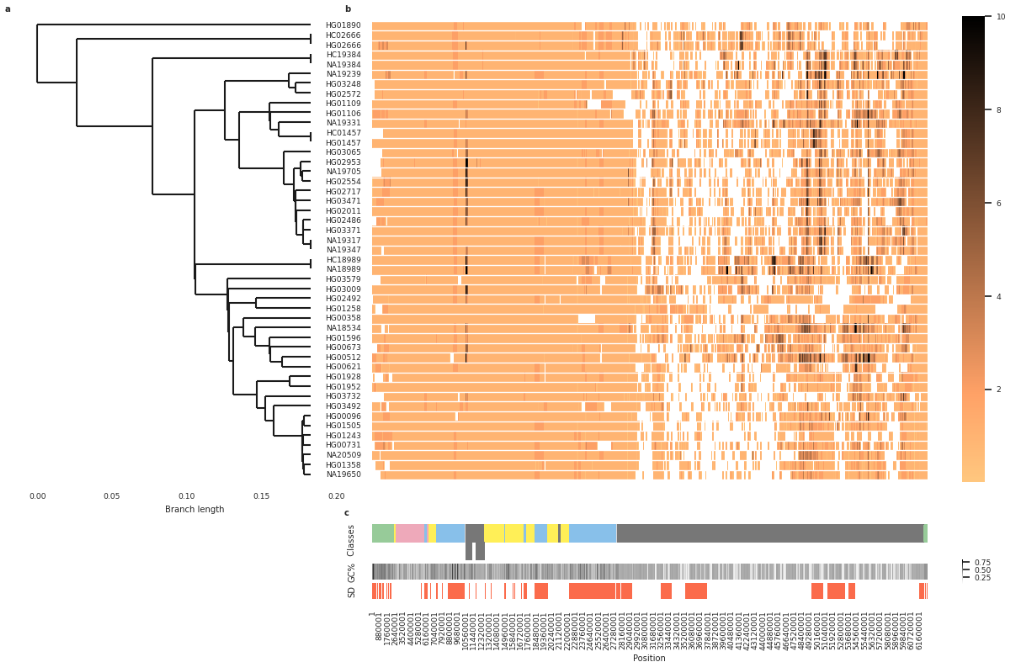


Figure S54. Composite plot depicting the Y contig alignments to the T2T Y reference sequence across the whole Y chromosome span. **a.** Phylogenetic relationships of the samples (see **Fig. S1** for details). Note that two assemblies are visualized for 4 samples for which both high- and lower-coverage assemblies were generated (HG02666, NA19384, HG01457 and NA18989; HC - refers to the high-coverage assembly; see Methods). **b.** The coverage from Y contig alignments to reference sequence, with coverage=1 (light orange) in well-aligning regions. Darker shades indicate regions with multiple contig alignments, potentially indicating assembly errors or poor alignments, e.g., due to structural differences between the sample and reference or difficult sequence contexts such as high repeat content; white denotes regions with no coverage i.e., no contig to reference alignments. **c.** Y-chromosomal subregion locations as described in **Fig. 1a**; below in gray locations of *DYZ3* (on the left) and *DYZ17* (on the right) repeat arrays, followed by GC% and segmental duplication locations (**Methods**).

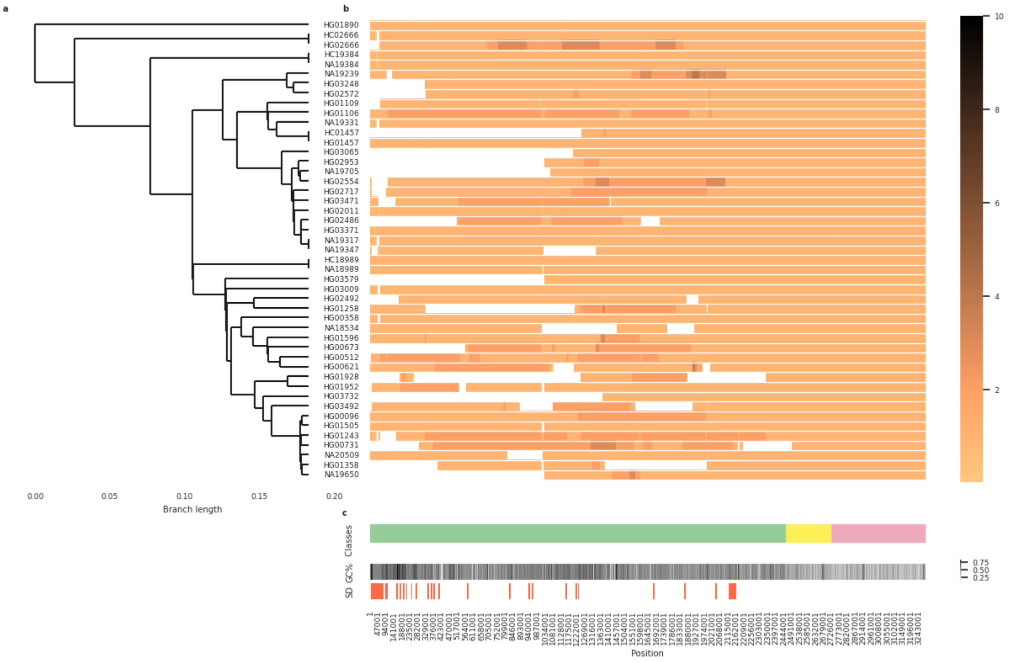


Figure S55. Composite plot depicting the Y contig alignments to the T2T Y reference sequence zooming into PAR1 subregion. **a.** Phylogenetic relationships of the samples (see **Fig. S1** for details). Note that two assemblies are visualized for 4 samples for which both high- and lower-coverage assemblies were generated (HG02666, NA19384, HG01457 and NA18989; HC - refers to the high-coverage assembly; see Methods). **b.** The coverage from Y contig alignments to reference sequence, with coverage=1 (light orange) in well-aligning regions. Darker shades indicate regions with multiple contig alignments, potentially indicating assembly errors or poor alignments, e.g., due to structural differences between the sample and reference or difficult sequence contexts such as high repeat content; white denotes regions with no coverage i.e., no contig to reference alignments. **c.** Y-chromosomal subregion locations as described in **Fig. 1a**, followed by GC% and segmental duplication locations (**Methods**).

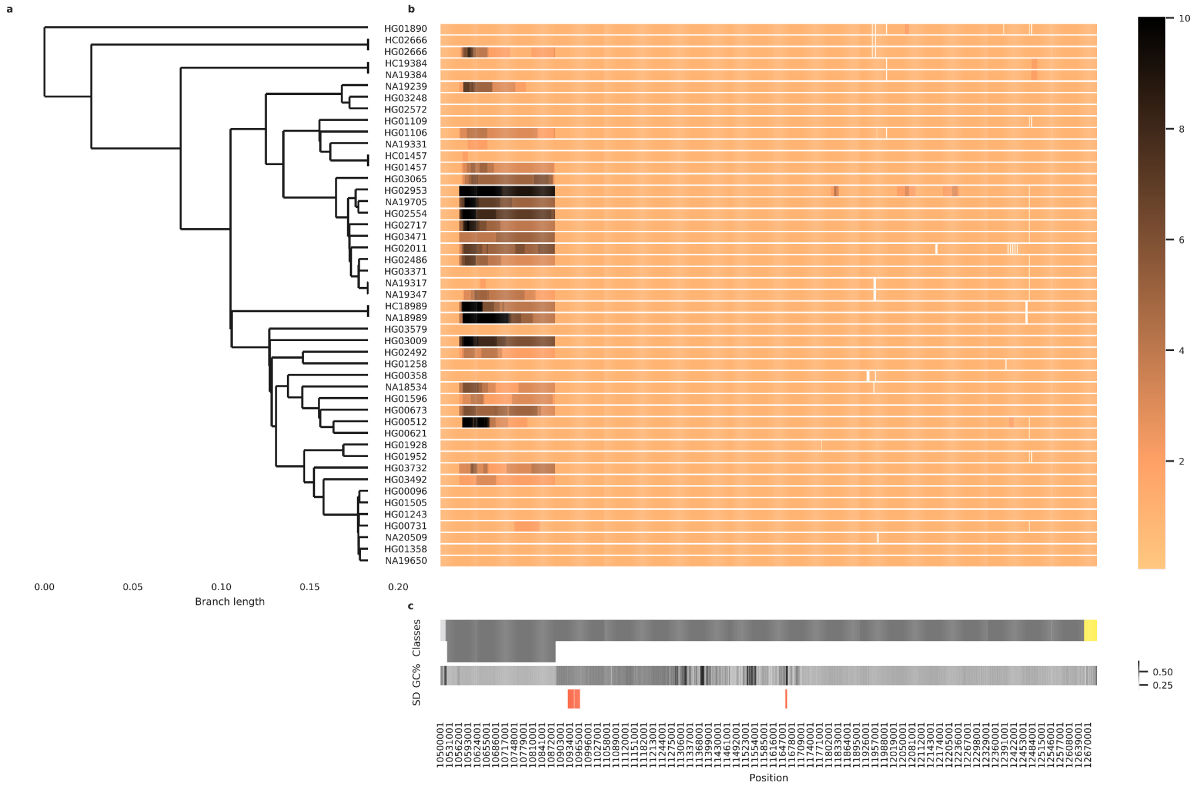


Figure S56. Composite plot depicting the Y contig alignments to the T2T Y reference sequence zooming into the (peri-)centromeric region. **a.** Phylogenetic relationships of the samples (see **Fig. S1** for details). Note that two assemblies are visualized for 4 samples for which both high- and lower-coverage assemblies were generated (HG02666, NA19384, HG01457 and NA18989; HC - refers to the high-coverage assembly; see Methods). **b.** The coverage from Y contig alignments to reference sequence, with coverage=1 (light orange) in well-aligning regions. Darker shades indicate regions with multiple contig alignments, potentially indicating assembly errors or poor alignments, e.g., due to structural differences between the sample and reference or difficult sequence contexts such as high repeat content; white denotes regions with no coverage i.e., no contig to reference alignments. **c.** Location of the (peri-)centromeric region (above) and the *DYZ3* α-satellite repeat array (below) shown in dark gray, followed by GC% and segmental duplication locations (**Methods**).

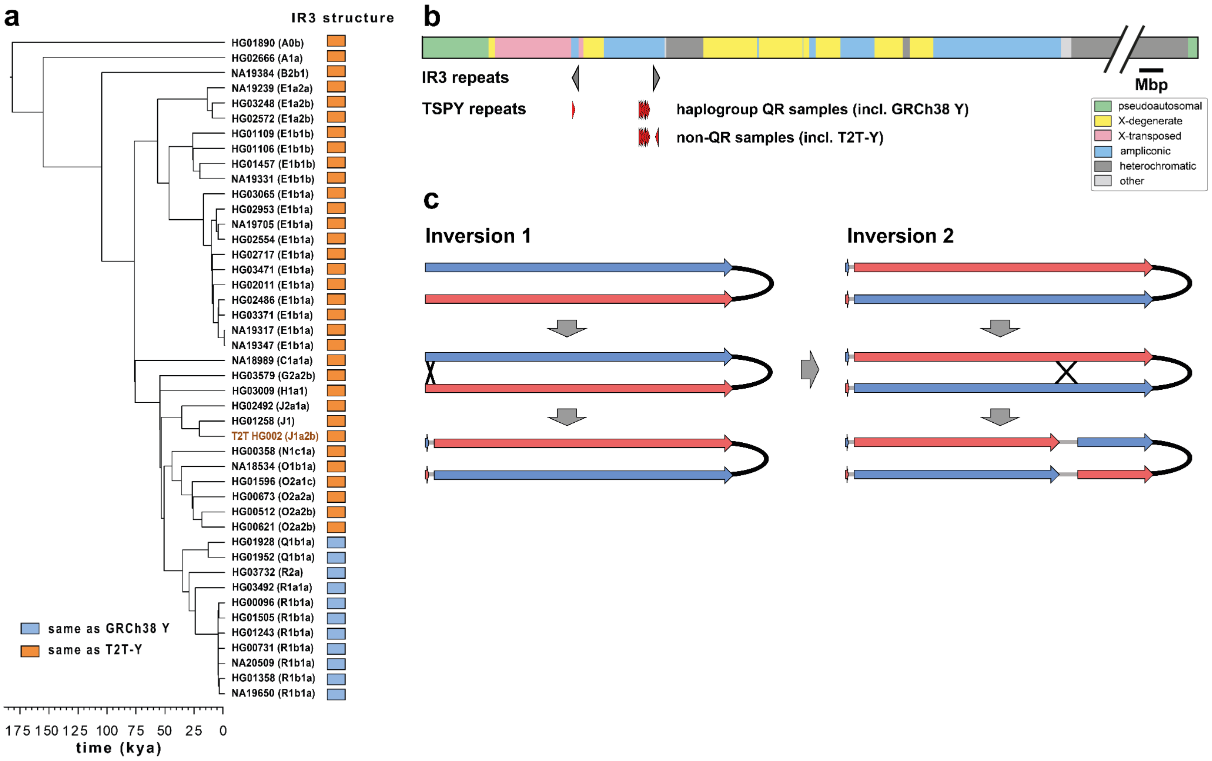


Figure S57. Phylogenetic distribution of different IR3 repeat compositions and the responsible IR3 inversion. **a.** Distribution of two different IR3 repeat compositions in the Y-chromosomal phylogeny. In orange - samples containing a single TSPY repeat in the proximal IR3 repeat in inverted orientation, in blue - samples containing a single TSPY repeat in the distal IR3 repeat in direct orientation. **b.** Schematic representation of IR3 composition and approximate locations of TSPY repeats relative to the Y chromosome structure. **c.** Identified inversions in phylogenetically related QR haplogroup samples - one changing the location and orientation of the single TSPY repeat from proximal to distal IR3 repeat, and another reversing the orientation of the region in between IR3 repeats. The inversions are indicated by black crosses. Blue and red arrows indicate distal and proximal IR3 repeats, respectively.

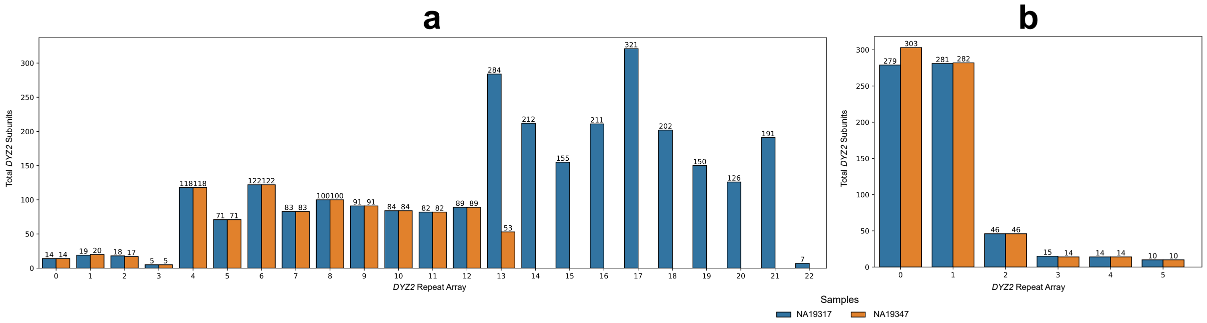


Figure S58. The bar plots show a comparison of the total *DYZ2* repeat copy numbers (y-axis) in each *DYZ2* array (x-axis) within the two most closely related genomes (NA19317 (blue) and NA19347 (orange)). **a.** *DYZ2* arrays within the proximally assembled contigs. **b.** *DYZ2* arrays within the distally assembled contigs. The analyses revealed an equal number of *DYZ2* repeats within 14 of 20 arrays.

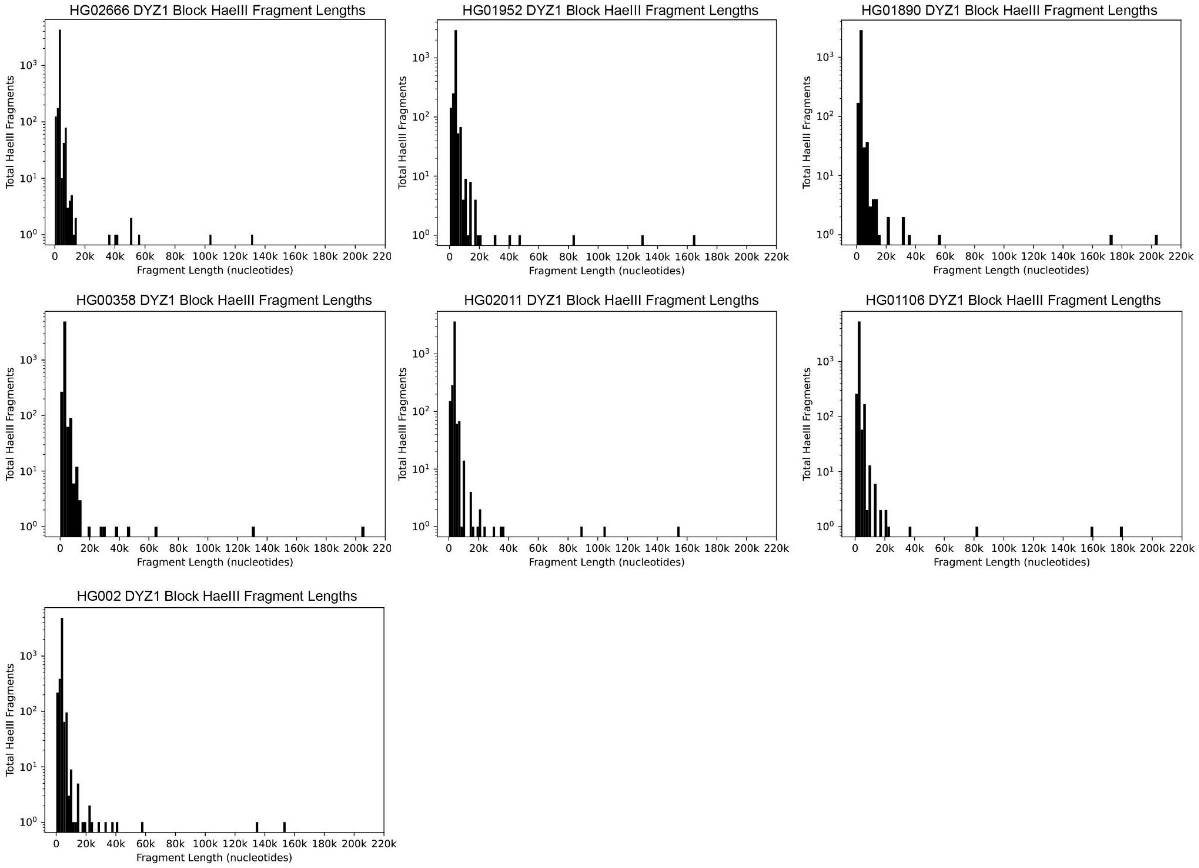


Figure S59. The distribution of total *DYZ1* array HaeIII virtual restriction digestion fragments (y-axis) and their lengths (x-axis) for each genome with a completely assembled Yq12 region is shown in the histograms. The majority of *DYZ1* repeat units were between 3-4 kbp in length within each genome.

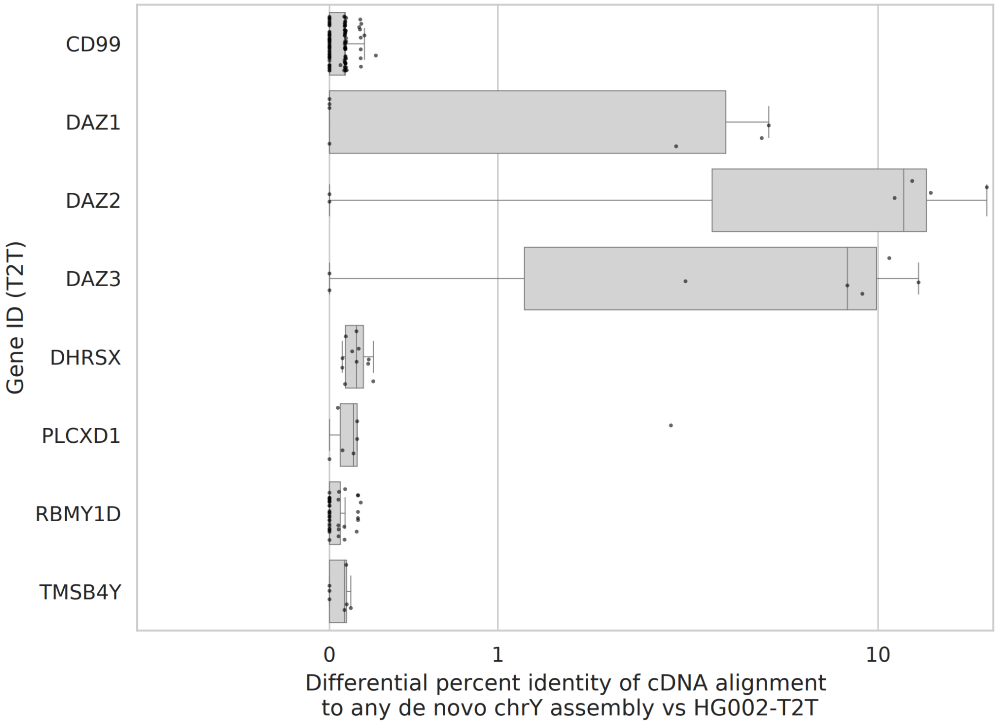
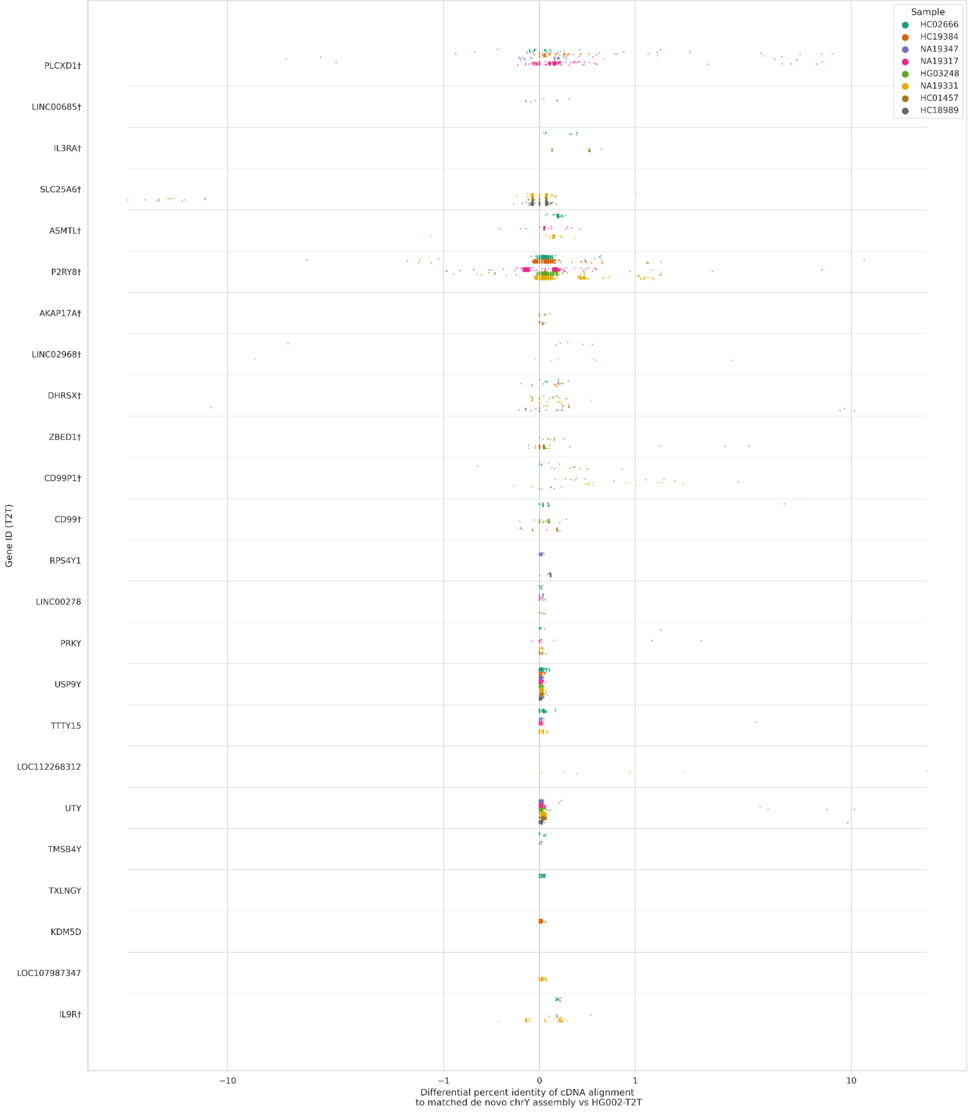


Figure S60. Testis Iso-seq percent identity to *de novo* assemblies compared to the T2T Y reference sequence. Each dot represents an individual cDNA read, and its position on the x-axis represents the numeric difference between percent identity of the read alignment to the T2T Y reference and the alignment to the best-matching *de novo* Y assembly. Gene IDs are based on alignment position to the reference.



I

Figure S61: Iso-seq percent identity to matched *de novo* assemblies compared to the T2T Y reference sequence. Each dot represents an individual cDNA read, and its position on the x-axis represents the numeric difference between percent identity of the read alignment to the T2T Y reference and the alignment to its sample-matched *de novo* Y assembly. Gene IDs are based on alignment position to the reference, with † indicating genes located in either PAR. Colors specify the sample for both the Iso-Seq library and *de novo* assembly.

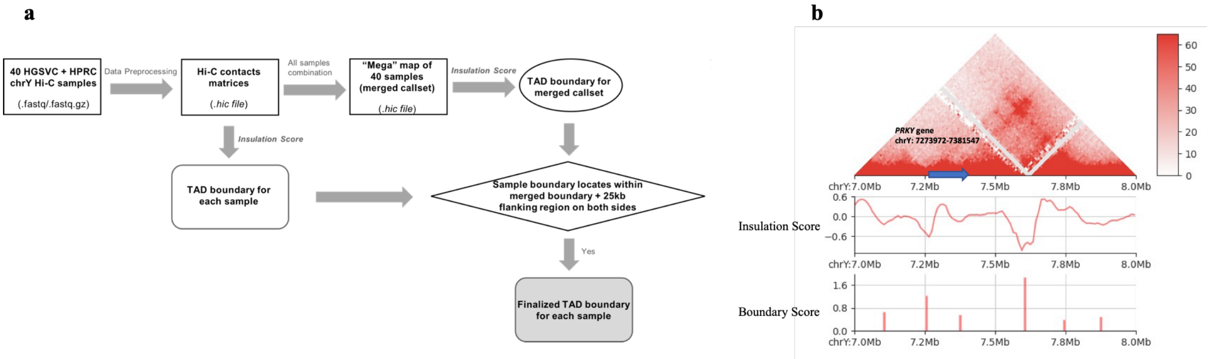


Figure S62. A step-by-step workflow to generate TAD boundaries in our chrY Hi-C analysis pipeline and a visualization of the chrY Hi-C merged callset calling results generated in our pipeline. **a.** 40 samples' raw reads were used as an input in Juicer to do preprocessing and create Hi-C maps which were binned at multiple resolutions. Insulation score algorithm was applied to call TAD boundaries for each sample on each of those 40 .hic files separately. All 40 .hic files were then merged together to create a "mega" map and used as an input of insulation score algorithm to call TAD boundaries for the chrY merged callset. A finalized TAD boundary results for each sample were defined as those sample boundaries located within the merged boundary plus 25 kb flanking regions on the left side of the boundary start site and the right side of the boundary end site. **b.** The Hi-C contact map, the insulation score and the boundary strength for the merged callset over the region chrY: 7Mb-8Mb. The blue arrow shows where the *PRKY* gene is.

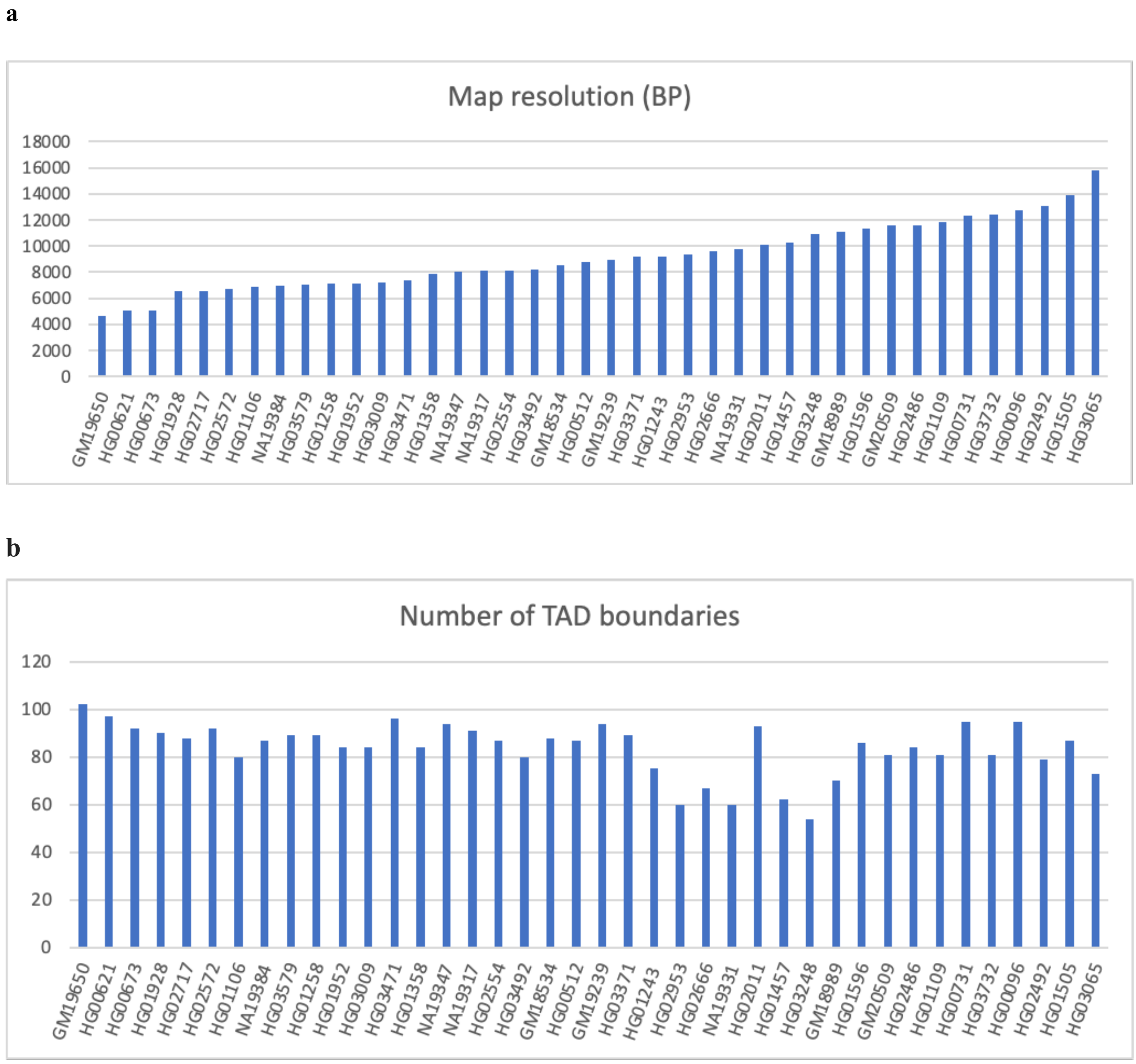


Figure S63. The map resolution (bp) for 40 Hi-C samples and the corresponding TAD boundaries detected by our strategy. **a.** As described in [30], the map resolution was calculated by calculate_map_resolution.sh script given by Juicer. The highest resolution is 4,650 bp while the lowest resolution is 15,800 bp. To average, 10 kbp resolution was chosen for the further analysis. **b.** The number the TAD boundaries for each sample which were redefined from the workflow shown in **Figure S62**.

References

1. 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).
2. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
3. 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).
4. Rhie, A. *et al.* The complete sequence of a human Y chromosome. bioRxiv 2022
5. Repping, S. *et al.* A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics* **83**, 1046–1052 (2004).
6. 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).
7. 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).
8. Shi, W. *et al.* Evolutionary and functional analysis of RBMY1 gene copy number variation on the human Y chromosome. *Hum. Mol. Genet.* **28**, 2785–2798 (2019).
9. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
10. Porubsky, D. *et al.* Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* **185**, 1986–2005.e26 (2022).
11. Repping, S. *et al.* Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.* **71**, 906–922 (2002).
12. Cooke, H. J., Schmidtke, J. & Gosden, J. R. Characterisation of a human Y chromosome repeated sequence and related sequences in higher primates. *Chromosoma* **87**, 491–502 (1982).
13. Schmid, M., Guttenbach, M., Nanda, I., Studer, R. & Epplen, J. T. Organization of DYZ2 repetitive DNA on the human Y chromosome. *Genomics* **6**, 212–218 (1990).
14. 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).
15. Cooke, H. J. & McKay, R. D. Evolution of a human Y chromosome-specific repeated sequence. *Cell* **13**, 453–460 (1978).
16. 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).

17. 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).
18. Yadav, S. K., Kumari, A., Javed, S. & Ali, S. DYZ1 arrays show sequence variation between the monozygotic males. *BMC Genet.* **15**, 19 (2014).
19. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
20. Cooke, H. J., Fantes, J. & Green, D. Structure and evolution of human Y chromosome DNA. *Differentiation* **23 Suppl**, S48–55 (1983).
21. 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).
22. Ray, D. A., Xing, J., Salem, A.-H. & Batzer, M. A. SINEs of a nearly perfect character. *Syst. Biol.* **55**, 928–935 (2006).
23. Smith, G. P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).
24. Stevison, L. S., Hoehn, K. B. & Noor, M. A. F. Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* **3**, 830–841 (2011).
25. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
26. 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.
27. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
28. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
29. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29**, 279–286 (2001).
30. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).