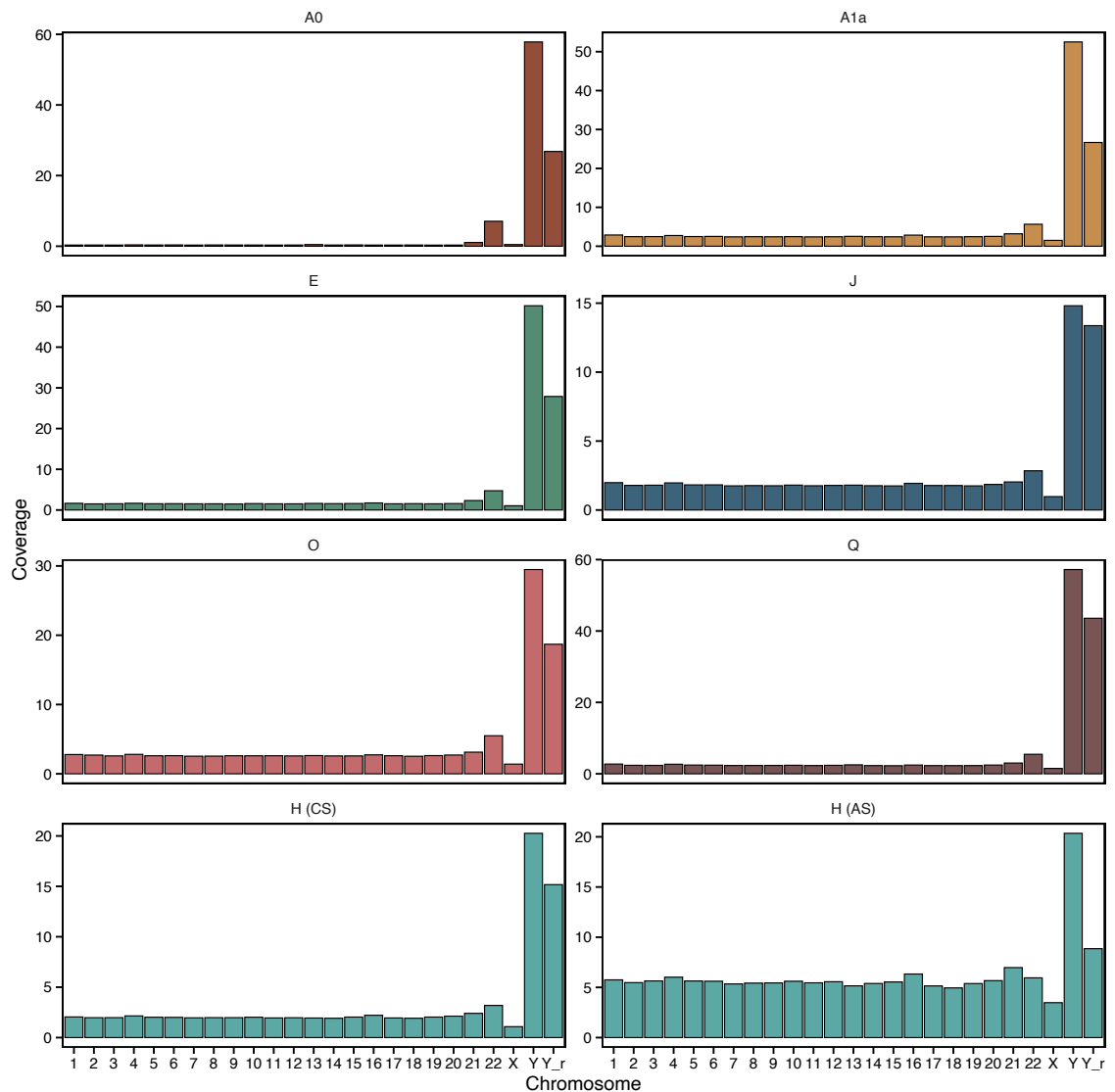


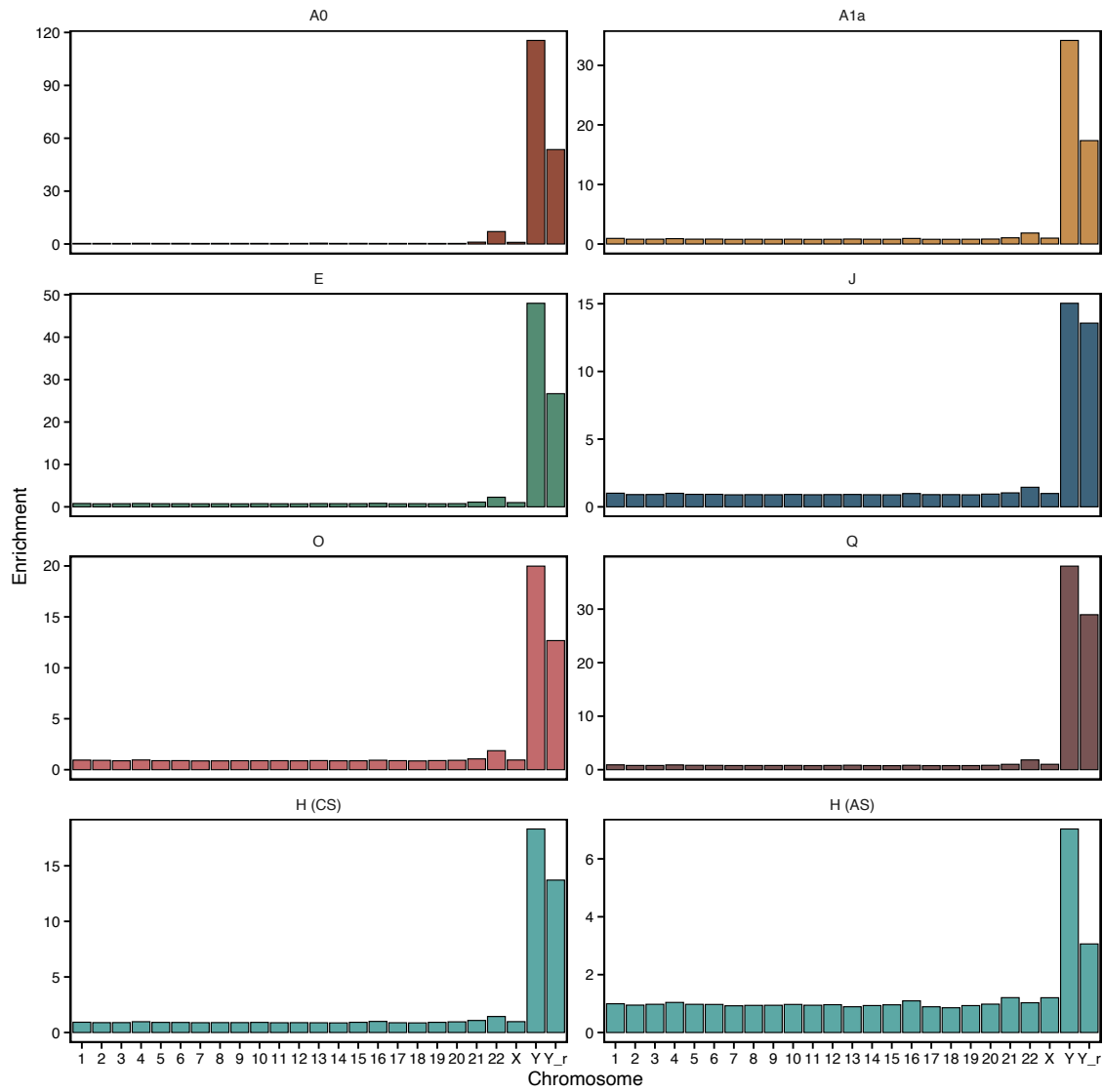
Y chromosome sequence and epigenomic reconstruction across human populations

Paula Esteller-Cucala^{*#}, Marc Palmada-Flores^{*}, Lukas F. K. Kuderna, Claudia Fontserè, Aitor Serres-Armero, Marc Dabad, María Torralvo, Armida Faella, Luis Ferrández-Peral, Laia Llovera, Oscar Fornas, Eva Julià, Erika Ramírez, Irene González, Jochen Hecht, Esther Lizano, David Juan, Tomàs Marquès-Bonet[#]

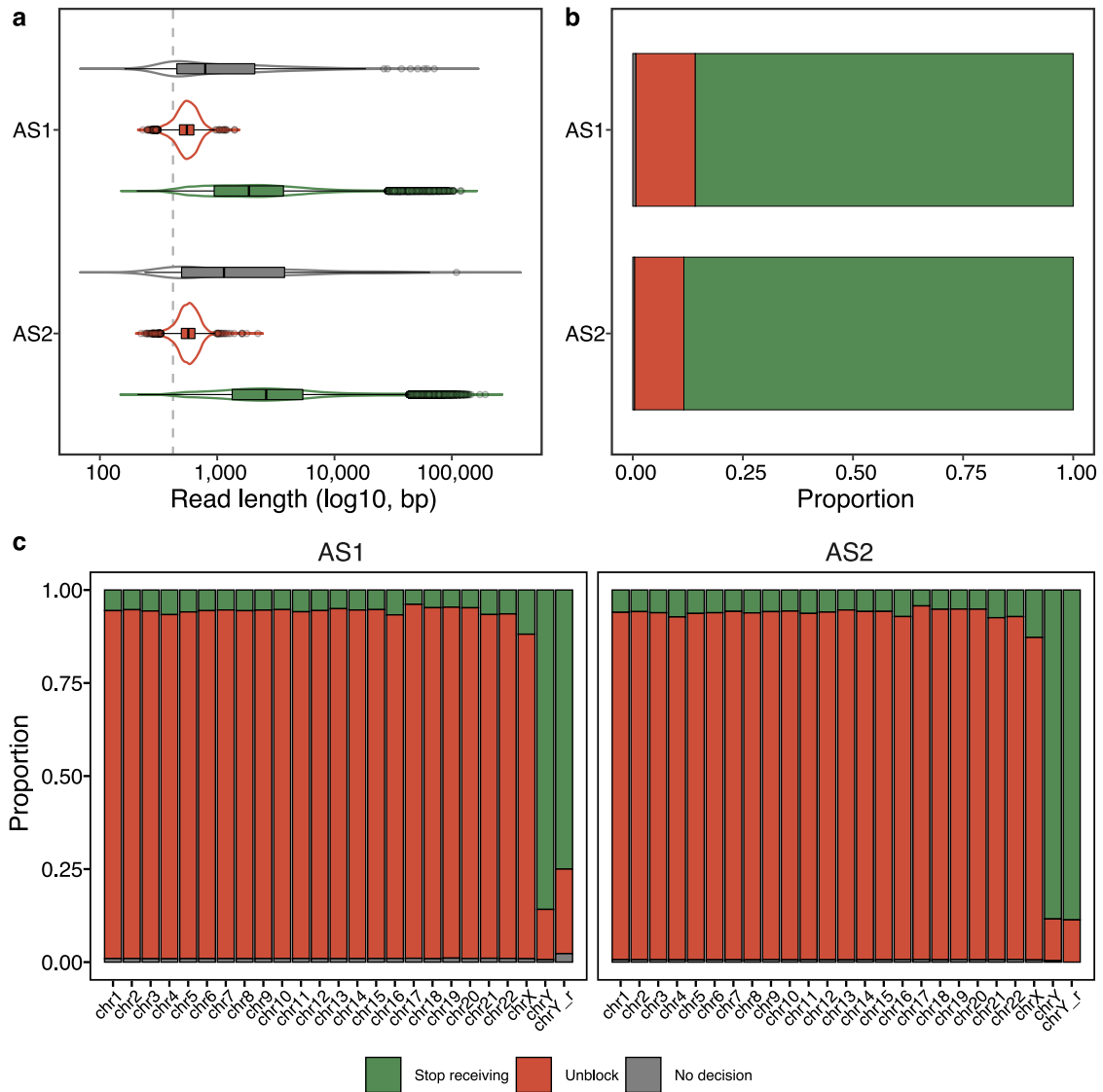
Supplementary Figures:



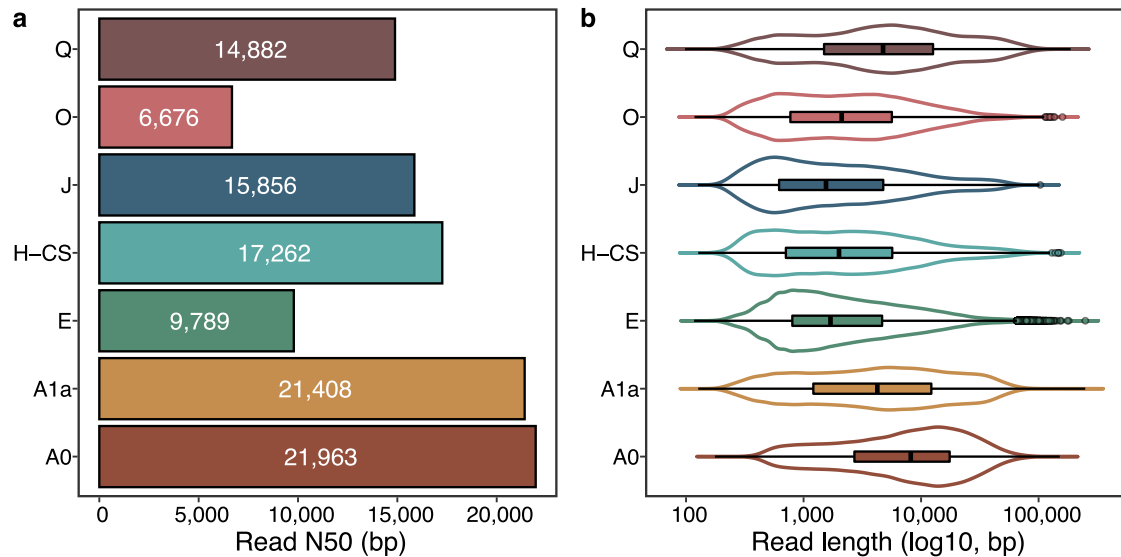
Supplementary Figure 1. Coverage of the 22 autosomal chromosomes, X chromosome, Y chromosome, and the random contig of the Y chromosome (Y_r) by haplogroup. H haplogroup separated by methodology: CS refers to chromosome sorting and AS refers to adaptive sampling.



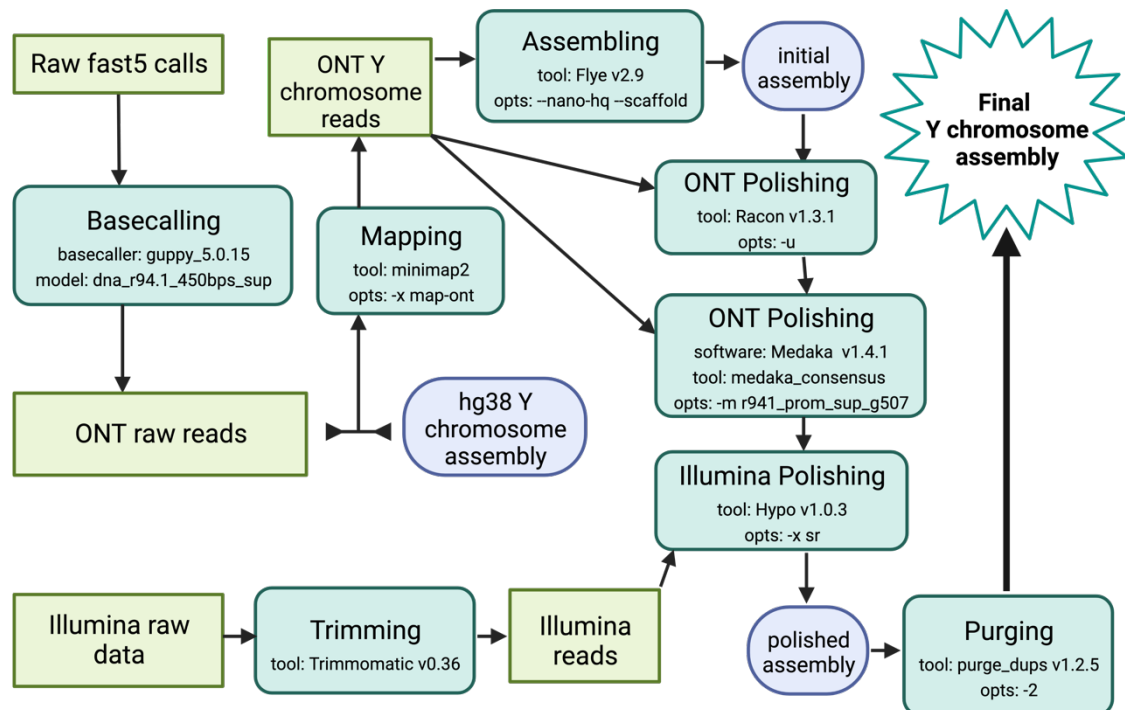
Supplementary Figure 2. Enrichment specificity of the sequencing data generated for the 22 autosomal chromosomes, X chromosome, Y chromosome and the random contig of the Y chromosome (Y_r) by haplogroup. H haplogroup separated by methodology: CS refers to chromosome sorting and AS refers to adaptive sampling.



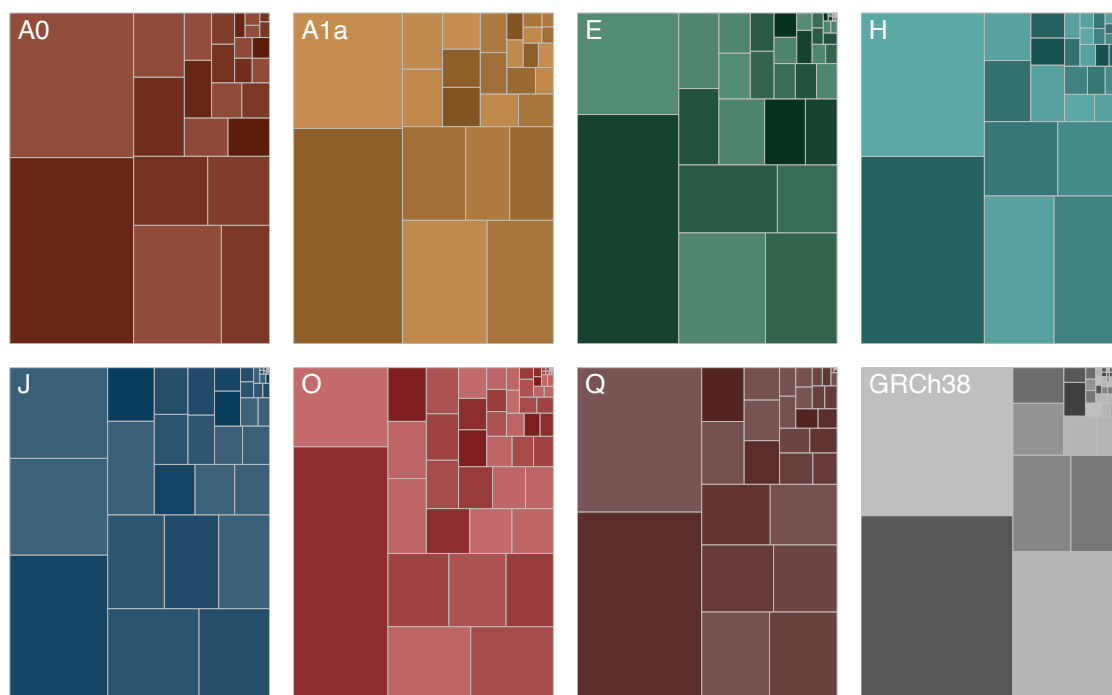
Supplementary Figure 3. Adaptive sampling (AS) metrics in the two runs generated for the H haplogroup sample (AS1 and AS2). (a) Read length and (b) Proportion of reads mapping to the GRCh38 chromosome Y (chrY) or the random chromosome Y contig (chrY_r) grouped by AS read decision in each AS run. The number of reads in AS1 is $n_{\text{stop-receiving}} = 57,790$, $n_{\text{unblock}} = 9,084$, and $n_{\text{no-decision}} = 450$ reads, and in AS2 $n_{\text{stop-receiving}} = 41,928$, $n_{\text{unblock}} = 5,316$, and $n_{\text{no-decision}} = 188$ reads. (c) Chromosome-wise proportion of mapping reads in each AS run grouped by read decision type. The dashed grey line in (a) indicates 420 bases, the theoretical length each read is sequenced before the sequencer makes a decision. AS decision types: *unblock* reads are rejected by the pore, *no decision* reads are inconclusive, and *stop receiving* reads are further sequenced. Box plots in (a) show medians and the first and third quartiles (the 25th and 75th percentiles), respectively. The upper and lower whiskers extend the largest and smallest value no further than $1.5 \times \text{IQR}$.



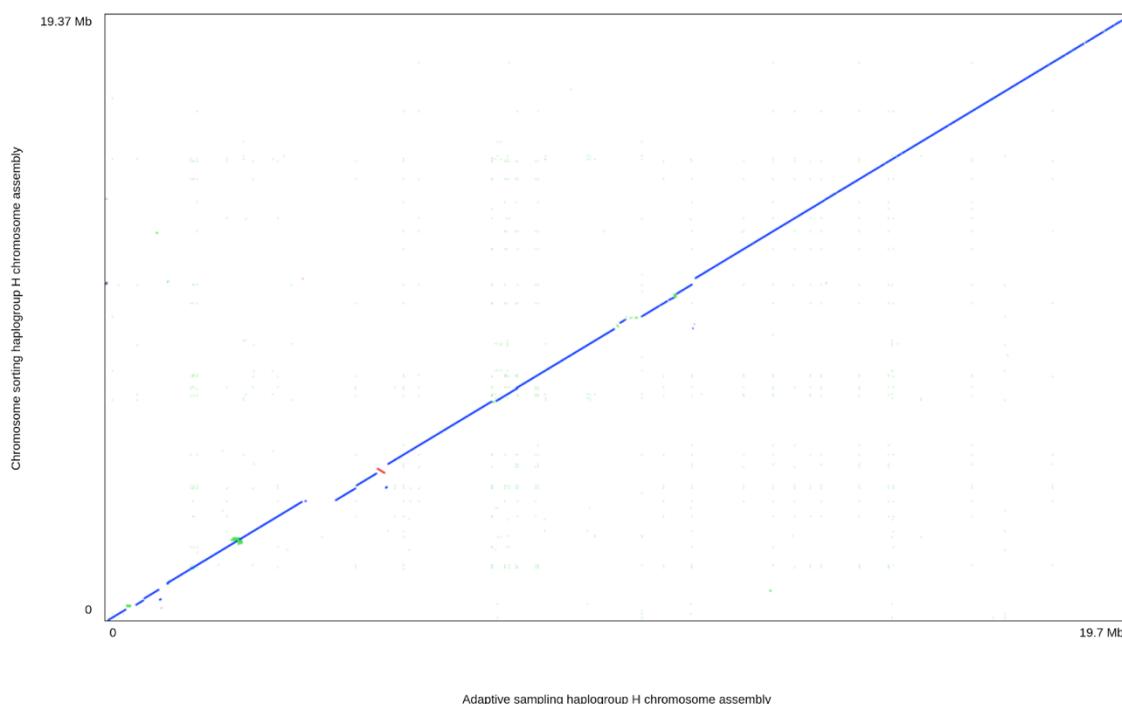
Supplementary Figure 4. (a) Read length N50 values. (b) Read length distributions of reads mapping to the GRCh38 chromosome Y or the random chromosome Y contig for each haplogroup. In (b) read lengths are shown in log10 scale. Only the chromosome sorting data for the H haplogroup is shown (H-CS). Box plots in (b) show medians and the first and third quartiles (the 25th and 75th percentiles), respectively. The upper and lower whiskers extend the largest and smallest value no further than 1.5×IQR. Number of reads per haplogroup: $n_{A0} = 77,701$, $n_{A1a} = 161,109$, $n_E = 322,225$, $n_{H-CS} = 103,606$, $n_J = 90,717$, $n_O = 178,476$, and $n_Q = 149,573$ reads.



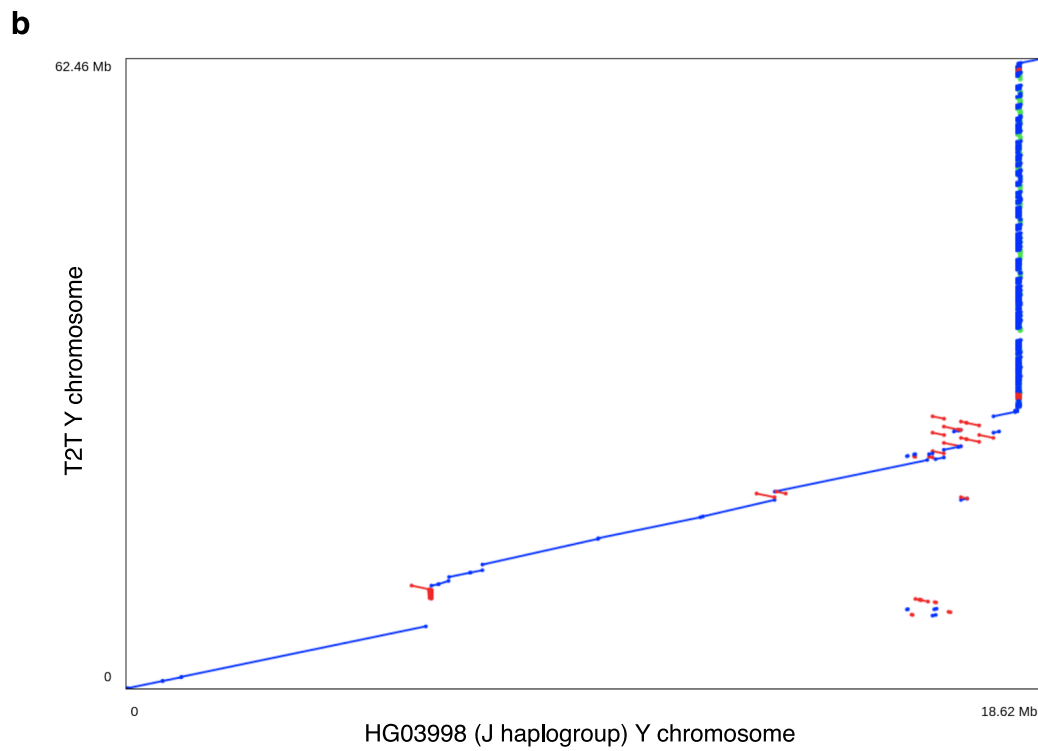
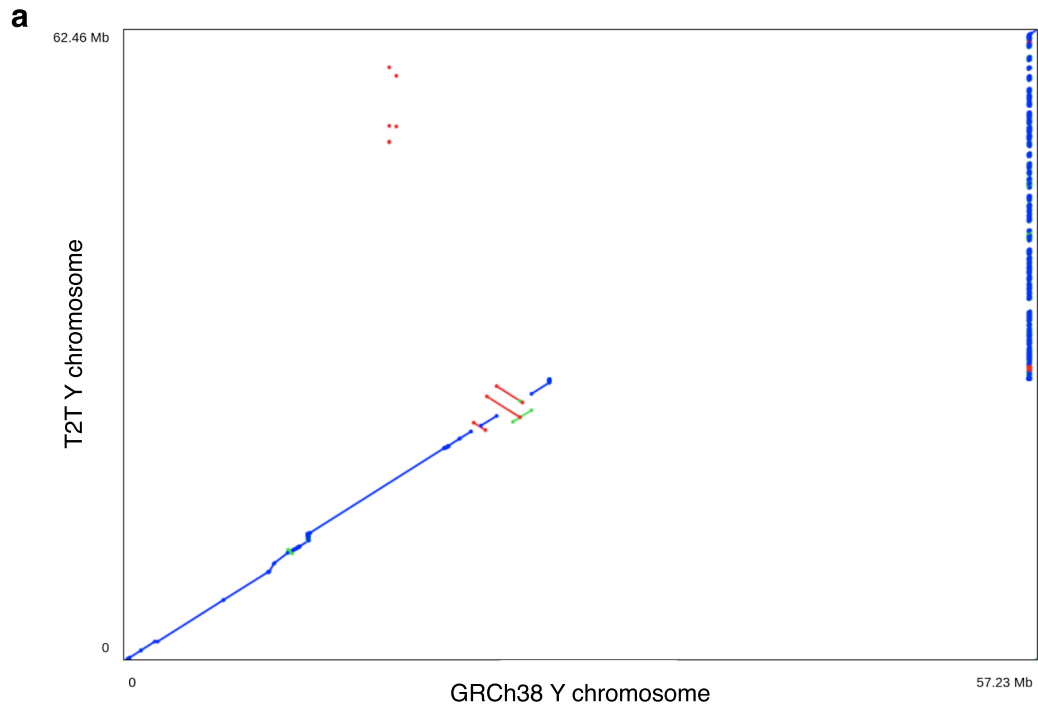
Supplementary Figure 5. Reference-guided Y chromosome assembly pipeline using chromosome sorting data. We consider both chrY and random chrY contig as the hg38 Y chromosome assembly. opts = running options; ONT = Oxford Nanopore Technologies. Created with BioRender.com.



Supplementary Figure 6. Treemaps comparing the contiguity of the assemblies obtained of the different chrY haplogroups and GRCh38 chrY. The size of each rectangle corresponds to the size of a contig within each of the assemblies.



Supplementary Figure 7. Dot-plot comparing haplogroup H (cell line GM21113) assemblies, manually scaffolded based on alignments to GRCh38, obtained using same amount of data of two different selective sequencing methods. The assembly generated from adaptive sampling is shown in the X axis and the assembly generated from chromosome sorting data is shown in the Y axis.



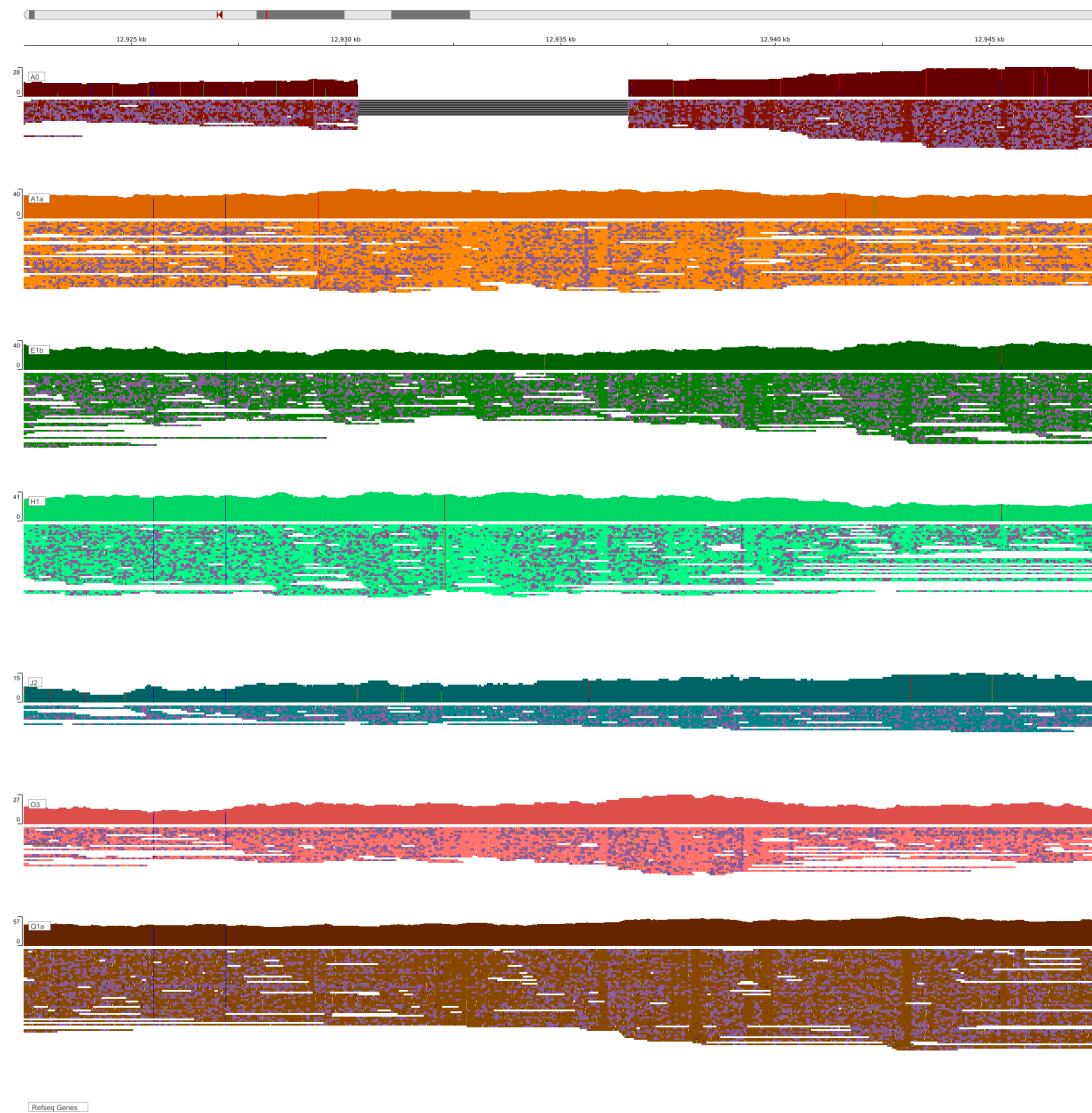
Supplementary Figure 8. Dot-plot comparing Y chromosome assemblies. (a) T2T Y chromosome vs GRCh38 Y chromosome assembly and (b) T2T Y chromosome vs J haplogroup (HG03998 cell line) assembly manually scaffolded based on alignments to GRCh38. Segments are colored blue, red or green, which denote forward, reverse or repetitive alignments, respectively.



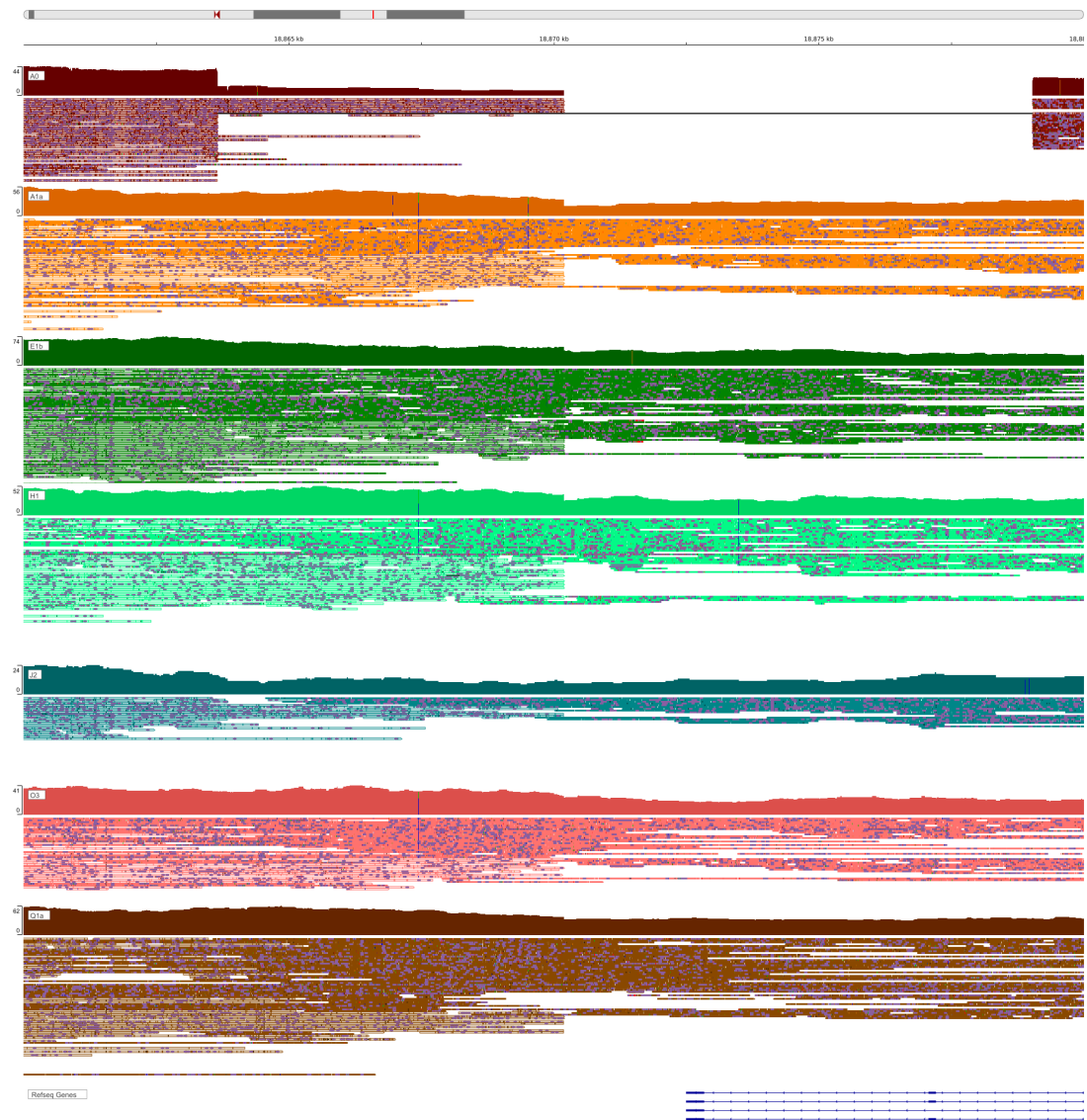
Supplementary Figure 9. IGV-screenshot showing ONT reads of 7 different Y haplogroups mapping to GRCh38 region with the longest insertion (shown in violet) found using *Sniffles*. It is only detected in the A0 haplogroup reads and not in the rest.



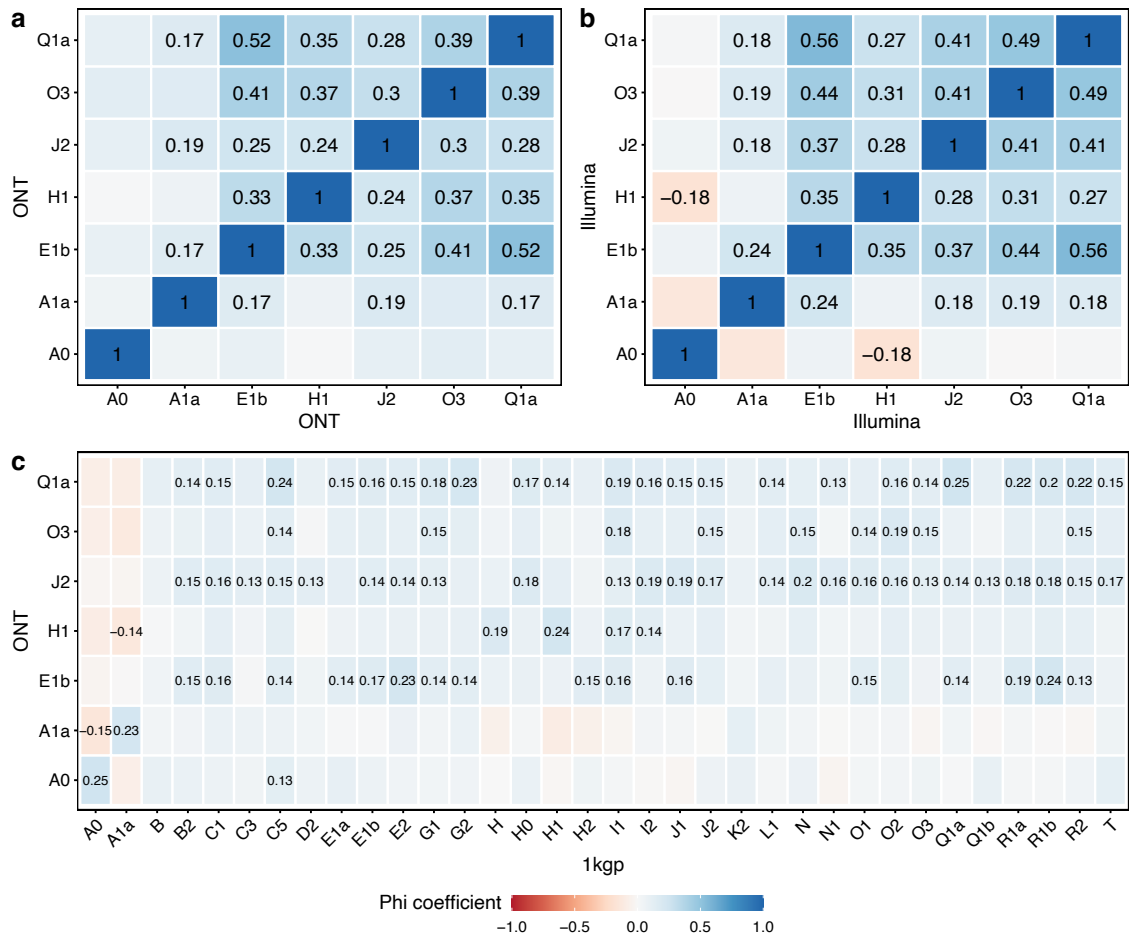
Supplementary Figure 10. IGV-screenshot showing ONT reads of 7 different Y haplogroups mapping to GRCh38 region with the longest deletion found using *Sniffles*. It is observed in the A0 haplogroup.



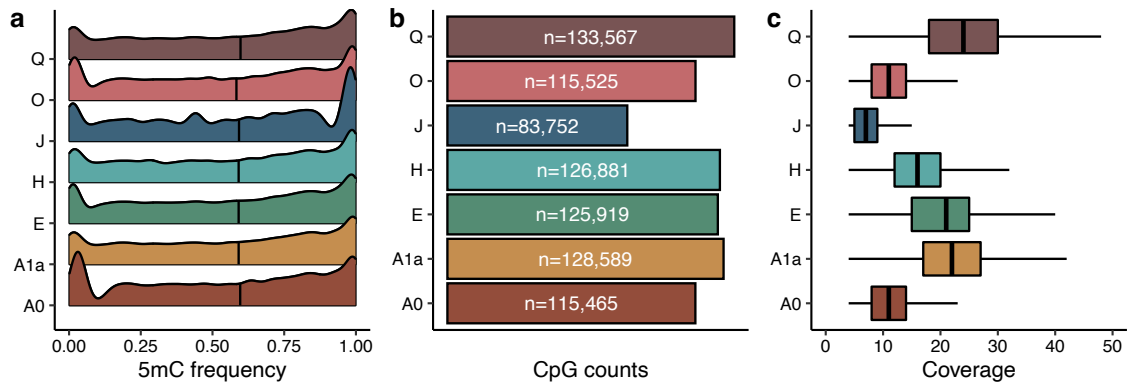
Supplementary Figure 11. IGV-screenshot showing ONT reads of 7 different Y haplogroups mapping to GRCh38 for a 1000 Genomes Project variant for HG02982 (A0): Deletion nearby DDX3Y. Observed only in A0.



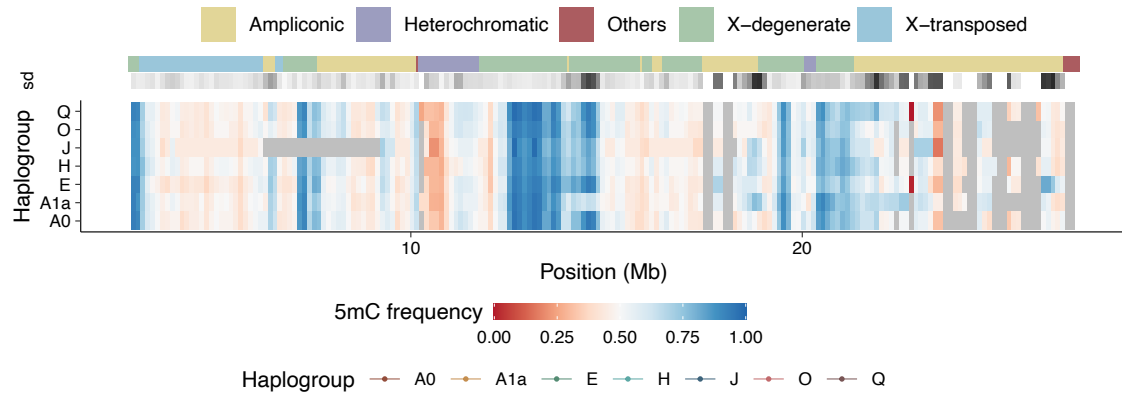
Supplementary Figure 12. IGV-screenshot showing ONT reads of 7 different Y haplogroups mapping to GRCh38 for a 1000 Genomes Project variant for HG02982 (A0): The deletion overlaps with the TTTY14 locus, nearby a segmental duplication.



Supplementary Figure 13. Correlations of structural variant genotypes. (a) Correlations of genotypes obtained with *Sniffles v2* on ONT data. (b) Correlations of genotypes obtained using *graphTyper* on Illumina data. (c) Correlation of genotypes obtained using *graphTyper* on 1kbp short read data with the ONT data. Structural variants found in frequencies > 0.2 for the 1kbp data were considered as present and lower frequencies as absent. Only correlation values that are statistically significant (p-value < 0.05) after Bonferroni multiple testing correction are shown.

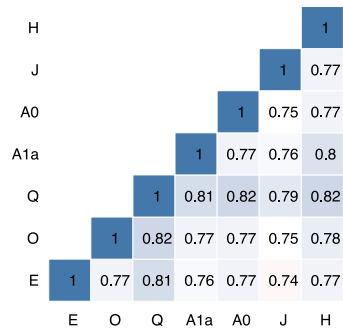


Supplementary Figure 14. (a) 5mC frequency levels, (b) Number of CpGs and (c) Coverage for each sample after quantile normalization (minimum coverage considered is 4x). Vertical black lines in (a) represent median 5mC frequency values. Box plots in (c) show medians and the first and third quartiles (the 25th and 75th percentiles), respectively. The upper and lower whiskers extend the largest and smallest value no further than $1.5 \times \text{IQR}$.

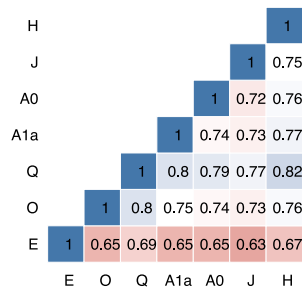


Supplementary Figure 15. Consistency in the frequency of 5mC in the seven cell lines along the resolved MSY of the hg38 is shown as tiles. The methylation levels are calculated as the median 5mC frequency value in 250kb sliding windows for each cell line. Each window requires a minimum of 10 CpG sites with methylation information. The sequence classes and the standard deviation (sd) of the methylation levels across cell lines are also shown. The standard deviation of the 5mC frequency is represented in a white-to-black scale, in which a darker color denotes a higher standard deviation value.

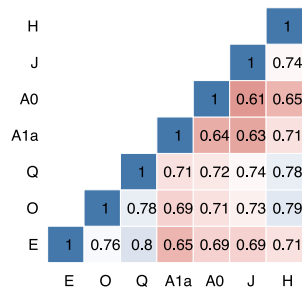
a All positions



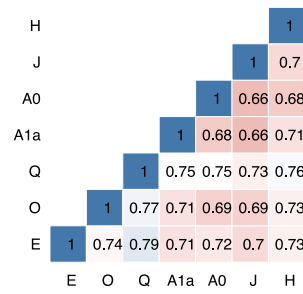
b Ampliconic



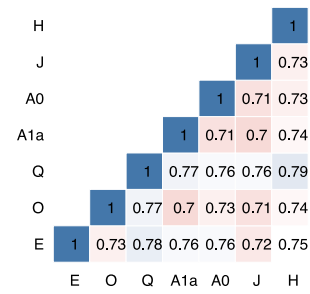
Pseudo-autosomal



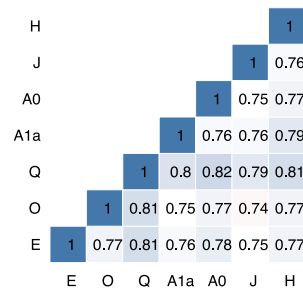
Heterochromatic



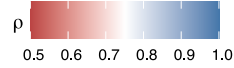
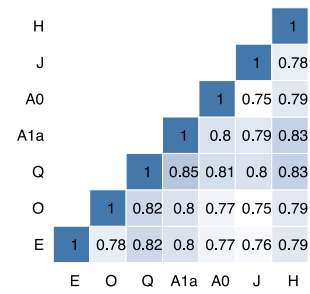
Others



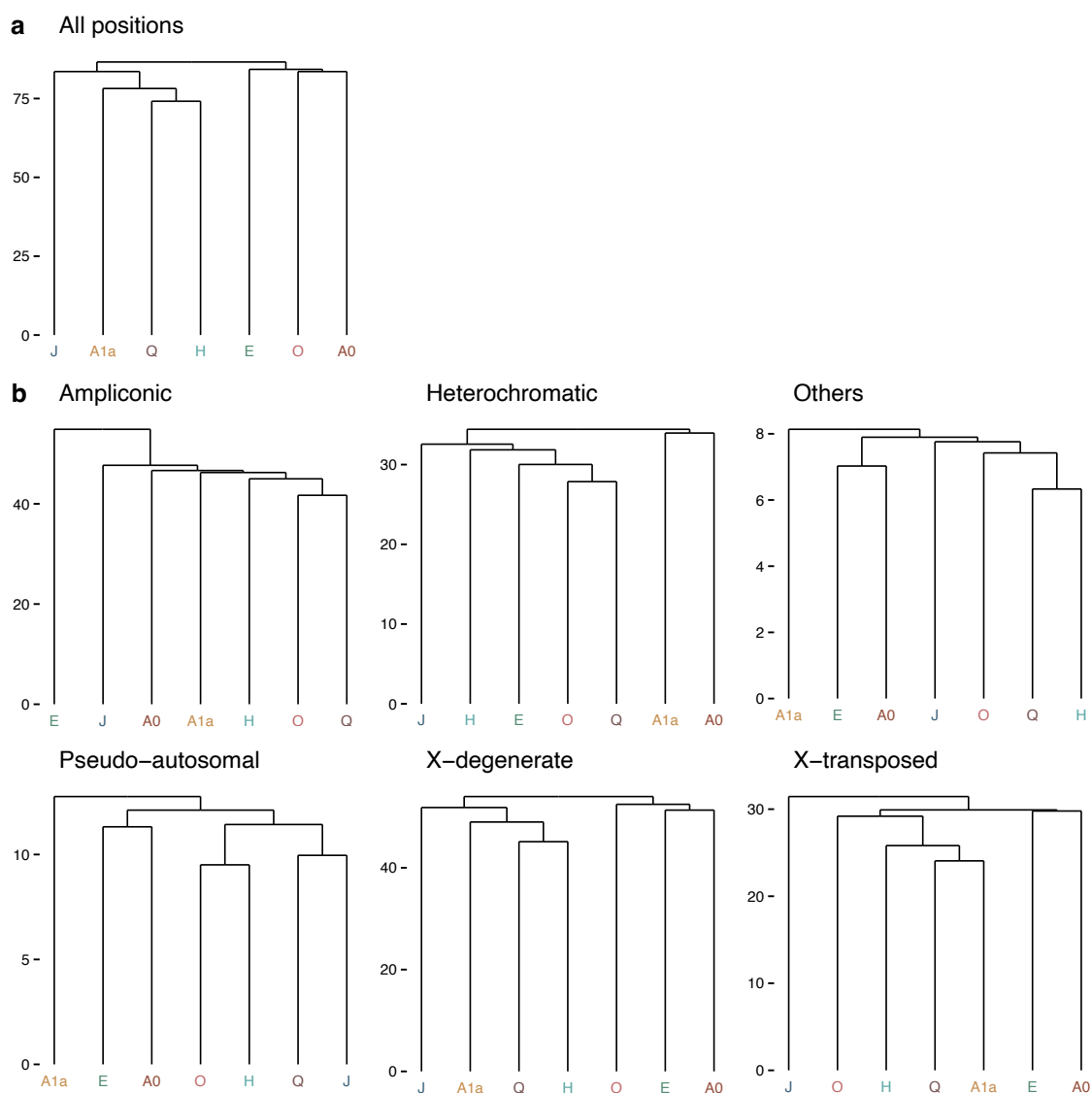
X-degenerate



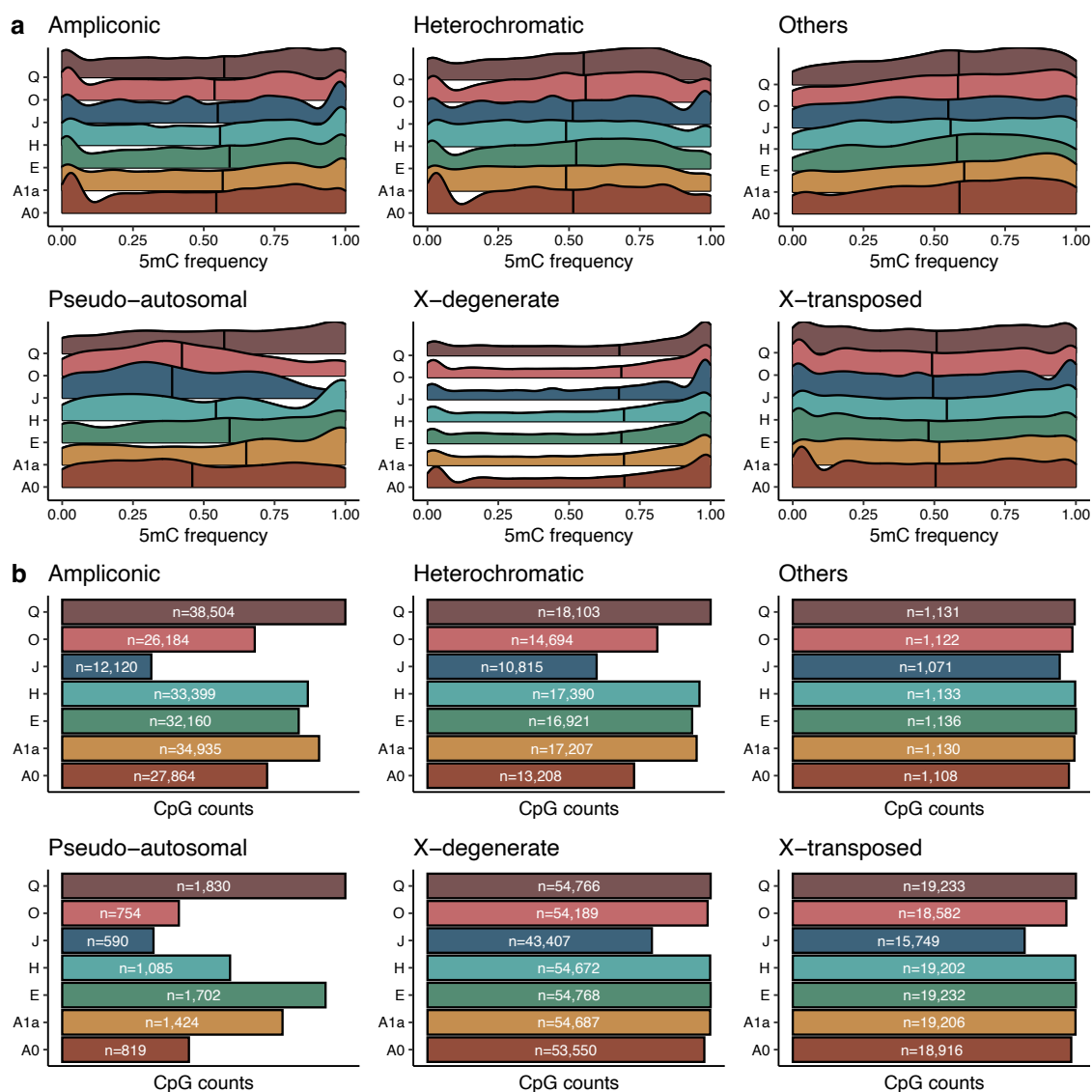
X-transposed



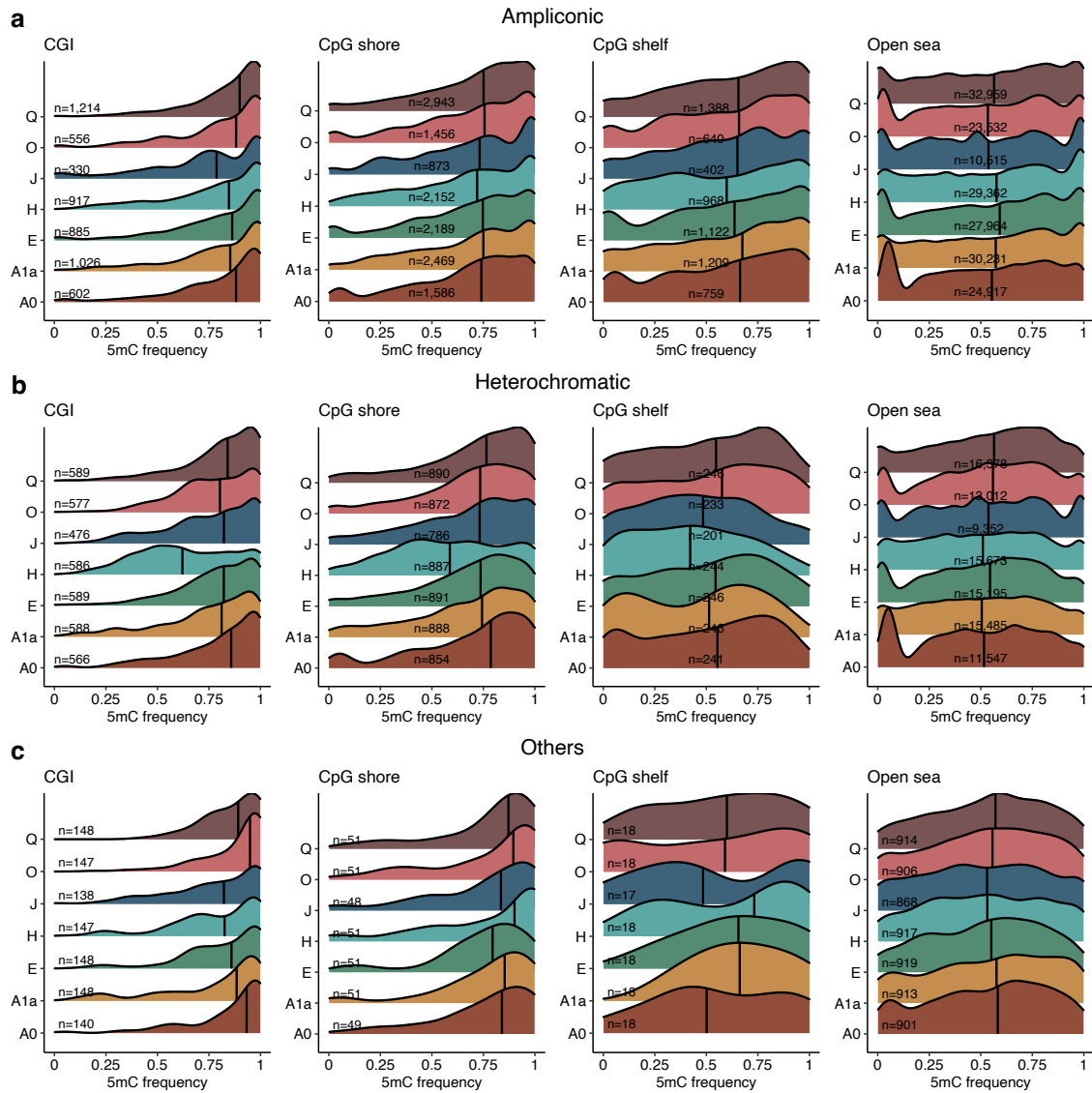
Supplementary Figure 16. High correlation of 5mC frequency in the seven cell lines calculated using the Spearman pair-wise correlation in (a) all CpG positions and in (b) sequence classes.



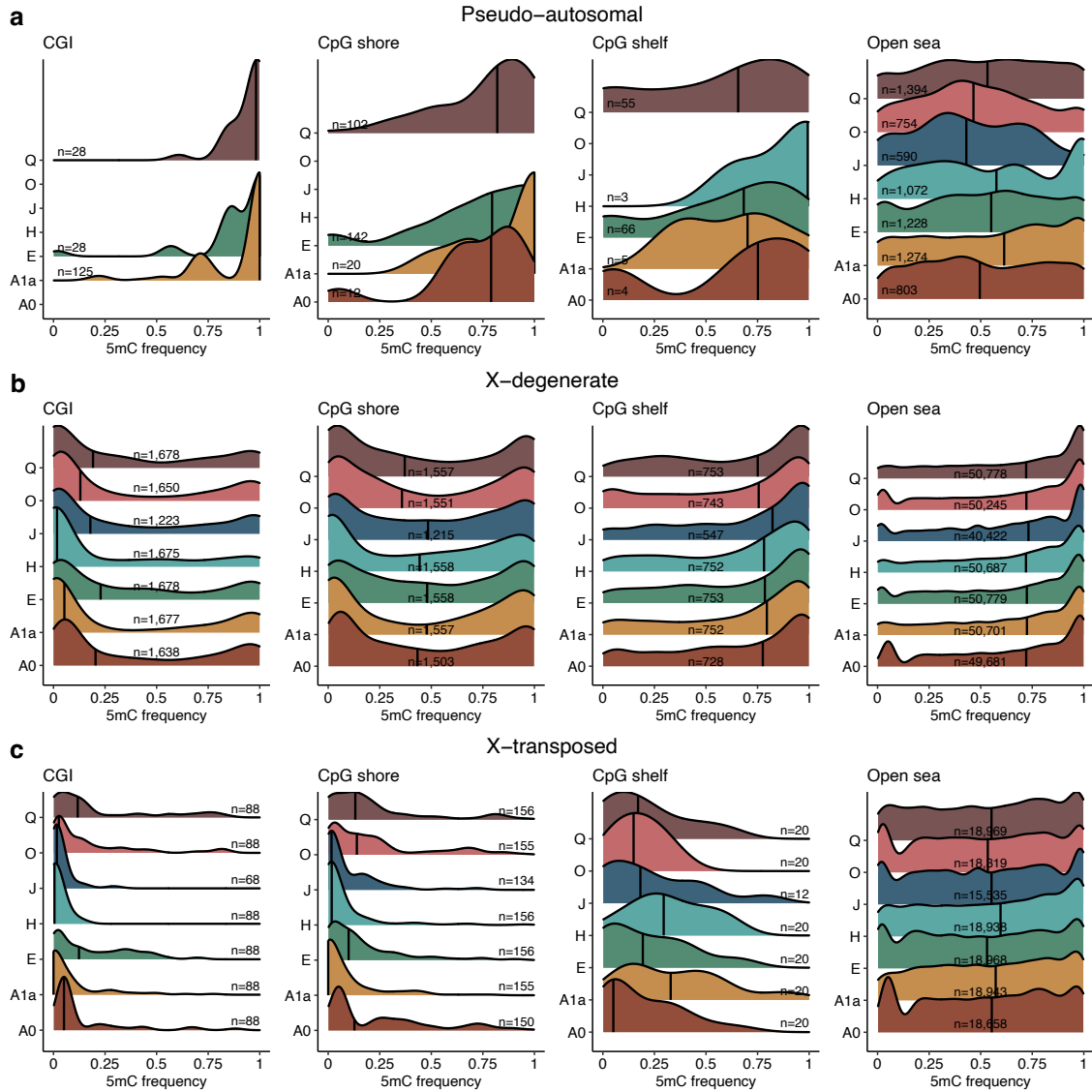
Supplementary Figure 17. Hierarchical cluster tree calculated using the Euclidean distance of 5mC frequency in the seven cell lines using (a) all CpG positions and (b) sequence classes.



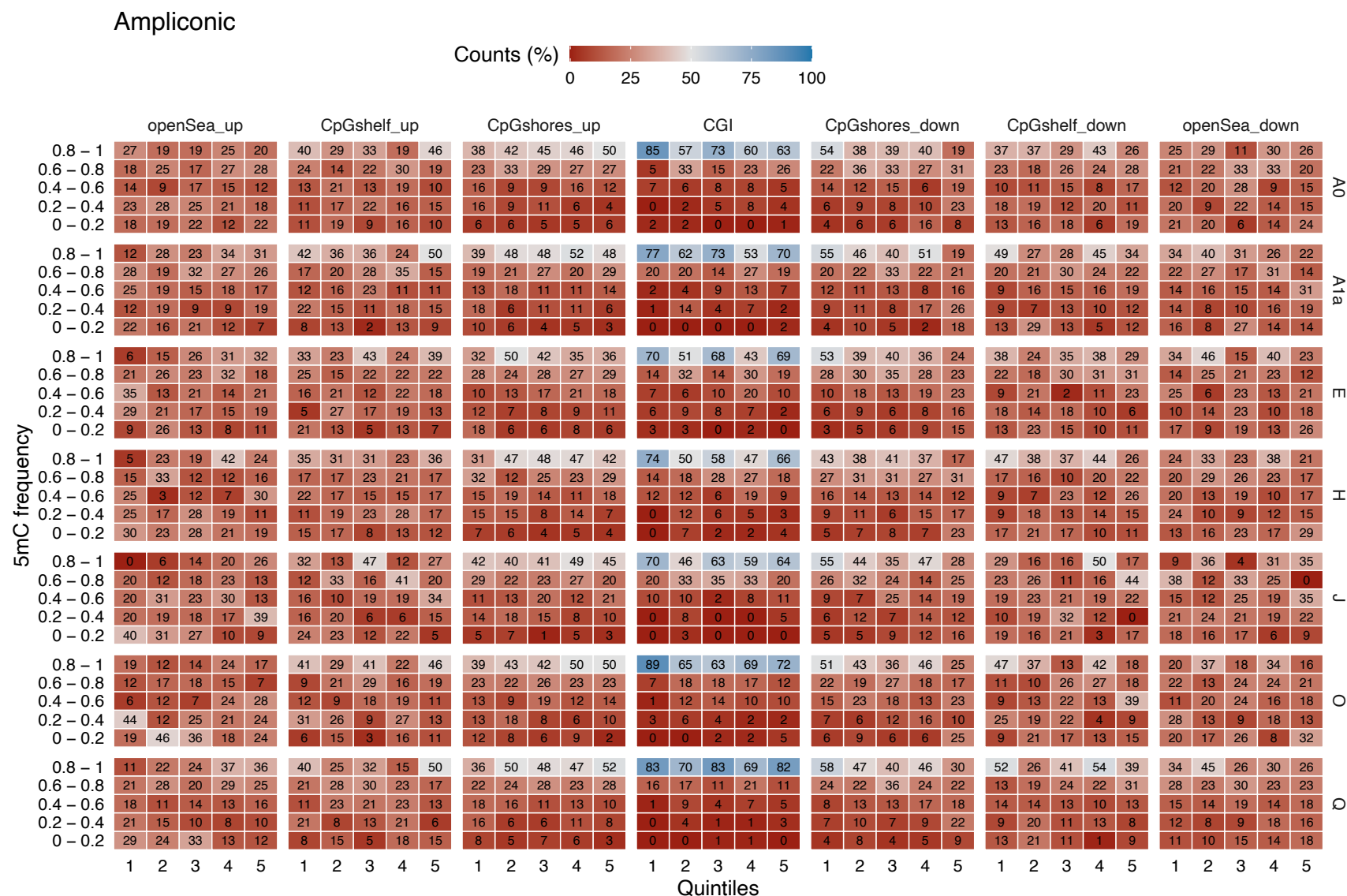
Supplementary Figure 18. Methylation profiles of each haplogroup per chromosome Y sequence classes. (a) 5mC frequency levels, (b) number of CpG in each sequence class used in (a). Vertical black lines in (a) represent 5mC frequency median values.



Supplementary Figure 19. Methylation profiles per chromosome Y sequence classes and CpG annotation: (a) ampliconic, (b) heterochromatin, and (c) other regions. CpG annotations are mutually exclusive regions that comprise: CpG islands (CGI), CpG shores (up to 2kb away from the end of the CGI), CpG shelves (up to 2kb away from the end of the CpG shores), and inter-CpG or open sea regions (where all remaining CpG are allocated). The number of CpG in each annotation is shown within the plot. Vertical black lines represent median 5mC frequency values.

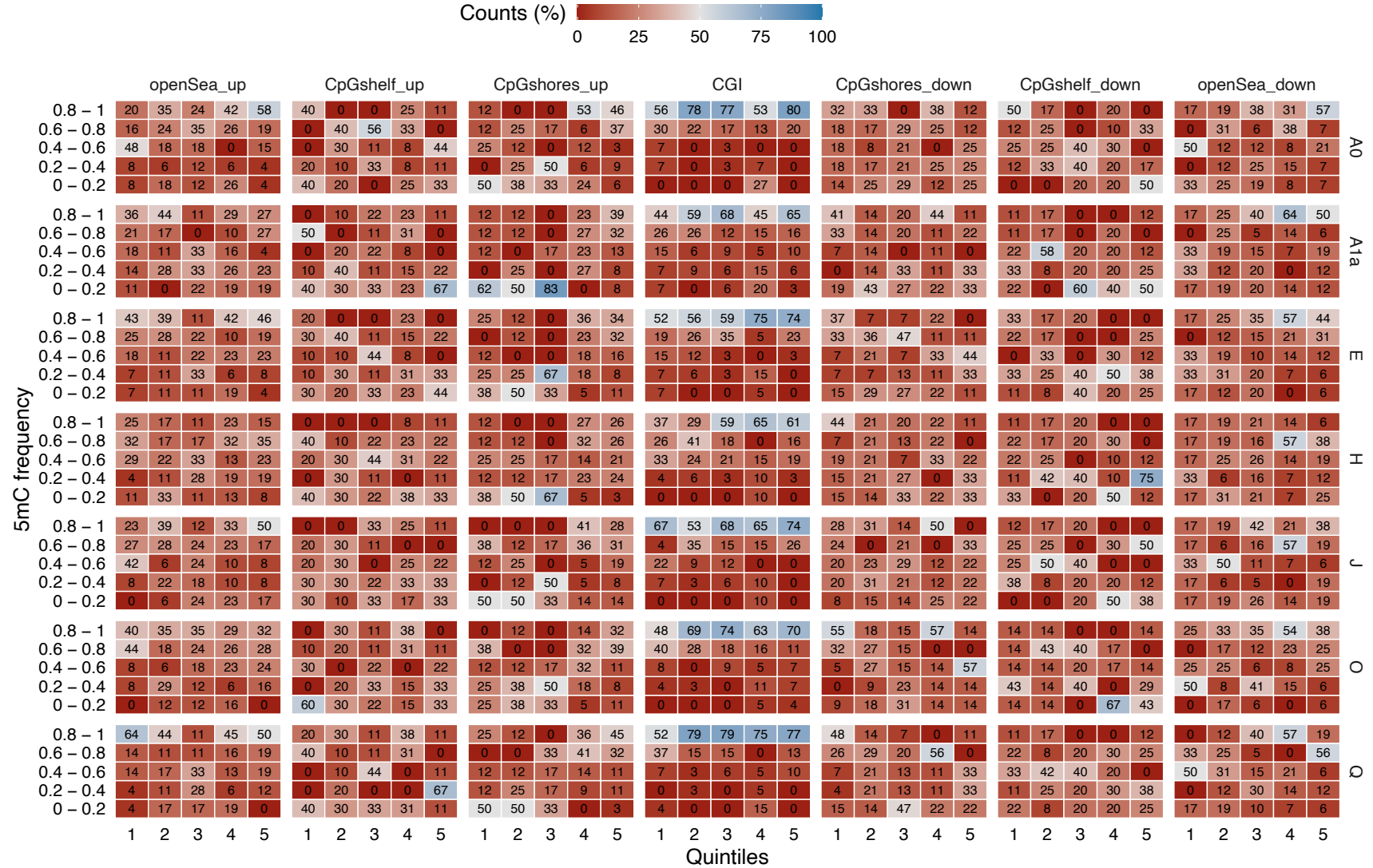


Supplementary Figure 20. Methylation profiles per chromosome Y sequence classes and CpG annotation: (a) pseudo-autosomal, (b) X-degenerate, and (c) X-transposed region. CpG annotations are mutually exclusive regions that comprise: CpG islands (CGI), CpG shores (up to 2kb away from the end of the CGI), CpG shelves (up to 2kb away from the end of the CpG shores), and inter-CpG or open sea regions (where all remaining CpG are allocated). The number of CpG in each annotation is shown within the plot. Vertical black lines represent median 5mC frequency values.



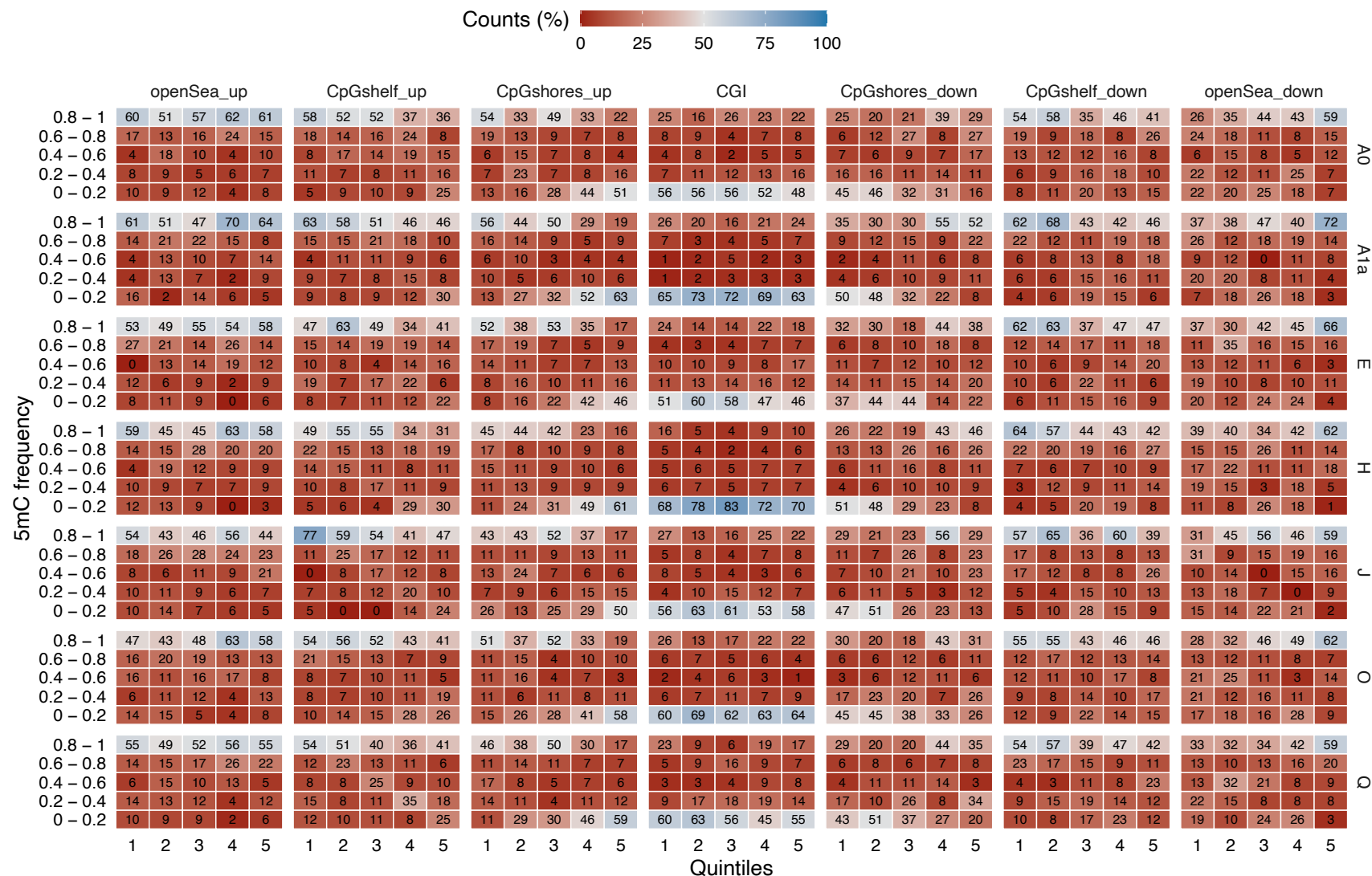
Supplementary Figure 21. Methylation frequency dispersion per CpG annotation in the ampliconic sequence class. For each CpG annotation, tiles show the proportion of CpG with a specific 5mC frequency. For that, CGI were only considered if they presented the full landscape (CpG shores, shelves and open sea) upstream and downstream of the CGI. Open sea was restricted to the 2kb regions closest to the CpG shelves. To account for the progression of methylation in each annotation, each CpG annotation was divided in 5 bins of equal length (quintiles). Pseudo-autosomal, X-transposed and other regions were not represented as they did not have enough CpG sites.

Heterochromatic



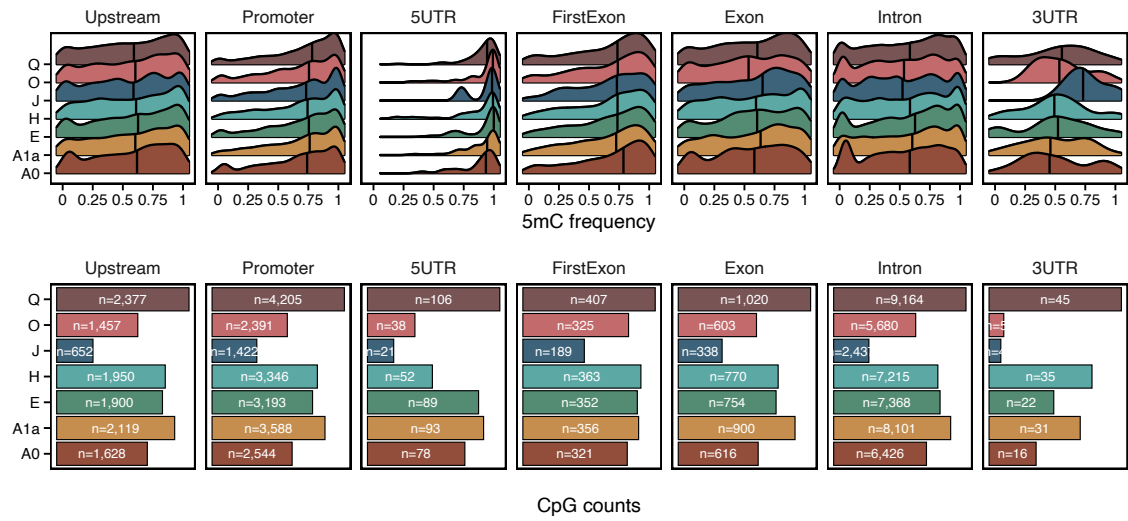
Supplementary Figure 22. Methylation frequency dispersion per CpG annotation in the heterochromatic sequence class. For each CpG annotation, tiles show the proportion of CpG with a specific 5mC frequency. For that, CGI were only considered if they presented the full landscape (CpG shores, shelves and open sea) upstream and downstream of the CGI. Open sea was restricted to the 2kb regions closest to the CpG shelves. To account for the progression of methylation in each annotation, each CpG annotation was divided in 5 bins of equal length (quintiles). Pseudo-autosomal, X-transposed and other regions were not represented as they did not have enough CpG sites.

X-degenerate

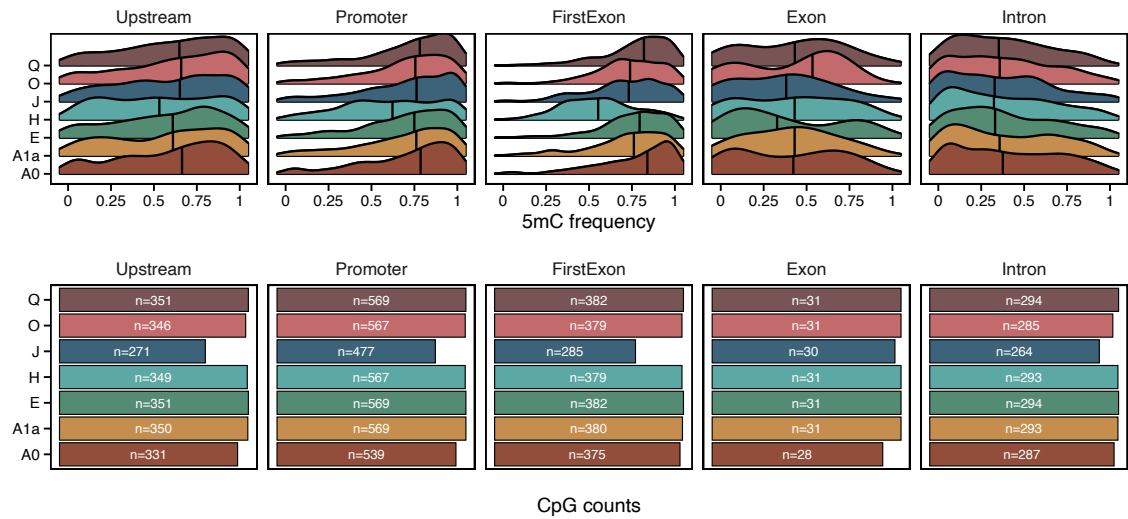


Supplementary Figure 23. Methylation frequency dispersion per CpG annotation in the X-degenerate sequence class. For each CpG annotation, tiles show the proportion of CpG with a specific 5mC frequency. For that, CGI were only considered if they presented the full landscape (CpG shores, shelves and open sea) upstream and downstream of the CGI. Open sea was restricted to the 2kb regions closest to the CpG shelves. To account for the progression of methylation in each annotation, each CpG annotation was divided in 5 bins of equal length (quintiles). Pseudo-autosomal, X-transposed and other regions were not represented as they did not have enough CpG sites.

a Ampliconic

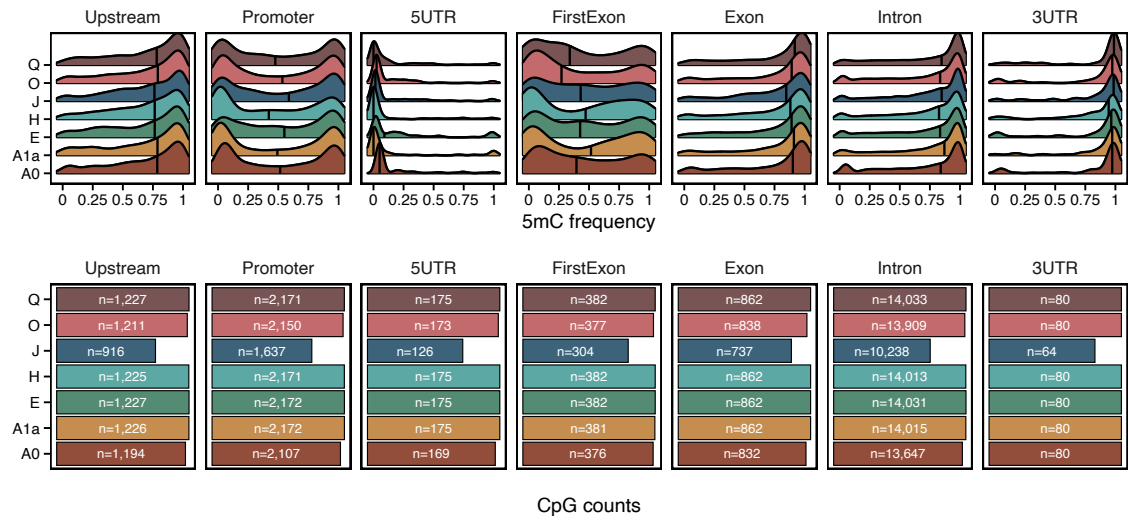


b Heterochromatic

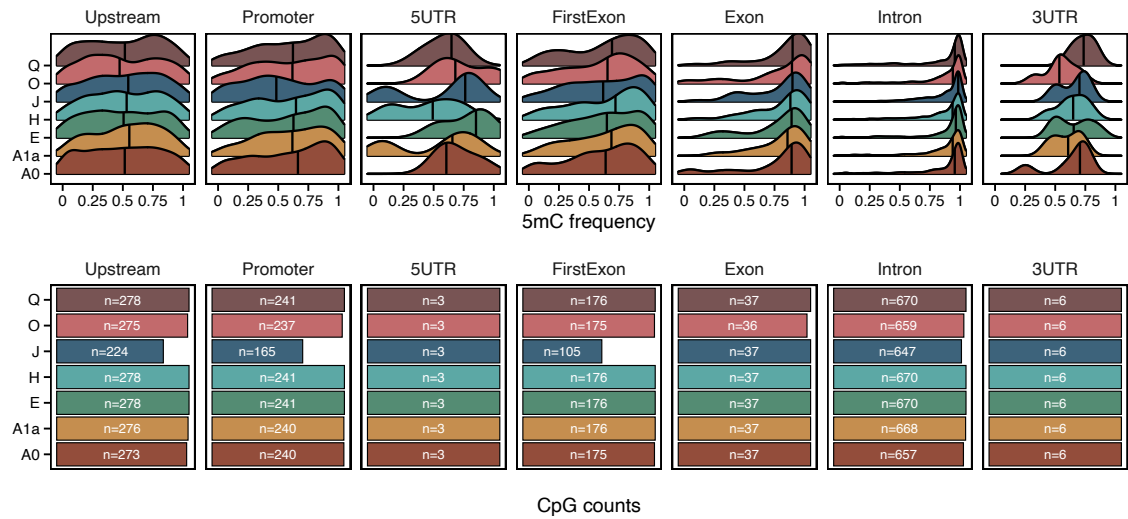


Supplementary Figure 24. Methylation profiles per chromosome Y sequence classes and CpG annotation: (a) ampliconic, and (b) heterochromatic. Gene annotations are mutually exclusive regions that comprise the upstream region (up to 3kb from the end of the promoter), promoter (2kb upstream of TSS), 5'UTR, first exon, exons, introns, 3'UTR. For the overlapping regions, the priority set was the following: promoters, UTRs (5', 3'), first exon, other exons, all introns, and the upstream region. The number of CpG for each gene annotation is shown in the lower plot of each panel. Vertical black lines in each top plot represent median 5mC frequency values. Only chromosome Y sequence classes with most CpG sites were represented.

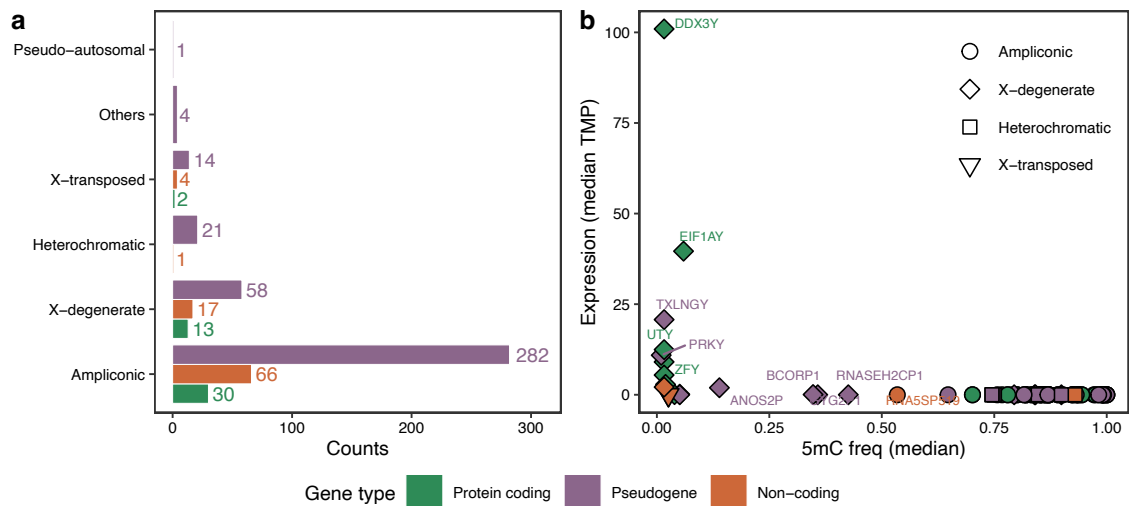
a X-degenerate



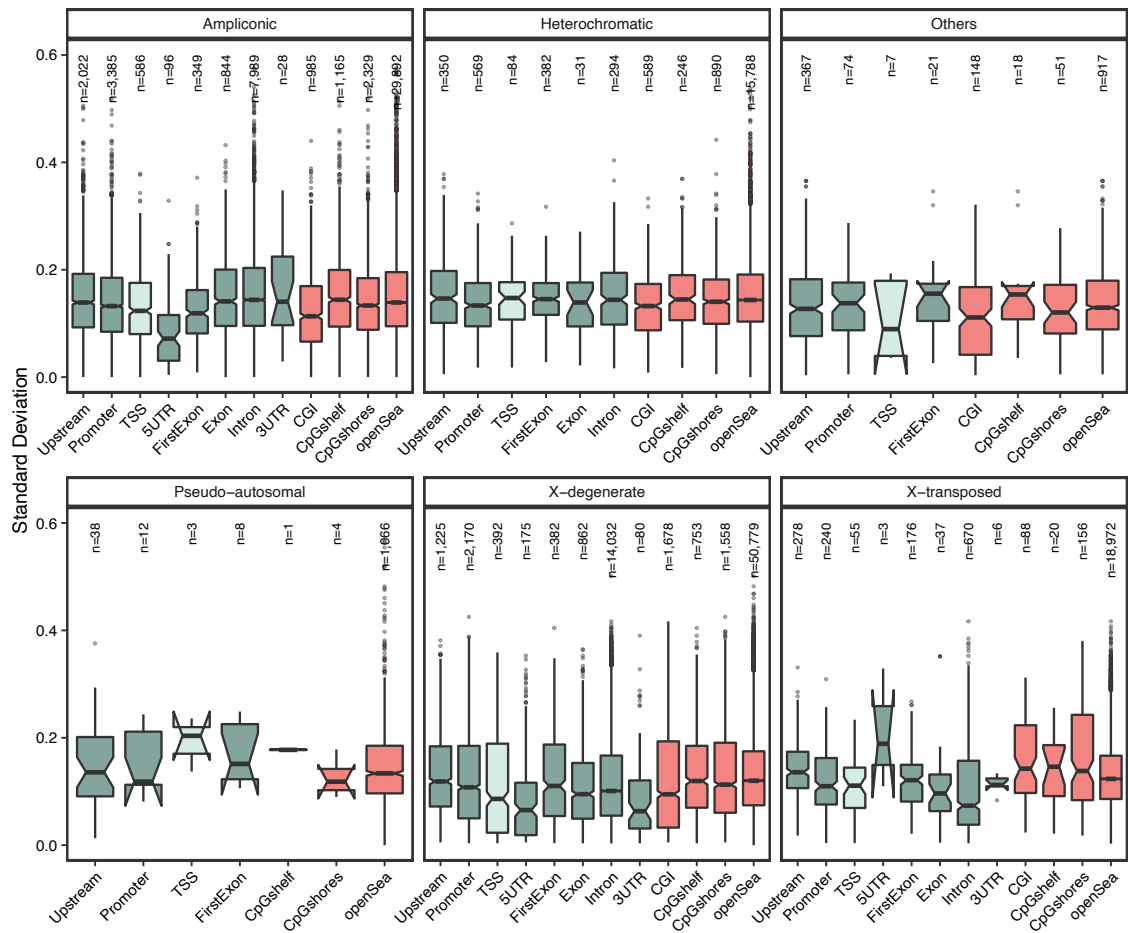
b X-transposed



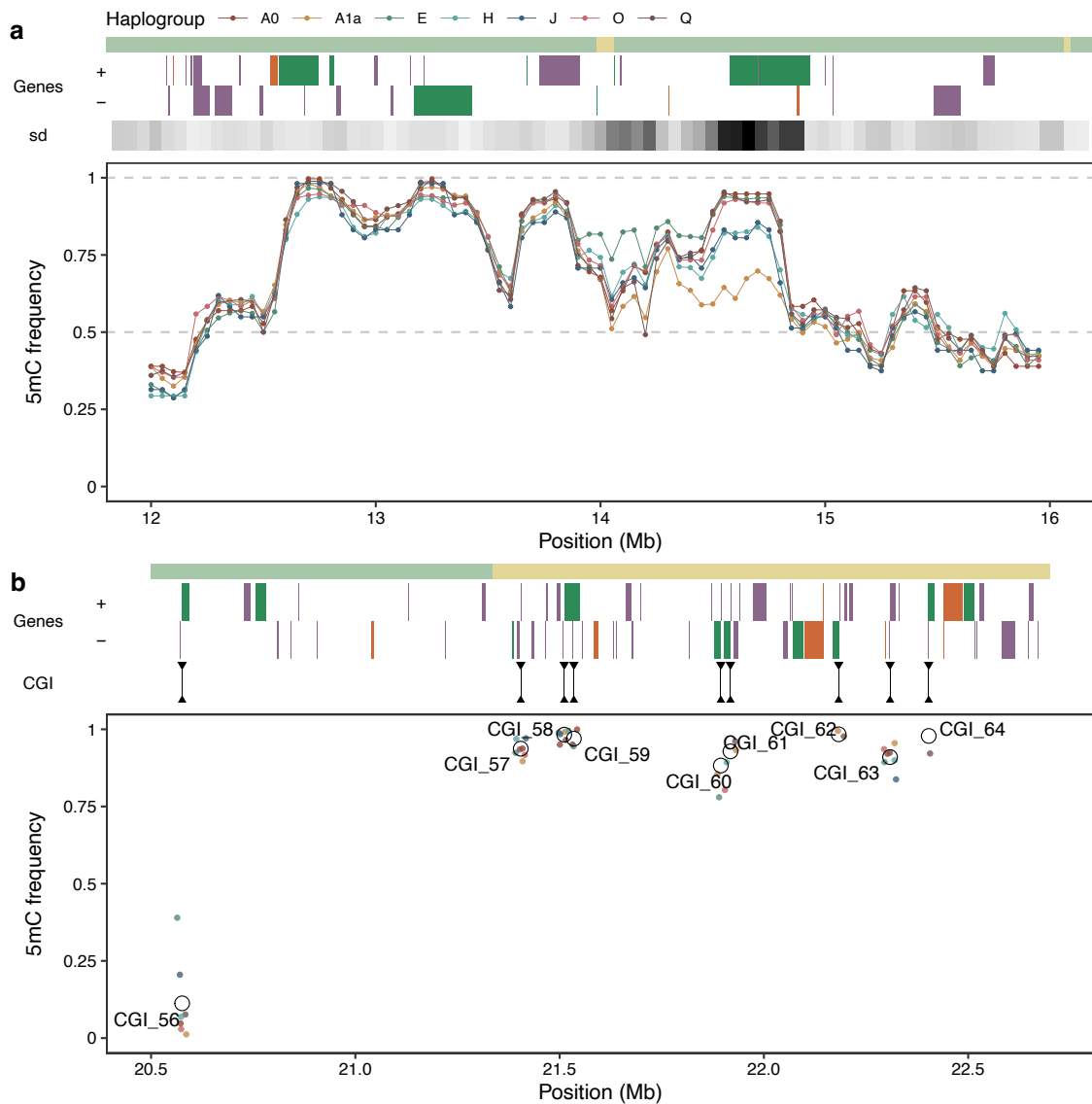
Supplementary Figure 25. Methylation profiles per chromosome Y sequence classes and CpG annotation: (a) X-degenerate, and (b) X-transposed region. Gene annotations are mutually exclusive regions that comprise the upstream region (up to 3kb from the end of the promoter), promoter (2kb upstream of TSS), 5'UTR, first exon, exons, introns, 3'UTR. For the overlapping regions, the priority set was the following: promoters, UTRs (5', 3'), first exon, other exons, all introns, and the upstream region. The number of CpG for each gene annotation is shown in the lower plot of each panel. Vertical black lines in each top plot represent median 5mC frequency values. Only chromosome Y sequence classes with most CpG sites were represented.



Supplementary Figure 26. (a) Number of genes per sequence class in the chromosome Y. The two most gene-dense regions are the ampliconic and X-degenerate regions. (b) Negative association between CGI methylation and gene expression. Each gene is associated with one upstream CGI and the median 5mC frequency is calculated for the 7 samples. Gene expression obtained is from GTEx data (v8) on LCLs derived from male samples and represented as median TMP.



Supplementary Figure 27. CpG 5mC frequency dispersion between cell lines across sequence class and annotation. Dispersion is measured as the standard deviation of 5mC frequencies in CpG sites with methylation information for 3 or more cell lines. Gene annotation is coloured in blue while CpG annotation is coloured in salmon. Gene annotations are mutually exclusive regions as described in Supplementary Figures 24 and 25 in all instances but in TSS, which is a subset of the promoter region comprising the 200bp immediately upstream the TSS. The number of CpG sites used in each category is shown. Box plots show medians and the first and third quartiles (the 25th and 75th percentiles), respectively. The upper and lower whiskers extend the largest and smallest value no further than 1.5xIQR.



Supplementary Figure 28. (a) Frequency of 5mC in the seven cell lines along a specific region of the resolved MSY of the hg38: coordinates from 12 to 16 Mb. The methylation levels are calculated as the median 5mC frequency value in 100kb sliding windows for each cell line. The sequence classes, the genes annotated and the standard deviation (sd) of the methylation levels across cell lines are also shown. The standard deviation of the 5mC frequency (sd) is represented in a white-to-black scale, in which a darker color denotes a higher standard deviation value. The greatest dispersion region which has a higher standard deviation coincides with the protein-coding gene *NLGN4Y* (14.5 to 14.9 Mb). The cell line derived from haplogroup A1a shows lower 5mC values in this region. (b) Methylation frequencies in different CpG islands (CGI) in a specific region of the resolved MSY of the hg38: coordinates from 20.5 to 22.5 Mb. Empty circles denote the mean 5mC frequency per CGI whereas smaller colored points indicate the individual value in each cell line. CGI 56, which shows overall low 5mC frequency values is consistent with the regulation of a neighboring protein-coding gene in the positive strand, *EIF1AY*. On the other hand, CGI in the ampliconic region shows overall high 5mC values, also consistent with the silencing of most genes present in this region, of which the protein-coding genes belong to the *RBMV* gene family.