Y chromosome sequence and epigenomic reconstruction across human populations

Paula Esteller-Cucala*[1], Marc Palmada-Flores*[1], Lukas F. K. Kuderna[1], Claudia Fontsere[1], Aitor Serres-Armero[1], Marc Dabad[2], María Torralvo[1], Armida Faella[1], Luis Ferrández-Peral[1], Laia Llovera[1], Oscar Fornas[3,4], Eva Julià[3], Erika Ramírez[3], Irene González[3], Jochen Hecht[3], Esther Lizano[1,5], David Juan[1], Tomàs Marquès-Bonet[1,2,4-6]

[1] Institut de Biologia Evolutiva, (CSIC-Universitat Pompeu Fabra), Doctor Aiguader 88, Barcelona, Spain

[2] CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, Barcelona, Spain

[3] Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Doctor Aiguader 88, Barcelona, Spain

[4] Universitat Pompeu Fabra (UPF), Doctor Aiguader 88, Barcelona, Spain

[5] Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, Cerdanyola del Vallès, Spain

[6] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain

* These authors contributed equally

Correspondence to: tomas.marques@upf.edu and paula.esteller@upf.edu

Abstract

Recent advances in long-read sequencing technologies have allowed the generation and curation of more complete genome assemblies, enabling the analysis of traditionally neglected chromosomes, such as the human Y chromosome (chrY). Native DNA was sequenced on a MinION Oxford Nanopore Technologies sequencing device to generate genome assemblies for 7 major chrY human haplogroups. We analyzed and compared the chrY enrichment of sequencing data obtained using two different selective sequencing approaches: adaptive sampling and flow cytometry chromosome sorting. We show that adaptive sampling can produce data to create assemblies comparable to chromosome sorting while being a less expensive and time-consuming technique. We also assessed haplogroup-specific structural variants, which would be otherwise difficult to study using short-read sequencing data only. Finally, we took advantage of this technology to detect and profile epigenetic modifications amongst the considered haplogroups. Altogether, we provide a framework to study complex genomic regions with a simple, fast, and affordable methodology that could be applied to larger population genomics datasets.

Introduction

Human sex chromosomes have been traditionally excluded from genome-wide studies[1,2]. This exclusion is particularly pronounced for the Y chromosome, the study of which could be key in understanding differences in disease susceptibility between men and women[3–5]. However, the Y chromosome is now considered important not only for male-specific traits but also for the study and characterization of common complex diseases[4]. Sex-limited chromosomes, defined as those unique to a heterogametic genome[6], are usually harder to assemble since they are haploid and thus have half the sequencing depth when sequenced together with other autosomal and homogametic chromosomes. Moreover, their repetitive nature, filled with ampliconic regions and heterochromatin, poses an additional challenge for assemblers[7].

The first Y chromosome assemblies were generated by means of bacterial artificial chromosomes (BACs) which are labor-intensive and time-consuming approaches[8–10]. Indeed, the Y chromosome sequences in the GRCh38 assembly[11–13] are a composite of BAC clones[11] from a male that belongs to the R1b haplogroup[14] and pseudoautosomal (PAR) regions from the X-chromosome.

To facilitate the assembly process, and also to avoid the use of such costly techniques, one can decrease the potential interchromosomal assembly overlaps by specifically enriching the chromosome of interest. This can be done by physically isolating the chromosome using flow cytometry (chromosome sorting)[6,15–17]. Alternatively, other selective sequencing methods such as adaptive sampling on Oxford Nanopore Technologies (ONT) devices[18] can potentially be used.

Chromosome sorting allows the chromosome of interest to be sequenced on different platforms after its physical isolation by flow cytometry. This separation is possible because different chromosomes have specific fluorescence intensity[19]. On the other hand, adaptive sampling allows for the sequencing of specific DNA regions by locus-specific enrichment or depletion of off-target reads without the need for previous chromosome enrichment[20,21]. To obtain a *de novo* assembly, it is also important to avoid whole genome amplification (WGA), as this process can introduce chimeras, bias the assembly process[22] and prevent the detection of epigenetic modifications.

Long-read whole-genome sequencing enables the assessment of previously unsolved repeats, and thus allows generating more contiguous assemblies. Currently, ONT can achieve the longest read lengths compared to any other existing sequencing technology[23–25]. Moreover, ONT allows the detection of DNA (and RNA) modifications based on the different current signals of the nanopores[26,27]. Taken together, this technology is able to resolve gaps, allowing for the true completion of chromosomes or even genomes[28–30]. Here we assess the performance of two enrichment methods to

77 sequence and assemble the Y chromosomes from 7 major human haplogroups.
78 Moreover, we provide insights into their structural variation and epigenomic landscape
79 showing that enrichment techniques coupled with ONT can be used to study variation
80 between population datasets.

81 **Results**

82 Data production

83 Complete Y chromosomes from 6 different human haplogroups were isolated as
84 previously described[17] (Fig. 1A). In brief, chromosomes were obtained from
85 lymphoblastoid cell lines (LCLs) used in the 1000 Genomes Project[31] (1kgp) and
86 sequenced on the ONT MinION. We also made use of the Y chromosome sorted ONT data
87 generated by Kuderna et al.[17], whose haplogroup (A0) represents one of the deepest-
88 rooting known haplogroups. Additionally, we also generated Illumina short-read data for
89 the same flow-sorted chromosomes (Supplementary Table 1).

90 The ONT data available for the cell lines ranged from 6.4 to 10.33 Gb, of which 7 to 33%
91 mapped to the Y chromosome in the reference. Moreover, we also generated 6.4 to 35
92 Gb of Illumina data for all the chromosome sorting extractions. This is a notably high
93 amount of data, especially considering that the Y chromosome sequence represents less
94 than 1% of the known sequence in GRCh38.

95 The Y chromosome enrichment specificity was assessed by aligning the basecalled data
96 to the human reference genome assembly GRCh38 and calculating the normalized
97 coverage on each chromosome accounting for the gaps of the reference genome and the
98 ploidy of each chromosome (Methods, Fig. 1B and Supplementary Figs. 1 and 2). The Y
99 chromosome-specific enrichment factor of the six samples showed high variability, as it
100 ranged from 15 to 50-fold, whereas the A0 haplogroup was over 100-fold enriched
101 (Supplementary Table 2). As noted in Kuderna et al.[17], we found that chromosome 22
102 partially co-sorts with chromosome Y, showing enrichment values slightly higher than 1.

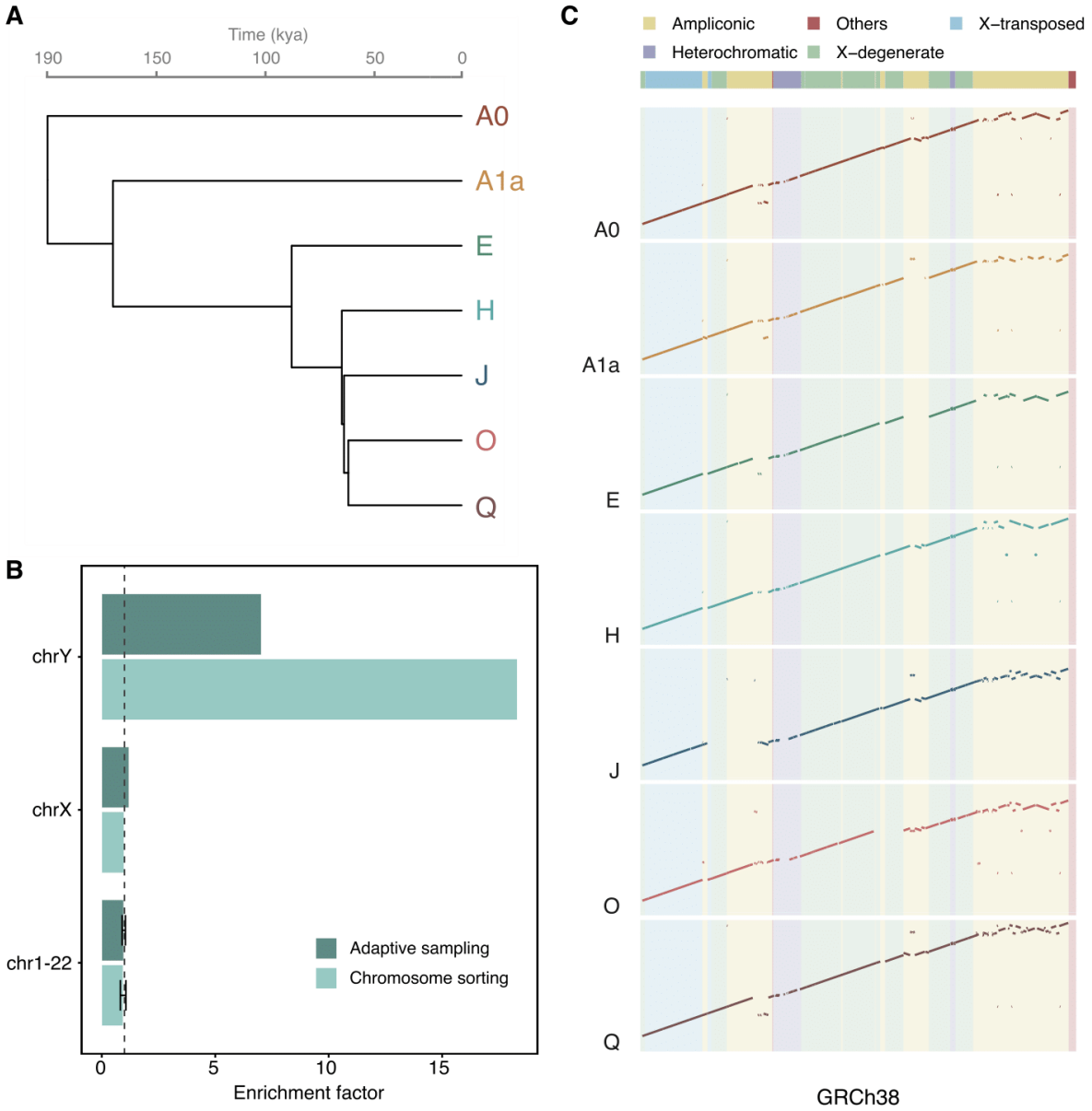

Figure 1. Study design, enrichments, and assemblies. (A) Phylogenetic tree of the human Y chromosomes used in the study. Split times taken from Jobling & Tyler-Smith, 2017[32]. kya, kilo years ago. (B) Enrichment factor values of the H haplogroup from data generated using chromosome sorting and adaptive sampling. The chrY shows higher enrichment with chromosome sorting than with adaptive sampling for the haplogroup compared. The dashed vertical line equal to 1 denotes no chromosomal enrichment. (C) Dot-plots of the manually scaffolded Y chromosomes compared to the resolved MSY region of GRCh38. The large-scale deletion in the J haplogroup is most likely due to its low coverage.

Adaptive sampling as a strategy to enrich specific chromosomes

A limiting factor in chromosome sorting is the need to culture hundreds of millions of cells in order to enrich the chromosome of interest effectively[16,17]. To overcome this limitation, we explored the potential of adaptive sampling to specifically enrich the Y chromosome. This approach was done for one of the cell lines (haplogroup H) for which

117 chromosome sorting data had also been generated. We used the nucleotide sequences
118 of the Y chromosome (chrY) and the contig *chrY_KI270740v1_random* (chrY_random) as
119 provided in the GRCh38 assembly as the target sequences to enrich. To obtain
120 comparable coverages using the two methodologies (~18x), we ran two ONT MinION
121 flowcells with adaptive sampling. In both experiments, we showed that the Y
122 chromosome was preferentially enriched to the other chromosomes (Supplementary Fig.
123 3). Although the Y chromosome enrichment factor value with chromosome sorting
124 doubles the one in adaptive sampling for this cell line (Fig. 1B and Supplementary Table
125 2), adaptive sampling proves to be a cheaper and less time-consuming strategy.

126 Y chromosome assembly across haplogroups and enrichment techniques

127 We obtained Y chromosome assemblies corresponding to 7 different Y chromosome
128 haplogroups using chromosome sorting on distinct cell lines (Methods, Supplementary
129 Figs. 4 to 6 and Supplementary Table 3). The coverage used by the assembler to generate
130 each assembly ranged from 13 to 50x, with a mean assembly coverage of over 28x. The
131 resulting assemblies spanned from 18.95 - 22.23 Mb in length, being 16 to 28% shorter
132 than the length of chromosome Y in GRCh38. We also observed that assemblies with
133 higher continuity (contig N50) tend to have higher values of read length N50 and mean
134 read lengths (Supplementary Table 3).

135 Our assemblies had a similar amount of contigs compared to a previously published
136 African Y chromosome assembly (haplogroup A0)[17], which is lower than the number of
137 contigs of the GRCh38. The N50 across our assemblies ranged from 1.40 - 2.67 Mb, and
138 are thus within the same order of magnitude as the Y chromosome in GRCh38 (6.9 Mb).
139 These results suggest that creating *de novo* assemblies primarily based on long reads
140 mapping to a reference chromosomal assembly might lead to shorter assemblies, less
141 fragmented but with lower continuity values (such as lower contig N50). However, these
142 results also show that by using ONT and Illumina platforms it is possible to generate
143 assemblies almost as continuous as the current GRCh38 reference, for which much more
144 effort and resources were devoted[11–13,33].

145 Furthermore, we manually scaffolded each haplogroup to create a single scaffold based
146 on genome-to-genome alignments to the GRCh38 Y chromosome (Fig. 1C). Comparing
147 our assemblies to the GRCh38 in the male-specific region of the Y chromosome (MSY)
148 shows how most of the MSY sequence classes were assembled for the most part. As
149 previously reported[17], the ampliconic region was the most fragmented and the least
150 complete, most likely due to collapsed repeats. Of note, for the J haplogroup, we observed
151 a big gap in the region comprising 6.7 - 9.3 Mb of the Y chromosome. This region includes
152 an X-degenerate region and most of its adjacent ampliconic region. When inspecting the
153 reads mapping to this region, we found few reads present, thus possibly explaining why
154 we could not accurately assemble this region.

155 Compared to the previously assembled African chrY that also made use of chromosome
156 sorting data, we were able to generate a longer and more contiguous assembly starting
157 from the same raw fast5 reads (Supplementary Table 3). This demonstrates the value of
158 combining up-to-date basecalling and assembly tools, which are constantly evolving for
159 long-read data[34–36].

160 Apart from chromosome sorting (CS), data from the H haplogroup (GM21113 cell line)
161 was also generated using adaptive sampling (AS). In order to generate and compare the
162 assemblies between the two enrichment methods, and given the unequal amount of data
163 generated between them (83.5Mb difference), we restricted the comparison to
164 assemblies generated using the same number of bases (Methods). The resulting
165 assemblies showed similar values in metrics such as genome span (CS: 21.8 Mb, AS: 22.0
166 Mb), contig N50 Mb (CS: 2.7 Mb, AS: 2.6 Mb), and L50 (both L50 = 3 scaffolds). Moreover,
167 the AS-based assembly led to a slightly more fragmented assembly (44 sequences)
168 compared to the CS-based one (31 sequences) (Table 1 and Supplementary Fig. 7).

169 **Table 1.** Assembly metrics of Y chromosome assemblies for the GM21113 cell line (haplogroup H)
170 using adaptive sampling and chromosome sorting, and an assembly using all the data available
171 for GM21113.

Selective sequencing method to generate the long-read data	Assembly span (bp)	Scaffold N50 (bp)	Scaffold L50	Number of sequences
Adaptive sampling (AS)	21,955,745	2,612,207	3	44
Chromosome sorting (CS)	21,794,102	2,666,112	3	31
AS + CS	22,007,578	2,640,901	3	42

172

173 Altogether, we have generated assemblies for 7 Y chromosome haplogroups with similar
174 contiguity to previously published assemblies. Moreover, we also show that adaptive
175 sampling can be used for generating assemblies that are comparable to those generated
176 by chromosome sorting.

177 <u>The landscape of structural variants across the human Y chromosome phylogeny</u>

178 As expected by the nature of these data, methods to detect structural variants which
179 make use of long reads show an overall better performance than methods based on
180 short-read data[37]. Taking advantage of our data, we assessed the landscape of structural
181 variants in the Y chromosome in the 7 haplogroups. For that, we used two approaches:
182 one based on long-read mapping (*Sniffles*[38,39]) and another based on assembly
183 comparison (*Assemblytics*[40]).

184 First, we identified different structural variants based on how the reads align to a
185 reference genome using *Sniffles*. We used chrY and the chrY_random sequence from the
186 GRCh38 as the reference. After merging the indel calls (Methods), we identified 803
187 unique variants (801 indels), including 194 structural variants (at least 50 bp in size,
188 Supplementary Table 4). The number of structural variants ranged from 103 to 536 events
189 per haplogroup. Moreover, *Sniffles* detected 1 translocation in the H haplogroup and 1
190 duplication event shared between 5 haplogroups (all but A0 and A1a, which are basal
191 relative to the others). The detected duplication is located in the position *chrY:56,673,215*,
192 at the end of a gap. This indicates that the reference is missing a region of around 98,295
193 bp, similar to the sequence after the gap. Most of the events were indels of 10 to 50 bp
194 in size (Fig. 2A). Out of the 803 variants found, there were 320 insertions and 481 deletions
195 (including 3 insertions and 9 deletions from chrY_random). We manually investigated the
196 longest events detected using *Sniffles* and confirmed the longest insertion of 6,023 bp and
197 the longest deletion of 6,314 bp. Both were found in haplogroup A0 and belonged to
198 different X-degenerate regions (Supplementary Figs. 8 and 9). After merging and
199 regenotyping the panel of indels, we observed that the cell line belonging to haplogroup
200 J was the one having more undetermined genotypes (ie. positions with no genotype
201 information). This correlates with a lower sequencing depth for this sample. We also
202 observed that 14% of the variants genotyped in all haplogroups shared the same
203 genotype which was different from the reference (Fig. 2B). Haplogroups A0 and A1a
204 harbour the most haplogroup-specific variants, concordant to their genetic distance to
205 the reference. We also manually assessed previously reported events for the A0
206 haplogroup[17] and confirmed that they were restricted to this cell line (Supplementary Figs
207 10 and 11). This indicates that structural variants found in only one haplogroup might not
208 be representative of widespread structural variants of a chromosome, but rather
209 delimited to one specific population or group of individuals.

210 Second, we identified structural variants based on the comparison of the obtained chrY
211 assemblies to the reference chrY GRCh38 using*Assemblytics v1.2.1*[40] (Supplementary Table
212 5). This allowed for the detection of 557 to 1,019 putative variants, for which 202 to 258
213 were at least 50 bp in size (Fig. 2A). We also found between 1 to 4 structural variants
214 bigger than 50,000 bp for the cell lines studied, a type of variant that the mapping-based
215 method may not detect because it would require constant coverage along a long region
216 of the reference.

217 We observe 194 to 406 insertions per haplogroup with *Assemblytics*[40] compared to the 45
218 to 255 insertions detected with *Sniffles*[38,39], and 206 to 424 deletions against 57 to 288,
219 respectively. However, similar amounts of structural variant indels are detected by *Sniffles*
220 in all haplogroups (between 52 and 110) but for the J haplogroup (n = 21 variants)
221 compared to *Assemblytics* (between 47 and 80, J haplogroup having 55 variants). These
222 results, together with the fact that J haplogroup is the one with less data generated,

223 suggest that mapping-based structural variation detection methods may not be able to
224 detect as many structural variants compared to assembly-to-assembly comparison-
225 based methods when having limited sequencing depth. In that situation, generating a *de*
226 *novo* assembly and using *Assemblytics* can lead to the identification of larger indels.
227 Moreover, *Assemblytics* provides many other structural variant events such as tandem
228 and repeat expansions or contractions, while *Sniffles* was able only to capture one
229 duplication event and one translocation.

230 We further assess if the indels found using long reads could be similarly genotyped using
231 short-read data. For that, we genotyped the variants confidently called by the ONT
232 mapping-based approach in Illumina data generated for the same cell line extractions.
233 With the Illumina data, we were able to replicate over one-quarter of the indels found in
234 the nanopore data (214 out of the 801 indels). A significant positive association was seen
235 between the predicted genotypes using ONT to those observed using Illumina data for
236 each cell line (Fig. 2C). The observed phi coefficients (correlation values for binary
237 variables) range between 0.35 and 0.54, not close to the highest correlation value of
238 one[41,42]. This is expected, given that most of the variants in our panel cannot be called by
239 the Illumina data. The intra-haplogroup correlation is generally higher than that inter-
240 haplogroup. However, when testing inter-haplogroup associations for genotypes derived
241 from the same methodology, strong associations are also detected (Supplementary Fig.
242 12). To explore which of the indels could be observed in a panel of human variation we
243 genotyped the same indels in the male samples present in the 1000 Genomes Project
244 (Methods). As expected, intra-haplogroup associations were typically positive and
245 significant, generally having stronger associations than inter-haplogroups comparisons
246 (Supplementary Fig. 12).

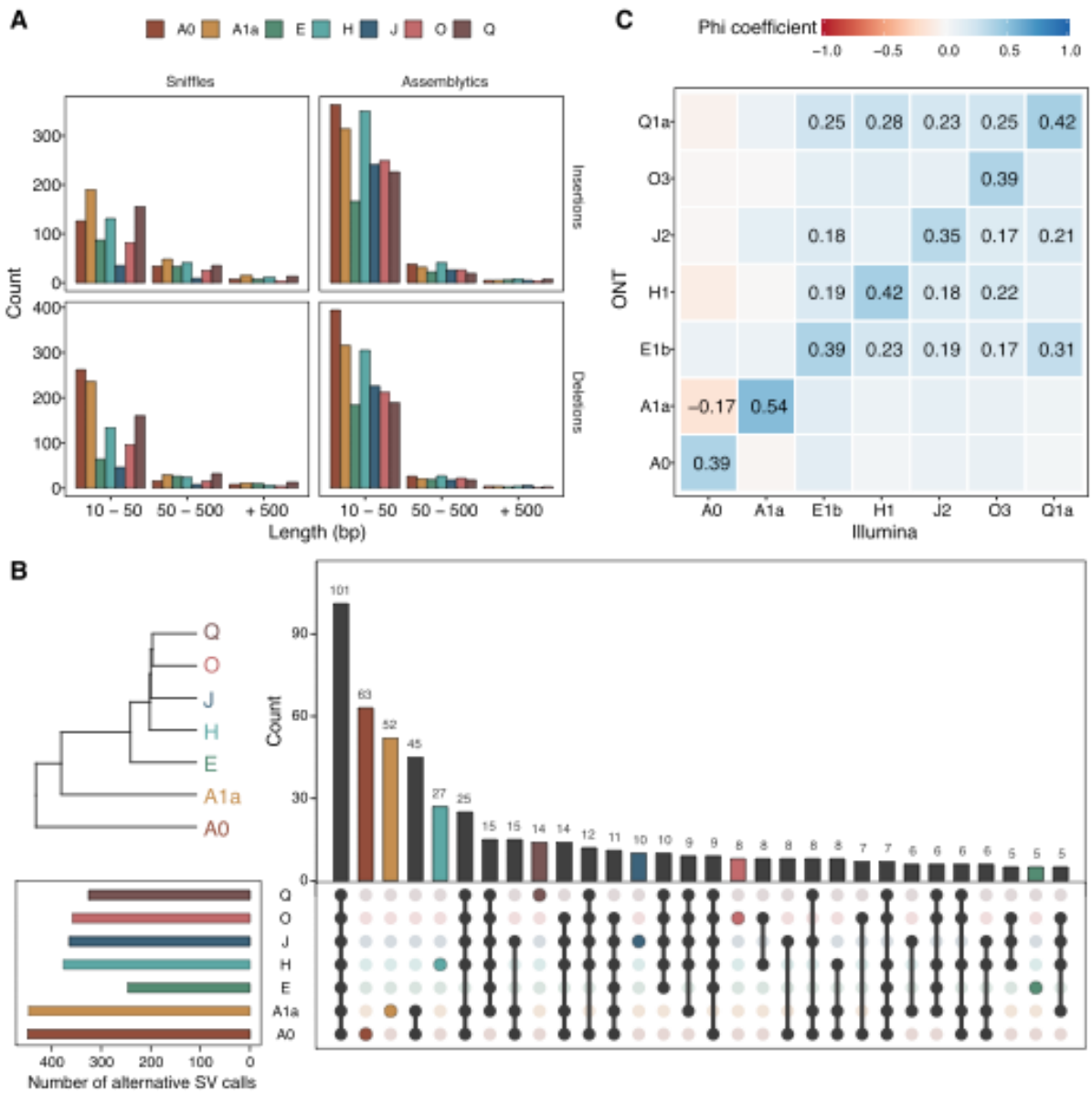

Figure 2. Profiling of structural variants. (A) Lengths of the structural variant events, insertions, and deletions called by *Sniffles* or *Assemblytics* for the different haplogroups. Variants are grouped into three categories depending on their length: from 10 up to 50 bp, from 50 up to 500 bp, and equal to or over 500 bp. (B) Overlap on the alternative calls between haplogroups. As expected by their evolutionary distance, haplogroups A0 and A1a show higher haplogroup-specific variants. Only variants with genotype calls for all haplogroups have been included (n = 726 variants). (C) Correlations between genotype calls using *Sniffles* (ONT-based) or *graphtyper* (Illumina-based) when calling the same set of structural variants. Phi coefficients range from 1 to -1, where 1 indicates complete association. Only correlation values that are statistically significant (p-value < 0.05) after Bonferroni multiple testing correction are shown.CpG methylation across the phylogeny.

CpG methylation across the phylogeny

ONT sequencing relies on the identification of different current signals when the DNA passes through the pore, so it is possible to go beyond the identification of the four canonical nucleotide bases and detect other modifications in the DNA. We used *nanopolish v1.12*[43] to call the methylation status of 5-methylcytosines (5mC) at CpG positions from the nanopore current signal. Assessing the Y chromosome methylome using long reads is beneficial for exploring regions that are traditionally inaccessible using short-read techniques, such as the PAR, X-transposed regions, and even the ampliconic regions.

For that, we performed quantile normalization on the methylation values across samples with a minimum coverage of 4x (Supplementary Fig. 13). We observed consistent methylation patterns along Y chromosomes across samples, indicating a strong overall correlation on the methylation status (Fig. 3A and Supplementary Figs. 14 and 15). However, 5mC frequency values could not recapitulate the expected phylogeny, either chromosome-wise or segregating by sequence class or epigenetic annotation (Supplementary Fig. 16). Given that methylation levels might vary within the population, age, environmental exposures, and cell culture conditions[44–46], and the absence of replicates for each of the haplogroups considered, this observation could be due to differences in any of these variables. However, given the uncertainties about the cell lines and age of the individuals from which were generated, we are unable to discern the 5mC variation which accounts for the different haplogroups from that which could be caused by other factors. As expected by the nature of the sequence classes[11], the X-degenerate region, which harbours single-copy genes and mostly ubiquitous expression, showed 5mC frequency values which resembled the most those normally seen in mammalian autosomal chromosomes[47] (Supplementary Fig. 17). X-degenerate regions showed the characteristic bimodal distribution of frequency values with a median close to 0.7, whereas all other regions showed less defined distributions. We also inspected the behavior of methylation according to the epigenetic annotation of the CpG of each of the sequence classes. For that, we divided the CpGs into four mutually exclusive categories (Fig. 3B and Supplementary Fig. 18): those in CpG islands (CGI), CpG shores, CpG shelves, and other inter-CGI regions (open sea). CGI in the X-degenerate and X-transposed regions were predominantly unmethylated, while all the other regions were mostly methylated. Open sea regions showed intermediate methylation levels for all sequence classes but the X-degenerate, whose median 5mC frequency reached 0.75. As expected by the dynamic nature of the human methylome, CpG shores and shelves showed intermediate values transitioning from CGI and open sea regions (Supplementary Fig. 19)[48–50].

296 DNA methylation is associated with gene expression[51], and so we also inspected the 5mC
297 frequency patterns across different gene annotations (Supplementary Fig. 20). Most
298 annotated genes are present in the X-degenerate and ampliconic sequence classes
299 (Supplementary Fig. 21), and consistent with the different expression profiles of the genes
300 in LCLs (retrieved from GTEx[52]) in each of these two sequence classes, we observed clear
301 distinct methylation patterns in their TSS, UTRs, and intragenic CpGs (Fig. 3C). Not
302 surprisingly, we found 5'UTRs to be the most constrained gene feature across samples,
303 which directly link its methylation status to gene expression (Supplementary Fig. 22).
304 Moreover, we found a direct relationship between upstream CGI methylation status with
305 gene expression (Supplementary Fig. 21). Finally, we explored those cases in which
306 differential methylation could have an effect on gene expression. We encountered a
307 region with high methylation dispersion fully spanning a protein-coding gene (Fig. 3D and
308 Supplementary Fig. 23). In that location, haplogroup A1a was found to be
309 undermethylated compared to the other haplogroups, and though this difference was
310 only modest, it could have an effect on the expression level of the gene located in this
311 region. This gene is *NLGN4Y*, which is a long gene that spans over 300 kb and is expressed
312 in brain and other tissues, including LCLs (τ_{NLGN4Y} = 0.714). Interestingly, this gene has been
313 proposed as a candidate for autism spectrum disorder[53,54]. As expected, we found CGIs
314 located upstream of this gene to be unmethylated (CGI_1 and CGI_2), whereas a CGI
315 potentially regulating an overlapping non-coding gene in the opposite strand and which
316 has no expression in LCLs was shown to be fully methylated in all cell lines (CGI_3).

317 Altogether, we show that ONT can be used to study 5mC across different cell lines, and it
318 can prove to be helpful for the study of traditionally challenging genomic regions,
319 particularly those present in the Y chromosome.

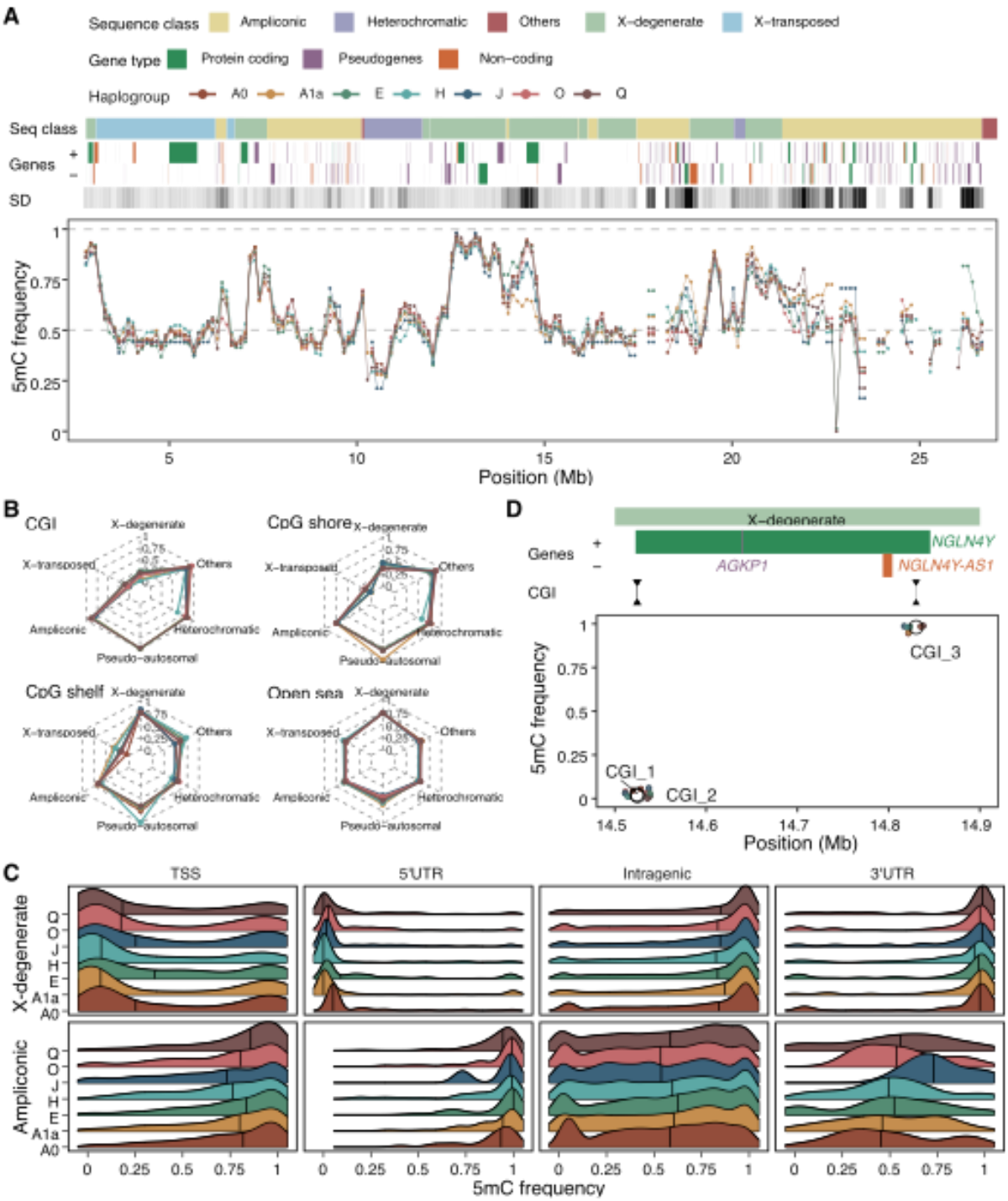

Figure 3. Methylation landscape across the Y chromosome phylogeny. (A) Frequency of 5mC in the seven cell lines along the resolved MSY of the GRCh38. The methylation levels are calculated as the median 5mC frequency value in 250kb sliding windows for each cell line. The sequence classes, the genes annotated and the standard deviation 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. (B) Median methylation value per cell line segregated by CpG annotation and sequence classes. 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-CGI or open sea regions (where all remaining CpG are allocated). (C) 5mC frequencies on different gene

331 features in X-degenerate and ampliconic sequence classes. Gene annotation features shown are
332 TSS (region of 200 bp upstream of the transcription start site), both UTRs, and intragenic regions
333 (which combine all exonic and intronic regions without considering the first gene exon). (D)
334 Methylation frequencies in 3 CpG islands (CGI) surrounding the *NLGN4Y* and *NLGN4Y-AS1* genes.
335 Empty circles show the mean 5mC frequency per CGI, whereas smaller colored points indicate the
336 individual value in each cell line.

337 Discussion

338 Here, we present a panel of ONT data for 7 cell lines that represent the major human Y
339 chromosome haplogroups. We have generated assemblies for each of them and studied
340 their diversity, focusing on structural variation and methylation. To generate this
341 resource, chromosome sorting data was employed and compared to adaptive sampling
342 data, an enrichment technique that is compatible with ONT sequencing data. After
343 generating and comparing the assemblies of the two enrichment techniques, we showed
344 that both methods can lead to comparable assemblies, while they require different time,
345 cost, and expertise. In terms of enrichment factor values, chromosome sorting shows co-
346 enrichment with chromosome 22. This is mainly due to the fact that both chromosomes
347 have similar sizes. However, this is not seen in adaptive sampling. In fact, samples
348 enriched with adaptive sampling show the lowest standard deviation of the enrichment
349 factor on autosomal chromosomes. Nevertheless, given the homology of the sex
350 chromosomes, and the fact that adaptive sampling is performed by providing the
351 genomic sequence of chromosome Y, chromosome X shows higher enrichment factor
352 compared to the other chromosomes and the samples enriched by chromosome sorting.
353 Altogether we show that adaptive sampling is a viable alternative strategy for the
354 enrichment of specific genomic regions. We also emphasize the importance of using high
355 molecular weight DNA or long DNA fragments, which are especially convenient for the
356 enrichment of small chromosomes with adaptive sampling. As such, at longer DNA
357 fragment sizes, the time the sequencer will be scanning for on-target regions (i.e., those
358 that belong to the Y chromosome) will be reduced. Therefore, we realize that having
359 started from higher DNA fragment sizes for the haplogroup H sample would have led to
360 higher enrichment efficiencies in the adaptive sampling enrichment method.

361 One major limitation of our work is the conservative filtering we have used to generate
362 the assemblies. Our approach uses the Y chromosome of the current genome of
363 reference GRCh38 as a backbone. All data obtained using adaptive sampling relies heavily
364 on the GRCh38 reference and may include a few reads of other chromosomes that start
365 with a similar sequence. On the other side, chromosome sorting produces data on
366 unresolved chromosomal regions but includes some undesired full chromosomes data.
367 As such, restricting our assembly to only those reads that map to the reference leads to
368 the loss of a fraction of Y chromosome potentially informative reads during the filtering
369 process. Conversely, this approach minimizes the retention of non-chromosome Y data

370 and limits the resulting assembly to the Y chromosome only. Compared to a previous
371 assembly created with the same data for the cell line that belongs to haplogroup A0, our
372 approach yielded a more contiguous assembly. As such, it shows the potential that re-
373 processing the same raw data with novel approaches might have in the future, especially
374 in the context of the big data era[55,56].

375 Due to its large fraction of heterochromatin, around half of the sequence in the current
376 Y chromosome assembly is unresolved. This limitation, together with the fact that we are
377 using a partial reference genome to generate assemblies of a specific chromosome,
378 hampers the possibility of reconstructing the totality of this chromosome. In the future,
379 telomere-to-telomere Y chromosome sequencing would undoubtedly avoid reference-
380 biases we encountered in this study[28].

381 We also took advantage of the long-read data generated to explore the landscape of
382 structural variants in each cell line. For that, we used two different methods for structural
383 variant calling: one based on long-read mapping and another based on assembly
384 comparison. The former allows for two rounds of genotyping and so the final candidates
385 are potentially more curated. The latter is based on genome-to-genome comparisons, so
386 it is able to detect longer genomics variants. We consider that for low data samples the
387 creation of a *de novo* Y chromosome assembly may allow the detection of structural
388 variants that cannot be recognized with a mapping-based method, considering the low
389 coverage of reads mapping in those regions.

390 Besides the potential to generate high-accuracy assemblies and resolve complex genomic
391 regions like structural variants, ONT also allows for studying the epigenome. We have
392 assessed the methylation status of cytosines in a CpG context for our panel of cell lines.
393 Despite the fact that the epigenome of the Y chromosome has not been deeply studied,
394 we were able to consistently replicate the methylation patterns that have been described
395 in other human autosomal chromosomes[57,58]. Not surprisingly, with the methylation
396 values obtained, we were unable to recapitulate our samples' expected phylogeny. Two
397 main factors can be attributed to this: the lack of replicates for each haplogroup and also
398 within-population variability[45,46,59,60] which, in our case, could also be confounded by
399 epigenetic drift[61]. Still, methylation differences at the population level are to be expected
400 to be small in magnitude[45]. In this line, we were able to detect small differences in
401 methylation in regions that could have an effect on the regulation of specific genes. This
402 is the case of gene *NYGN4Y*, which we found to fully overlap with a region with consistently
403 lower methylation in the cell line belonging to haplogroup A1a.

404 Nevertheless, we are using lymphoblastoid cell lines (LCLs), which are artificially
405 transformed cells, so caution must be taken when extrapolating these findings. But the
406 extent to which the generalization of our results could be biased is even more
407 consequential when reporting those findings that are sample-specific. As such, an

408 increase in the number of replicates would help to discern which of our findings are
409 artifacts from those which have a true biological meaning.

410 Taken together, here we provide a framework to study complex genomic regions. We
411 applied this simple, fast and affordable technology to study diverse human population
412 groups. Moreover, this approach can be applied to the generation of long-read data of
413 other regions or chromosomes of interest. As such, it could be used for the
414 characterization of virtually any species, although it would be especially advantageous for
415 those rich in complex genomic features.

416 **Methods**

417 Flow chromosome sorting followed by ONT or Illumina sequencing

418 Chromosome preparation was performed as previously described[16,17]. The libraries to
419 obtain the Illumina paired-end data were constructed using a SureSelect V6-Post Library
420 Kit. Raw data generated for the haplogroup A0 (HG02982 cell line) was retrieved from
421 Kuderna et al.[17] The data generated in each MinION run was basecalled using *Guppy*
422 *v5.0.15*[35] with the super accuracy model *dna_r94.1_450bps_sup*.

423 Adaptive sampling for the enrichment of a specific chromosome

424 We extracted DNA from cultured cells of haplogroup H (GM21113 cell line) using the
425 Qiagen MagAttract HMW Kit. DNA libraries for ONT sequencing were obtained using the
426 Ligation Sequencing Kit (SQK-LSK110) and sequenced in two ONT MinION flowcells (FLO-
427 MIN106 R9.4.1) using a MinION Mk1C with MinKNOW v21.02-beta4~xenial. We aimed for
428 the specific enrichment of the chrY and the chrY_random, by adding their nucleotide
429 sequence as provided in the GRCh38 assembly. This method bioinformatically labels the
430 reads that are being sequenced for enrichment or depletion. After a DNA strand enters
431 the pore, the sequencer only needs one second (around 420 bases) to decide whether to
432 continue sequencing the DNA if it matches the region of interest or to eject it if it does
433 not. Each of the strands that enter a pore will be labeled as *unblock* and *no decision* when
434 they are rejected by the pore or they are so short that their status remains inconclusive,
435 respectively. They will be labeled as *stop receiving* when they are on target, thus further
436 sequenced. Only reads labeled as *stop receiving* were used in this project.

437 The enrichment obtained with adaptive sampling highly depends on the fragment length
438 of the library. Longer DNA library lengths are preferred, as the adaptive sampling
439 enrichment algorithm takes a fixed amount of time to recognize whether to enrich a DNA
440 strand. Because of this, in order to target a specific region of the genome that is
441 particularly small (the Y chromosome represents ~1% of the genome), it will always be
442 better to have few long DNA fragments, rather than many short DNA fragments, as the
443 time spent by the sequencer scanning for on-target regions will be reduced.

444 Assessing the performance of two different enrichment methodologies

445 The coverage and enrichment factor for each chromosome were calculated as follows:

446 $$Coverage\ of\ chrN = \frac{Mapped\ bp\ in\ chrN}{Size\ chrN\ (bp, without\ N)} \text{ (Eq. 1)}$$

447 $$Enrichment\ factor\ of\ chrN = \frac{\frac{Mapped\ bp\ in\ chrN}{Total\ mapped\ bp}}{\frac{Size\ chrN \times n\ (bp, without\ N)}{Diploid\ genome\ size\ (bp, without\ N)}} \text{ (Eq. 2)}$$

448 Since more than 50% of the Y chromosome in GRCh38 is composed of long stretches of
449 unknown sequence (that in the assembly is seen as N), it is important to exclude these
450 regions from the coverage and enrichment calculations. Because of that, Eq. 1 and Eq. 2
451 only consider chromosome sizes without Ns. Moreover, for calculating the enrichment,
452 and in order to account for the real target space of each chromosome, the size of each of
453 them is multiplied by its ploidy.

454 Assembly generation

455 Basecalled passed reads (Q > 10) were mapped to the human GRCh38 genome assembly
456 using *minimap2 v2.17-r941*[62] with the option *-x map-ont*. The resulting bam was indexed
457 using *SAMTOOLS v1.12*[62,63] and the reads mapping either to chrY or chrY_random (chrY
458 specific reads) were retrieved.

459 We ran *Flye v2.9*[34] using the chrY specific reads with the option *--nano-hq* as suggested by
460 the developers while using data basecalled using *Guppy v5*[35] onwards with the super
461 accuracy model. We added the option *--scaffold* to enable scaffolding based on the
462 assembly graph, and included two internal rounds of polishing with the argument *-i 2*.

463 As we used uncorrected long reads to obtain the draft assemblies, we polished the initial
464 assemblies by first using ONT reads. We started with 2 rounds of *Racon v1.3.1*[64], using
465 *minimap2 v2.9-r720*[62] with the option *-x ont* to obtain the mapping file, and adding to *Racon*
466 the argument *-u* to keep any unpolished sequences. To further improve the assembly, we
467 then ran *medaka v1.4.1*[65] using the *medaka_consensus* program with default settings and
468 the model *-m r941_prom_sup_g507*.

469 Additionally, to polish the assemblies with Illumina data we used *HyPo v1.0.3*[66], mapping
470 the Illumina reads to the polished assembly. For mapping short reads to the existing
471 assembly we used *minimap2 v2.9-r720*[62] with the option *-x sr*.

472 Once we polished the assemblies we purged them using *purge_dups v1.2.5*[62,67], with
473 default parameters and the *-2* option. This was done to remove any haplotig present in
474 the assemblies. We obtained the mapping files using *minimap2 v2.14-r883*[62] with the
475 option *-x ont* to map the ONT reads to the polished assembly, and with the options *-xasm5*
476 *-DP* to map the split polished assembly to itself.

477 For the comparison of the assemblies generated from the two selective sequencing
478 methods (chromosome sorting and adaptive sampling), we downsampled the data of the
479 adaptive sampling experiment. For that, we used *Filtlong v0.2.0*[68] with the option --
480 *keep_percent 87.8* so as to retrieve 87.8% of the AS sequencing data. From that point on,
481 the assembly process was the same as the one explained (Supplementary Fig. 5).

482 Genome-to-genome comparisons

483 To obtain genome-to-genome alignments we used MuMmer v3.23[69] *nucmer* tool with
484 options *–-maxmatch -l 100 -c 100*. To manually scaffold the Y chromosome assemblies, we
485 used the dot-plot viewer *dot*[70].

486 Structural variant detection with long reads

487 Structural variation was called using *Sniffles v2.0.2*[38,39] with a minimum number of reads
488 that support an SV of *-s 10*, fed with the bam files for which we calculated and added MD
489 tags using *SAMTOOLS v1.9*[63], with the program *samtools calmd* adding options *-uAr -Q*. We
490 summarized the amount of SVs per type and filtered out the SVs considered 'IMPRECISE'
491 by *Sniffles*.

492 We merged the insertions and deletion separately with a maximum permitted distance
493 of 100 bp (so that indels located 100 bp upwards or downwards will be considered a single
494 event) found independently in all the cell lines using *SURVIVOR v1.0.7*[71]. We removed any
495 genotype with quality under 25 (MQ) and the events that were homozygous for the
496 reference genotype in all samples.

497 We used *Assemblytics v1.2.1*[40] to find structural variants in the different assemblies
498 generated by comparing them to the reference GRCh38. We looked for structural variants
499 with sizes between 10 to 100,000 and the unique sequence length required to call a
500 variant of 1,000.

501 Structural variant genotyping with short reads

502 We genotyped, using the indels obtained using Sniffles and *SURVIVOR* as reference, the
503 structural variants based on Illumina data with the program *graphtyper* v2.7.5 with the
504 option "genotype_sv" and only kept the indels with a quality > 0. We genotyped them with
505 the Illumina data generated in this study and with the Illumina data of the *1kgp*[31] available
506 for the Y chromosome.

507 Correlation between structural variant detection using long or short reads

508 To assess the reproducibility of the structural variant calls obtained with ONT data in
509 short-read data. We took the indels genotyped in the Illumina data and compared them
510 between platforms. For each structural variant, the ONT genotypes were assumed to be

511 true positives and all genotype calls were binarised into presence (1) or absence (0). For
512 each structural variant, and given that Y chromosomes are hemizygous, homozygous and
513 heterozygous alternative calls were considered present, and homozygous reference
514 genotypes absent.

515 As we wanted to study the correlation of binary variables, we made use of phi coefficients,
516 also known as Matthews correlation coefficient or MCC. Phi coefficients should be
517 interpreted similarly to a Pearson correlation coefficient. For the *1kgp* data, we
518 considered a structural variant to be present if it was at a frequency higher than 0.2.

519 Studying methylation using ONT

520 The methylation status was called using *nanopolish v0.13.2*[43], which assigns a log-
521 likelihood ratio to each individual CpG site. To avoid adding noise to the methylation
522 results, we only used reads with the highest mapping quality as provided by minimap2
523 (mapQ = 60) and filtered out all others. We used the default log-likelihood threshold of 2
524 as implemented in *nanopolish v0.12* onwards. As suggested by the developers, we called
525 methylation with the option --min-separation 5 to help calling CpG dense regions. The
526 methylation frequency was calculated for each site as the number of mapped reads
527 predicted as methylated divided by the number of total mapped reads.

528 We filtered out the few instances in which alternative alleles were present in a genomic
529 position with a cytosine in the reference sequence. We performed quantile normalization
530 on the methylation values across samples with a minimum coverage of 4x using the R
531 package *preprocessCore v1.56.0*[72]. CpG and genic annotations were obtained using the R
532 package *annotatr v1.24.0*[73]. Minor modifications were made to these annotations for
533 different analyses. All these modifications have been specifically described when used in
534 the text. For the overlapping regions in the genic annotations, the priority set was the
535 following: promoters, UTRs (5', 3'), first exon, non-first exons, all introns, and upstream
536 region.

537 **Contributions**

538 T.M.B and L.F.K.K. conceived the study. O.F., E.J., and E.V. cultured cells and performed
539 the flow cytometry. A.F. cultured cells. P.E.-C., C.F., and L.L. performed adaptive sampling.
540 I.G. and J.H. performed sequencing. L.F.K.K., A.S., L.F.-P., and E.L. provided analytical
541 support. M.D. and M.T. helped with data curation and analyses. M.P.-F. generated
542 assemblies and structural variant calls. P.E.-C. performed methylation analyses and
543 supervised analyses. D.J. designed and supervised analyses. P.E.-C. and M.P.-F. wrote the
544 manuscript with input from all co-authors.

Acknowledgments

MPF has support of an INPhINIT Retaining Fellowship from "La Caixa" Foundation (ID 100010434) with code LCF/BQ/DR20/11790032. TMB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 864203), "Unidad de Excelencia María de Maeztu", funded by the AEI (CEX2018-000792-M) and NIH 1R01HG010898-01A1.

References

1. Accounting for sex in the genome. *Nature Medicine* vol. 23 1243–1243 Preprint at https://doi.org/10.1038/nm.4445 (2017).
2. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
3. Wilson, M. A. The Y chromosome and its impact on health and disease. *Human Molecular Genetics* vol. 30 R296–R300 Preprint at https://doi.org/10.1093/hmg/ddab215 (2021).
4. Anderson, K., Cañadas-Garre, M., Chambers, R., Maxwell, A. P. & McKnight, A. J. The Challenges of Chromosome Y Analysis and the Implications for Chronic Kidney Disease. *Frontiers in Genetics* vol. 10 Preprint at https://doi.org/10.3389/fgene.2019.00781 (2019).
5. Molina, E., Clarence, E. M., Ahmady, F., Chew, G. S. & Charchar, F. J. Coronary Artery Disease: Why We should Consider the Y Chromosome. *Heart, Lung and Circulation* vol. 25 791–801 Preprint at https://doi.org/10.1016/j.hlc.2015.12.100 (2016).
6. Mank, J. E. The W, X, Y and Z of sex-chromosome dosage compensation. *Trends in Genetics* vol. 25 226–233 Preprint at https://doi.org/10.1016/j.tig.2009.03.005 (2009).
7. Tomaszkiewicz, M., Medvedev, P. & Makova, K. D. Y and W Chromosome Assemblies: Approaches and Discoveries. *Trends in Genetics* vol. 33 266–282 Preprint at https://doi.org/10.1016/j.tig.2017.01.008 (2017).
8. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
9. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* vol. 483 82–86 Preprint at https://doi.org/10.1038/nature10843 (2012).
10. Soh, Y. Q. S. *et al.* Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).
11. 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).
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. International Human Genome Sequencing Consortium. Finishing the euchromatic

583 sequence of the human genome. *Nature* **431**, 931–945 (2004).
584 14. Mendez, F. L., David Poznik, G., Castellano, S. & Bustamante, C. D. The Divergence of
585 Neandertal and Modern Human Y Chromosomes. *Am. J. Hum. Genet.* **98**, 728–734
586 (2016).
587 15. Tomaszkiewicz, M. *et al.* A time- and cost-effective strategy to sequence mammalian
588 Y Chromosomes: an application to the de novo assembly of gorilla Y. *Genome Res.* **26**,
589 530–540 (2016).
590 16. Kuderna, L. F. K. *et al.* Flow Sorting Enrichment and Nanopore Sequencing of
591 Chromosome 1 From a Chinese Individual. *Front. Genet.* **10**, 1315 (2019).
592 17. Kuderna, L. F. K. *et al.* Selective single molecule sequencing and assembly of a human
593 Y chromosome of African origin. *Nat. Commun.* **10**, 4 (2019).
594 18. Martin, S. *et al.* Nanopore adaptive sampling: a tool for enrichment of low abundance
595 species in metagenomic samples. *Genome Biol.* **23**, 11 (2022).
596 19. Doležel, J. *et al.* Chromosomes in the flow to simplify genome analysis. *Funct. Integr.*
597 *Genomics* **12**, 397–416 (2012).
598 20. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized
599 genomes. *Nat. Biotechnol.* **39**, 442–450 (2021).
600 21. Kovaka, S., Fan, Y., Ni, B., Timp, W. & Schatz, M. C. Targeted nanopore sequencing by
601 real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* **39**, 431–
602 441 (2021).
603 22. Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through
604 high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, 216
605 (2006).
606 23. Udaondo, Z. *et al.* Comparative Analysis of PacBio and Oxford Nanopore Sequencing
607 Technologies for Transcriptomic Landscape Identification of Penaeus monodon. *Life*
608 vol. 11 862 Preprint at https://doi.org/10.3390/life11080862 (2021).
609 24. Lang, D. *et al.* Comparison of the two up-to-date sequencing technologies for genome
610 assembly: HiFi reads of Pacbio Sequel II system and ultralong reads of Oxford
611 Nanopore. Preprint at https://doi.org/10.1101/2020.02.13.948489.
612 25. Tvedte, E. S. *et al.* Comparison of long-read sequencing technologies in interrogating
613 bacteria and fly genomes. *G3 Genes|Genomes|Genetics* vol. 11 Preprint at
614 https://doi.org/10.1093/g3journal/jkab083 (2021).
615 26. Rand, A. C. *et al.* Mapping DNA methylation with high-throughput nanopore
616 sequencing. *Nat. Methods* **14**, 411–413 (2017).
617 27. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing.
618 *Nat. Methods* **14**, 407–410 (2017).
619 28. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
620 29. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human
621 chromosome 8. *Nature* **593**, 101–107 (2021).
622 30. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X

623 chromosome. *Nature* **585**, 79–84 (2020).
624 31. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
625 32. Jobling, M. A. & Tyler-Smith, C. Human Y-chromosome variation in the genome-
626 sequencing era. *Nat. Rev. Genet.* **18**, 485–497 (2017).
627 33. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943–945
628 (2001).
629 34. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads
630 using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
631 35. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools
632 for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
633 36. Dida, F. & Yi, G. Empirical evaluation of methods for *de novo* genome assembly. *PeerJ*
634 *Computer Science* vol. 7 e636 Preprint at https://doi.org/10.7717/peerj-cs.636 (2021).
635 37. Mahmoud, M. *et al.* Structural variant calling: the long and the short of it. *Genome*
636 *Biol.* **20**, 246 (2019).
637 38. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-
638 molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
639 39. Smolka, M. *et al.* Comprehensive Structural Variant Detection: From Mosaic to
640 Population-Level. *bioRxiv* 2022.04.04.487055 (2022) doi:10.1101/2022.04.04.487055.
641 40. Nattestad, M. & Schatz, M. C. Assemblytics: a web analytics tool for the detection of
642 variants from an assembly. *Bioinformatics* **32**, 3021–3023 (2016).
643 41. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC)
644 over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6
645 (2020).
646 42. Yule, G. U. On the methods of measuring association between two attributes. *J. R.*
647 *Stat. Soc.* **75**, 579 (1912).
648 43. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de
649 novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
650 44. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and
651 tissues. *Genome Res.* **23**, 555–567 (2013).
652 45. Fraser, H. B., Lam, L. L., Neumann, S. M. & Kobor, M. S. Population-specificity of
653 human DNA methylation. *Genome Biol.* **13**, R8 (2012).
654 46. Husquin, L. T. *et al.* Exploring the genetic basis of human population differences in
655 DNA methylation and their causal impact on immune gene regulation. *Genome Biol.*
656 **19**, 222 (2018).
657 47. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression
658 variation in HapMap cell lines. *Genome Biol.* **12**, R10 (2011).
659 48. Illingworth, R. S. & Bird, A. P. CpG islands--'a rough guide'. *FEBS Lett.* **583**, 1713–1720
660 (2009).
661 49. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and
662 hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41**, 178–

663 186 (2009).
664 50. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human
665 genome. *Nature* **500**, 477–481 (2013).
666 51. Lowdon, R. F., Jang, H. S. & Wang, T. Evolution of Epigenetic Regulation in Vertebrate
667 Genomes. *Trends Genet.* **32**, 269–283 (2016).
668 52. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across
669 human tissues. *Science* **369**, 1318–1330 (2020).
670 53. Ross, J. L., Tartaglia, N., Merry, D. E., Dalva, M. & Zinn, A. R. Behavioral phenotypes in
671 males with XYY and possible role of increased NLGN4Y expression in autism features.
672 *Genes Brain Behav.* **14**, 137–144 (2015).
673 54. Chen, J., Yu, S., Fu, Y. & Li, X. Synaptic proteins and receptors defects in autism
674 spectrum disorders. *Front. Cell. Neurosci.* **8**, 276 (2014).
675 55. Dall'Alba, G., Casa, P. L., Abreu, F. P. de, Notari, D. L. & de Avila E Silva, S. A Survey of
676 Biological Data in a Big Data Perspective. *Big Data* **10**, 279–297 (2022).
677 56. Kamble, S. S., Gunasekaran, A., Goswami, M. & Manda, J. A systematic perspective on
678 the applications of big data analytics in healthcare management. *Int. J. Healthc.*
679 *Manag.* **12**, 226–240 (2019).
680 57. McCarthy, N. S. *et al.* Meta-analysis of human methylation data for evidence of sex-
681 specific autosomal patterns. *BMC Genomics* **15**, 981 (2014).
682 58. Johansson, A., Enroth, S. & Gyllensten, U. Continuous Aging of the Human DNA
683 Methylome Throughout the Human Lifespan. *PLoS One* **8**, e67378 (2013).
684 59. Palumbo, D., Affinito, O., Monticelli, A. & Cocozza, S. DNA Methylation variability
685 among individuals is related to CpGs cluster density and evolutionary signatures.
686 *BMC Genomics* **19**, 229 (2018).
687 60. Galanter, J. M. *et al.* Differential methylation between ethnic sub-groups reflects the
688 effect of genetic ancestry and environmental exposures. *Elife* **6**, (2017).
689 61. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock
690 theory of ageing. *Nat. Rev. Genet.* **19**, 371–384 (2018).
691 62. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
692 3094–3100 (2018).
693 63. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
694 2078–2079 (2009).
695 64. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome
696 assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
697 65. GitHub - nanoporetech/medaka: Sequence correction provided by ONT Research.
698 *GitHub* https://github.com/nanoporetech/medaka.
699 66. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super Fast & Accurate Polisher for Long Read
700 Genome Assemblies. *bioRxiv* 2019.12.19.882506 (2019)
701 doi:10.1101/2019.12.19.882506.
702 67. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome

assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
68. rrwick/Filtlong. *GitHub* https://github.com/rrwick/Filtlong.
69. Kurtz, S. *et al.* Genome Biology. vol. 5 R12 Preprint at https://doi.org/10.1186/gb-2004-5-2-r12 (2004).
70. Website. GitHub - MariaNattestad/dot: Dot: An interactive dot plot viewer for comparative genomics. GitHub https://github.com/marianattestad/dot.
71. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
72. Ben Bolstad <bmb@bmbolstad.com>. *preprocessCore*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.PREPROCESSCORE.
73. Cavalcante, R. G. & Sartor, M. A. annotatr: genomic regions in context. *Bioinformatics* **33**, 2381–2383 (2017).