

METHODOLOGY ARTICLE

In-solution Y-chromosome capture-enrichment on ancient DNA libraries

Diana I Cruz-Dávalos^{1,2,3}, María A Nieves-Colón⁴, Alexandra Sockell⁵, G David Poznik⁶, Hannes Schroeder^{7,8}, Anne C Stone^{4,9}, Carlos D Bustamante^{5,10}, Anna-Sapfo Malaspinas^{1,3,11} and María C Ávila-Arcos^{2*}

*Correspondence:

mavila@iigh.unam.mx

²International Laboratory for Human Genome Research, National Autonomous University of Mexico, Boulevard Juriquilla 3001, 762330 Juriquilla, Mexico
Full list of author information is available at the end of the article

Abstract

Background: As most ancient biological samples have low levels of endogenous DNA, it is advantageous to enrich for specific genomic regions prior to sequencing. One approach – in-solution capture-enrichment – retrieves sequences of interest and reduces the fraction of microbial DNA. In this work, we implement a capture-enrichment approach targeting informative regions of the Y chromosome in six human archaeological remains excavated in the Caribbean and dated between 200 and 3,000 years BP. We compare the recovery rate of Y-chromosome capture (YCC) alone, whole-genome capture followed by YCC (WGC+Y) versus non-enriched (pre-capture) libraries.

Results: We recovered 17–4,152 times more targeted unique Y-chromosome sequences after capture, where 0.01–6.2% (WGC+Y) and 0.01–23.5% (YCC) of the sequence reads were on-target, compared to 0.0002–0.004% pre-capture. In samples with endogenous DNA content greater than 0.1%, we found that WGC followed by YCC (WGC+Y) yields lower enrichment due to the loss of complexity in consecutive capture experiments, whereas in samples with lower endogenous content, WGC+Y yielded greater enrichment than YCC alone. Finally, increasing recovery of informative sites enabled us to assign Y-chromosome haplogroups to some of the archeological remains and gain insights about their paternal lineages and origins.

Conclusions: We present to our knowledge the first in-solution capture-enrichment method targeting the human Y-chromosome in aDNA sequencing libraries. YCC and WGC+Y enrichments lead to an increase in the amount of Y-DNA sequences, as compared to libraries not enriched for the Y-chromosome. Our probe design effectively recovers regions of the Y-chromosome bearing phylogenetically informative sites, allowing us to identify paternal lineages with less sequencing than needed for pre-capture libraries. Finally, we recommend considering the endogenous content in the experimental design and avoiding consecutive rounds of capture for low-complexity libraries, as clonality increases considerably with each round.

Keywords: capture-enrichment; Y chromosome; ancient DNA

Background

Uniparental markers such as those on the mitochondrial chromosome (mtDNA) and on Y-chromosome DNA (Y-DNA) are widely used to infer the demographic histories of specific human lineages [1]. Although much smaller than the nuclear genome, the inheritance mechanism and lack of recombination make them powerful

tools for inferring ancestry and estimating the ages of pedigrees and the times to the most recent common ancestors (TMRCA) of the mtDNAs and Y-DNAs of present-day populations [2]. Analyses of modern and ancient mtDNA and Y lineages have broadened our knowledge of diversification and founder events from human population history [3, 4, 5].

Due to the large number of copies in each cell, mtDNA has been at the forefront of ancient DNA research [6]. In contrast, each cell possesses just one copy of the Y chromosome. Thus, when analyzing ancient samples, the probability of retrieving any given portion of Y-chromosome DNA is much lower than for mtDNA. Furthermore, as endogenous DNA is often found highly fragmented and in low quantity, recovering ancient DNA (aDNA) from a pool of endogenous and contaminating environmental DNA is extremely challenging and costly. To overcome these challenges, methods have been developed to increase the endogenous DNA proportion of sequencing libraries. These methods target select genomic regions, such as SNPs, whole chromosomes, or mitochondrial or nuclear genomes [3, 7, 8, 9, 10] prior to sequencing. They consequently increase the proportion of genomic regions of interest while reducing sequencing costs.

There are two main types of enrichment methods: solid phase enrichment [11, 12], wherein DNA probes are fixed to a surface, and in-solution enrichment, wherein biotinylated DNA or RNA probes hybridize to the targeted molecules of a DNA library [7, 10, 13]. In the in-solution approach, the probe-target complex is captured with streptavidin-coated magnetic beads, and the remaining fragments, including those from microbial DNA contamination, are washed away. Capture-enrichment approaches have enabled DNA retrieval from samples that initially showed small amounts of endogenous DNA [9, 14, 15]. Consequently, these enrichment methods have positively impacted ancient genomics research by lowering endogenous content requirements, thereby increasing the number of samples that can be employed for profiling.

Capture-enrichment strategies have been applied to target genome-wide SNP sets and to specific subsets of the genome to study the phylogenetic context of ancient populations. Recent implementations [3, 5] include probes targeting thousands of autosomal and Y SNPs characterized by the Simons Genome Diversity Project [16] and the International Society of Genetic Genealogy (ISOGG, <https://isogg.org/>). However, due to the provenance of the samples of the ISOGG consortium, ISOGG SNPs are best suited to genotype present-day European haplogroups. Consequently, aDNA enrichment has been applied to study Y-chromosome variation in ancient European and Middle-Eastern individuals, while studies of Africans [17] and Native Americans have been restricted either to direct interrogation of known Y-DNA markers with targeted PCR-based sequencing [18] or to low and medium-depth whole-genome sequencing [17, 19, 20, 21].

An important consideration of enrichment designs targeting pre-selected SNPs is the ascertainment bias introduced and the impediment of discovering new variants. An ideal strategy would involve capturing the whole Y chromosome, however its abundance of repetitive sequences makes it less amenable for capture experiments [13]. To overcome these constraints, we made use of a probe design targeting 10.3 megabases (Mb) of the 57.2 Mb Y chromosome defined by Poznik

and colleagues [2]. The 41 regions within this 10.3 Mb were selected to fall within the non-recombining portion of the Y chromosome, be depleted of repeats, and well suited for genotype calling from short read sequence data [2]. Furthermore, we were interested in assessing whether Y-DNA could be enriched from libraries with very low endogenous content that had been subjected to WGC and for which we also had pre-capture libraries. We thus tested this approach and compared different enrichment strategies, on samples excavated from the Caribbean, a region that poses a particular challenge for DNA preservation. Previous studies on some of these samples failed to obtain enough Y-DNA data to reliably call a haplogroup even after WGC [22]. Consequently, we investigated the parameters affecting the quality and the quantity of the data and, at the same time, assessed the extent to which the enrichment improved the resolution of the Y-chromosome haplogroup assignment. Our results provide a better understanding of the parameters that affect Y-DNA enrichment experiments and illustrate its benefits for studying the paternal genetic ancestry of ancient human populations.

Methods

Samples

We performed 18 capture-enrichment experiments on DNA libraries obtained from the archaeological remains of six individuals excavated from Caribbean contexts. Two samples (STM1 and STM2) belong to 17th-Century enslaved males of African origin from Saint Martin (Lesser Antilles) and were previously reported in [22]. The other four (PI174, PI383, PI435, and PI437) were obtained from archaeological remains from the Paso del Indio site (PI) in Puerto Rico. These four dated between 824 and 1039 CE, as described in [23, 24].

Ancient DNA extraction

DNA from the STM samples was extracted from tooth roots, as described in [22]. Sampling and DNA extractions for PI samples were conducted at the Arizona State University Ancient DNA Laboratory, a Class 10,000 clean-room facility. Teeth were cleaned with a 1% sodium hypochlorite solution, and the outer surfaces of the tooth roots were mechanically removed with a Dremel tool. Teeth were sliced transversely at the cemento-enamel junction using the Dremel. The roots were then covered in aluminum foil and pulverized by blunt force with a hammer, as in [25]. To avoid contamination, additional precautions were taken, including single use of Dremel wheels, bleach decontamination and UV irradiation of tools and the work area before and between uses, as well as full body coverings for all researchers [26]. DNA was extracted following [14], using 50 mg of pulverized tooth material. Extracts and extraction blanks were quantified with the Qubit 2.0 High Sensitivity assay [27].

Ancient DNA library preparation

DNA extracted from STM samples was built into 6-bp-indexed double-stranded Illumina libraries, as described in [22]. For PI samples, double-stranded Illumina libraries were prepared following the protocol in [28]. Extraction blanks were also converted into libraries, and an additional negative library control containing only ddH₂O was also included. 1:100 dilutions of each library were prepared for

quality screening through Real-Time PCR (qPCR) using the Thermo Scientific Dynamo SYBR Green qPCR kit with ROX. Reactions were run in triplicate and prepared in final volumes of 20 μ l with the following conditions: 10 μ l of 2X Dynamo SYBR Green qPCR Master Mix with 0.3 \times ROX, 1 μ l of primer IS7 (5'-ACACTCTTTCCCTACACGAC-3') at 10 μ M, 1 μ l of primer IS8 (5'-GTGACTGGAGTTCAGACGTGT-3') at 10 μ M, 7 μ l of ddH₂O, and 1 μ l of library dilution. Reactions were heated to 95°C for 10 minutes for initial denaturation, and further denaturations were performed at 95°C for 15 seconds and for 40 1-minute cycles at 60°C. A final disassociation stage was added at the end of these cycles: 95°C for 15 seconds, 60°C for 15 seconds and 95°C for 15 seconds. Quantification was performed using an ABI7900HT thermocycler and analyzed with SDS software. After qPCR, all libraries were double-indexed as in [29]. To increase library complexity, four 100 μ l indexing reactions were performed per library with the following conditions: 10 μ l of *Pfu* Turbo Buffer, 2.50 μ l of 10 mM dNTPs, 1.50 μ l of 10 mg/ml Bovine Serum Albumin, 2 μ l of P5 indexing primer (5'-AATGATACGGCGACCACCGAGATCTACAC xxxxxxACACTCTTTCCCTACACGACGCTCTT-3') at 10,000 nM, 2 μ l of P7 indexing primer (5'-CAAGCAGAAGACGGCATACGAGAT xxxxxxGTGACTGGAGTTCAGACGTGT-3') at 10,000 nM, 72 μ l of ddH₂O, 1.00 μ l of *Pfu* Turbo enzyme (Agilent), and 9 μ l of DNA library. Reactions were heated to 95°C for 15 minutes for initial denaturation. Further denaturation, annealing, and elongation were performed at 95°C for 30 seconds, at 58°C for 30 seconds, and for 10 45-second cycles at 72°C. Final extension was performed at 72°C for 10 minutes and reactions were then kept at 10°C. All four aliquots of each amplified library were combined, and the library was purified with the Qiagen MinElute PCR purification kit following manufacturer's instructions with the following modification: the EB buffer was preheated to 65°C before use, and reactions were eluted in 30 μ l. A 1- μ l aliquot of each library was used for quantification with the Qubit 2.0 Broad Range assay. Purified libraries were further diluted to a factor of 1:1,000 and quantified with the KAPA Library Quantification kit (Kapa Biosystems) following manufacturer's instructions.

Indexed libraries were amplified a second time to increase the amount of DNA. To increase library complexity, four 100- μ l amplification reactions were performed per library. PCR conditions were: 10 μ l of 10X Accuprime *Pfx* reaction mix, 3 μ l of IS5 primer at 10 μ M, 3 μ l of IS6 primer at 10 μ M, 76 μ l of ddH₂O, 1 μ l of AccuprimeTM *Pfx* enzyme, and 7 μ l of DNA library. Reactions were heated to 95°C for 2 minutes for initial denaturation, and further denaturation, annealing, and elongation were performed at 95°C for 15 seconds, 60°C for 30 seconds, and for 7–13 1-minute cycles at 68°C. Final extension was performed at 68°C for 5 minutes and reactions were then kept at 4°C. All four aliquots of each amplified library were combined, and the library was purified with Qiagen MinElute PCR purification kit as detailed above. 1 μ l of each purified and amplified library were used for flourometric quantification. Purified libraries were further diluted to a factor of 1:10,000 and quantified with the KAPA Library Quantification kit (Kapa Biosystems) following the manufacturer's instructions. 1 μ l of each library was used for fragment analysis with the Agilent 2100 Bioanalyzer DNA 1000 chip.

Whole-Genome Capture (WGC)

Whole-Genome Capture was performed on each of the libraries obtained from the six archaeological samples (STM1, STM2, PI174, PI383, PI435, PI437) following published protocols. We used the human whole-genome enrichment kit MYbaits (MYcroarray, Ann Arbor, online version 1.3.8) to capture STM libraries, as reported in [15]. For PI libraries, we implemented the WISC approach [9], starting with 500 ng per library and hybridizing for 66 hours. The libraries were PCR amplified for 15–20 cycles.

Y-chromosome bait design

We used DNA biotinylated probes (baits) from Nimblegen's SeqCap EZ Choice XL Enrichment Kit for Y capture. Baits were designed using Roche's NimbleDesign algorithm to target 10.3 Mb well suited for reliable genotype calling and haplogroup inference [2]. The 2.1 million probes were 105 bp in length, with an average tiling overlap of 21 bp for adjacent probes. Probes were designed using the hg19/GRCh37 reference sequence for the Y chromosome, most of which is derived from a single European haplogroup.

Y-chromosome capture (YCC)

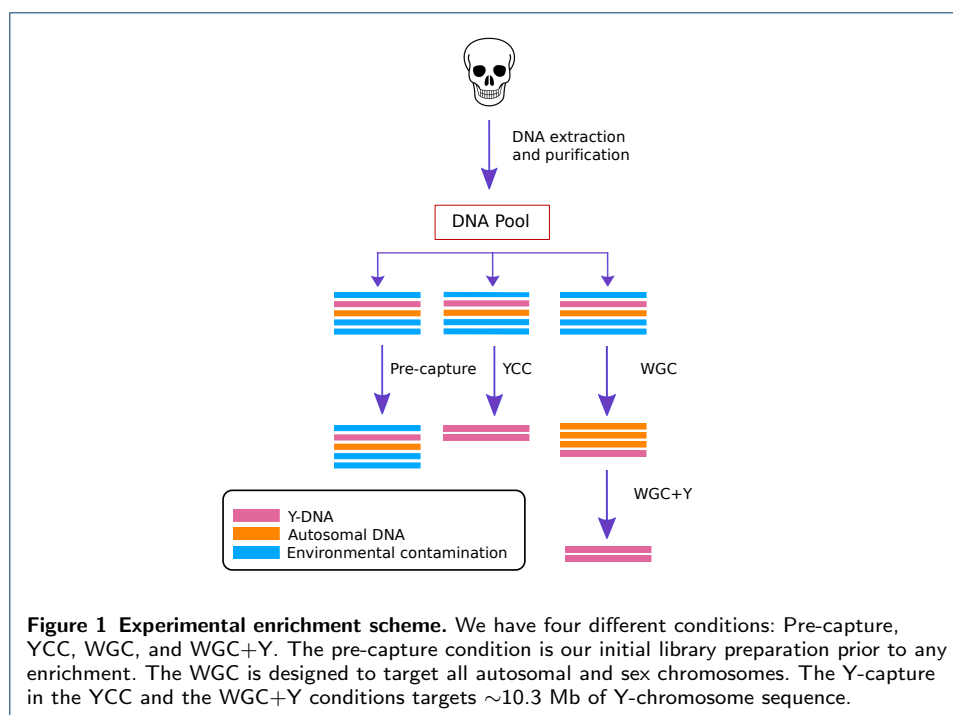
We performed Y-chromosome capture-enrichment experiments on both pre-capture and WGC libraries. Libraries were pooled in equal masses. Capture reactions were performed according to NimbleGen SeqCap EZ XL protocol, with the following modifications: due to limited sample availability, the total mass of the pooled libraries was ~500 ng rather than the recommended 1.25 µg; hybridization was performed for a total of 65 hours (48–72 h recommended); and the adapter-blocking oligonucleotides were IDT xGen blocking oligos, rather than those sold by Roche, due to increased on-target rates with the IDT xGen oligos. Following capture, libraries were amplified with 6 cycles of PCR, and quality was assessed using the Agilent Bioanalyzer High Sensitivity kit.

Illumina sequencing

Pre-capture and WGC libraries for the STM samples were sequenced at the National High Throughput DNA Sequencing Centre in Copenhagen, Denmark, on a HiSeq2000 platform using single-end 100-bp runs, as reported in [22] and [15]. PI pre-capture and WGC libraries were paired-end sequenced on the NextSeq500 using a High Output 150-cycle kit with paired-end 76-bp reads. All libraries subjected to YCC (i.e., YCC and WGC+Y libraries from all samples) were sequenced on the NextSeq500 using the High Output 150 cycle kit at Sanford University using paired-end 76-bp mode.

Sequence data processing and mapping

FASTQ-format reads from the pre-capture and WGC conditions are available for the STM samples through the European Nucleotide Archive, project PRJEB8269, experiment accession numbers ERX682089, ERX682243, ERX682248, and ERX682249 [22]. We processed these reads, as well as the reads generated for this study (YCC and WGC+Y libraries for the STM and all PI libraries) with the



following steps. To trim adapters and low quality bases, we used AdapterRemoval v2 with the default options in single-end mode for STM pre-capture and WGC libraries, and in paired-end mode for all PI and STM YCC libraries [30]. As the yield per experiment is variable and can bias the comparisons, we subsampled 10 times an equal number of reads for each experimental condition and for each individual using seqtk (<https://github.com/lh3/seqtk>). To determine the total number of sequences to subsample, we selected the lowest number of reads that passed the trimming filters for each individual across experiments (Table 1). The sequences were then aligned to the Homo sapiens reference genome build 37 (hg19) using the BWA aligner [31] implemented in PALEOMIX [32], with a mapping quality threshold set at 30. The quality of the aligned bases was rescaled with mapDamage2 [33] to lower the quality of mismatches to the reference sequence that likely derive from DNA damage.

To calculate the enrichment rate, we used the subsampled data and computed the average number of unique reads mapped to the on-target regions in each experiment. Then, we calculated fold-enrichment by dividing the on-target average of YCC or WGC+Y experiments by that of the pre-capture libraries. Specifically, we calculated this fold-enrichment for YCC using the pre-capture condition as a baseline and compared the WGC+Y experiments to both the pre-capture and WGC conditions. When we observed no reads aligning to the Y chromosome or target regions, we used the maximum number of reads observed across the replicates of a given library as a baseline. We computed binomial proportion confidence intervals for the mean endogenous content, the proportion of on- and off-target reads, and clonality, and we conducted a t test for the length estimate. Confidence intervals and chi-squared tests were computed in R software, version 3.3.1 [34].

Table 1 Total number of reads. Ten replicates per library were obtained by down-sampling to the minimum number of retained reads within each sample (underlined). The “Mapping to chrY” and “% of on-target” sections indicate the number of unique reads mapping to the Y-chromosome, and the percentage of unique on-target reads respect to the total reads mapping to the Y-chromosome, respectively.

		STM1	STM2	PI174	PI383	PI435	PI437
Pre-capture	Total reads	34,025,874	14,973,474	9,173,100	2,795,632	4,617,394	9,986,135
	mapping to chrY	1,384	301	4	3	6	6
	% of chrY on-target	74%	72%	25%	67%	50%	33%
YCC	Total reads	68,795	41,884	159,728	107,796	41,810	56,157
	mapping to chrY	16,430	2,562	17	27	26	12
	% of chrY on-target	99%	99%	100%	100%	96%	100%
WGC	Total reads	14,973,474	29,884,294	10,052,798	10,265,430	12,363,127	9,132,306
	mapping to chrY	6,407	1,925	5	8	57	7
	% of chrY on-target	62%	63%	60%	100%	65%	57%
WGC+Y	Total reads	98,540	30,629	17,212,495	14,557,575	10,181,433	11,356,277
	mapping to chrY	4,854	236	37	152	242	47
	% of chrY on-target	96%	95%	100%	99%	97%	96%

To call Y-chromosome genotypes for each sample, we first merged data across experiments. We used the haploid genotype caller implemented in ANGSD, retaining only bases with quality scores of at least 13 and sampling one random base at each site [35]. Finally, we performed a binary tree search with a custom script to find the most derived SNP that determines the haplogroup of the individuals. We used as input the phylogenetic tree constructed from the Y-SNPs reported in Phase 3 of the 1000 Genomes Project [4].

Sex determination

To determine the biological sex of the six individuals, we used the script in [36] to calculate the ratio (R_y) of reads mapping to the Y-chromosome to those mapping to both sex chromosomes [36]. R_y values above 0.075 are consistent with a male genotype.

Yield and enrichment curves

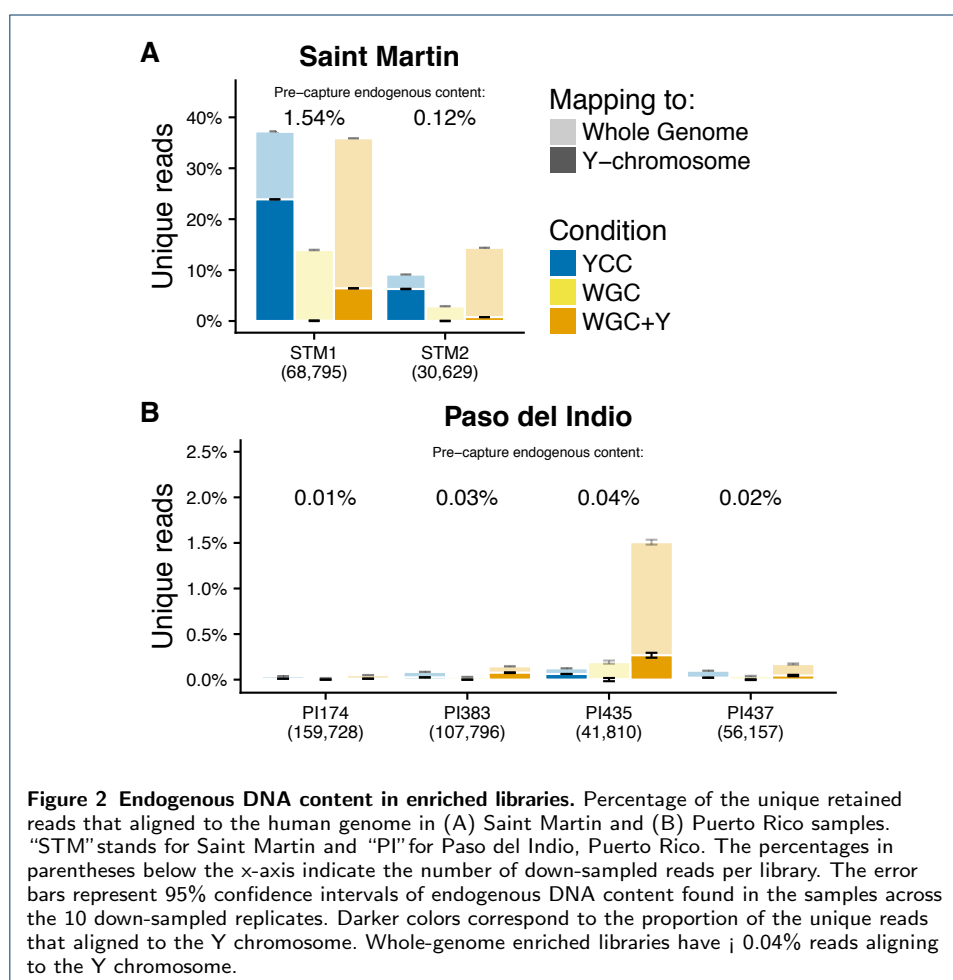
We estimated the yields and complexities of the libraries, with respect to the reads mapping to the targeted regions, with the preSeq package implemented in R (preseqR, [37]) and corrected the amount of required sequencing by the fraction of on-target reads in the libraries. Since the method relies on having a fraction of duplicated reads to estimate the yield, for the cases where the pre-capture libraries did not have duplicated on-target reads to adjust a yield curve, we instead assumed a linear relationship with a slope equal to the proportion of unique on-target reads present in the library. We then modeled an “enrichment curve” to explore the level of enrichment predicted for different amounts of sequencing. To this end, we used the median unique on-target reads estimated by PreSeq to calculate an expected fold-enrichment. We divided the median value estimated by PreSeq of each captured library by the median of its pre-captured counterpart (i.e., YCC vs. pre-capture, WGC+Y vs. pre-capture, and WGC+Y vs. WGC).

Results

Enrichment rates

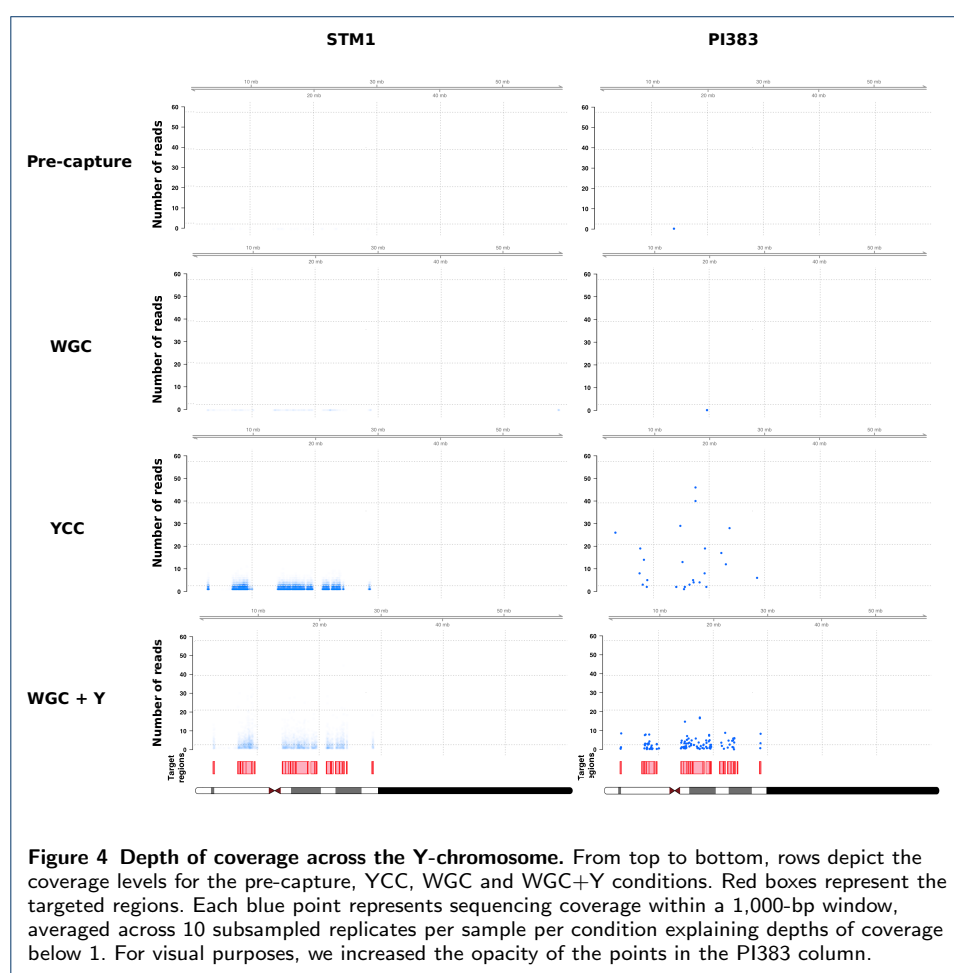
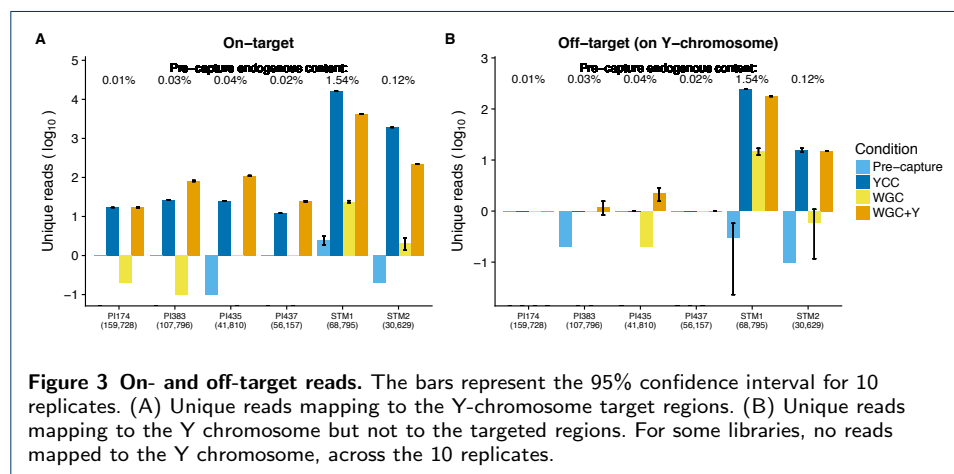
We tested the performance of Y-chromosome capture on Illumina sequencing libraries obtained from the archaeological remains of six individuals excavated in

the Caribbean islands of Saint Martin (STM1 and STM2) and Puerto Rico (PI174, PI383, PI435, and PI437) (Table 1). For each sample, we performed a series of enrichment experiments, as depicted in Figure 1. First, we shotgun-sequenced a DNA library without performing any enrichment. We then performed a capture reaction targeting a set of DNA probes covering 10.3 Mb of the non-recombining portion of the Y chromosome. These regions were validated by Poznik and colleagues in [2] as being well suited for unambiguous read mapping and for yielding reliable genotype and haplogroup calls from short-read sequencing. Additionally, we performed another set of capture experiments, either enriching only the whole-genome (WGC) or the on-target regions after having enriched the whole genome (WGC+Y). First, we confirmed the molecular sex of the samples and determined that the six individuals each had a karyotype consistent with XY [36]. We then assessed the performance of the capture experiments.



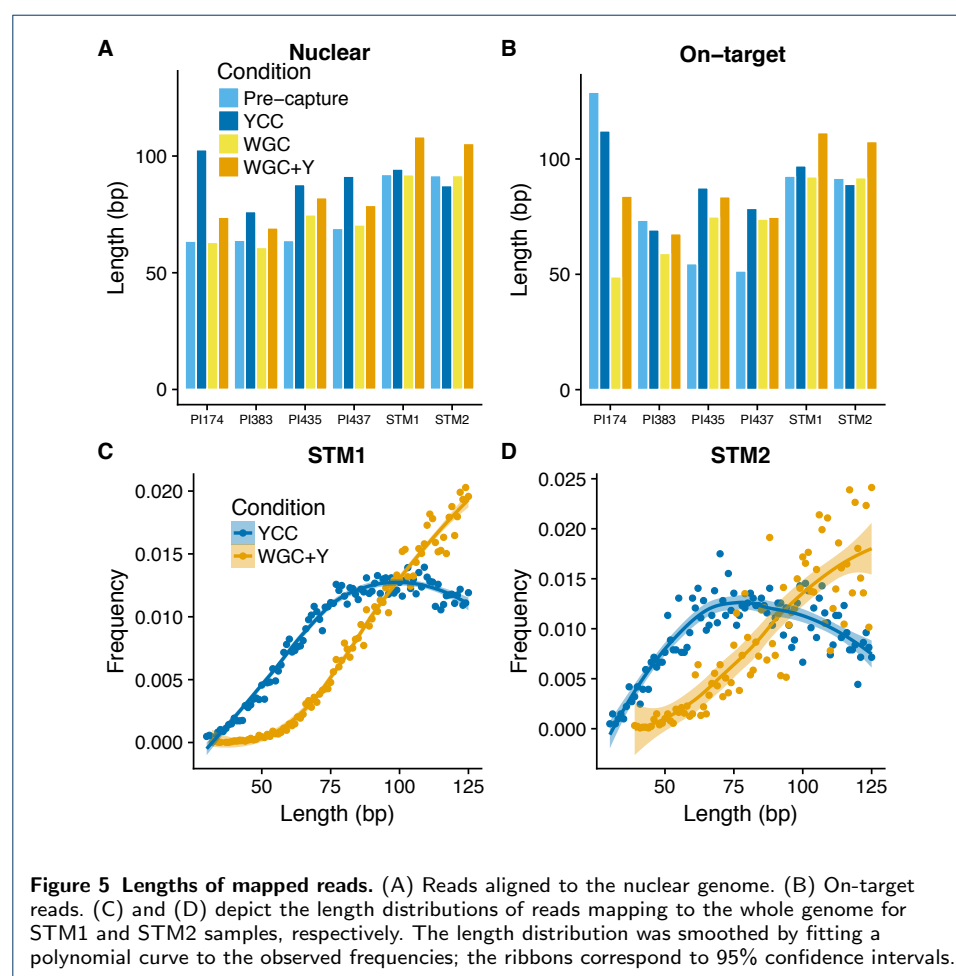
Since experiments yielded differing numbers of reads per sample, we down-sampled to equal numbers per individual as described in Methods (Table 1). The pre-capture libraries yielded just 0.01% to 1.54% unique reads aligning to the human genome (Figure 2), and less than 0.004% mapping to the Y-chromosome. After implementing the YCC and WGC+Y enrichments, the endogenous DNA content increased, on average, by factors of 24.2 to 122.0 for the STM samples (Figure 2A) and by factors

between 3.2 and 38.9 for the PI samples (Figure 2B). Moreover, in the YCC and WGC+Y libraries, 5.3% to 68.3% of the human STM reads mapped to the Y chromosome, with 17.7% to 50.0% the corresponding figures for the PI samples.



To evaluate whether the enrichment experiments effectively recovered the targeted regions, we compared the total number of unique reads mapping to the targeted

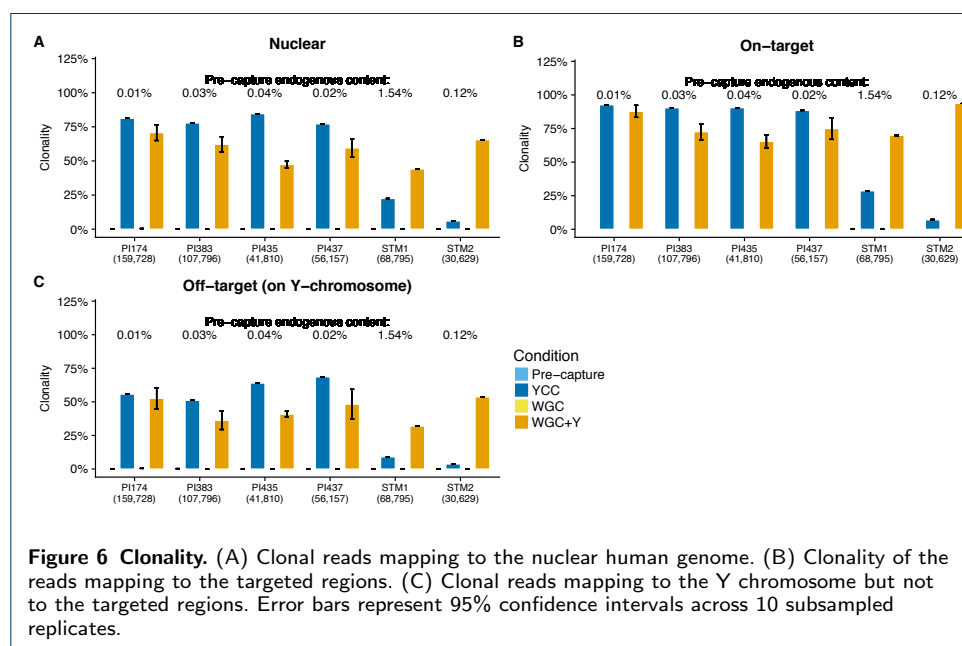
regions to the number of off-target reads (Figure 3). YCC experiments on the STM samples yielded 6,476 to 9,549-fold increases of on-target sequences compared to the pre-capture condition. The WGC+Y experiments on the same samples resulted in 105- to 176-fold-enrichment compared to WGC alone. For the PI samples, YCC experiments resulted in 12- to 250-fold enrichment, and we observed 85.5- to 813-fold enrichment for the WGC+Y approach. Although we observed an increase in the off-target content for the STM enrichments (Figure 3), it is one order of magnitude smaller than their respective on-target enrichment. Above 96% of the reads mapping to the Y chromosome were on target in both YCC and WGC+Y, in contrast to the pre-capture and WGC experiments where these figures range between 25% and 72% (Table 1). Overall, the distribution of the on-target sequences in all Y-chromosome enrichment experiments is qualitatively even (Figure 4, Supplementary figure 1). In summary, all Y-chromosome capture enrichment experiments consistently increased the number of unique on-target reads.



Length distribution and clonality

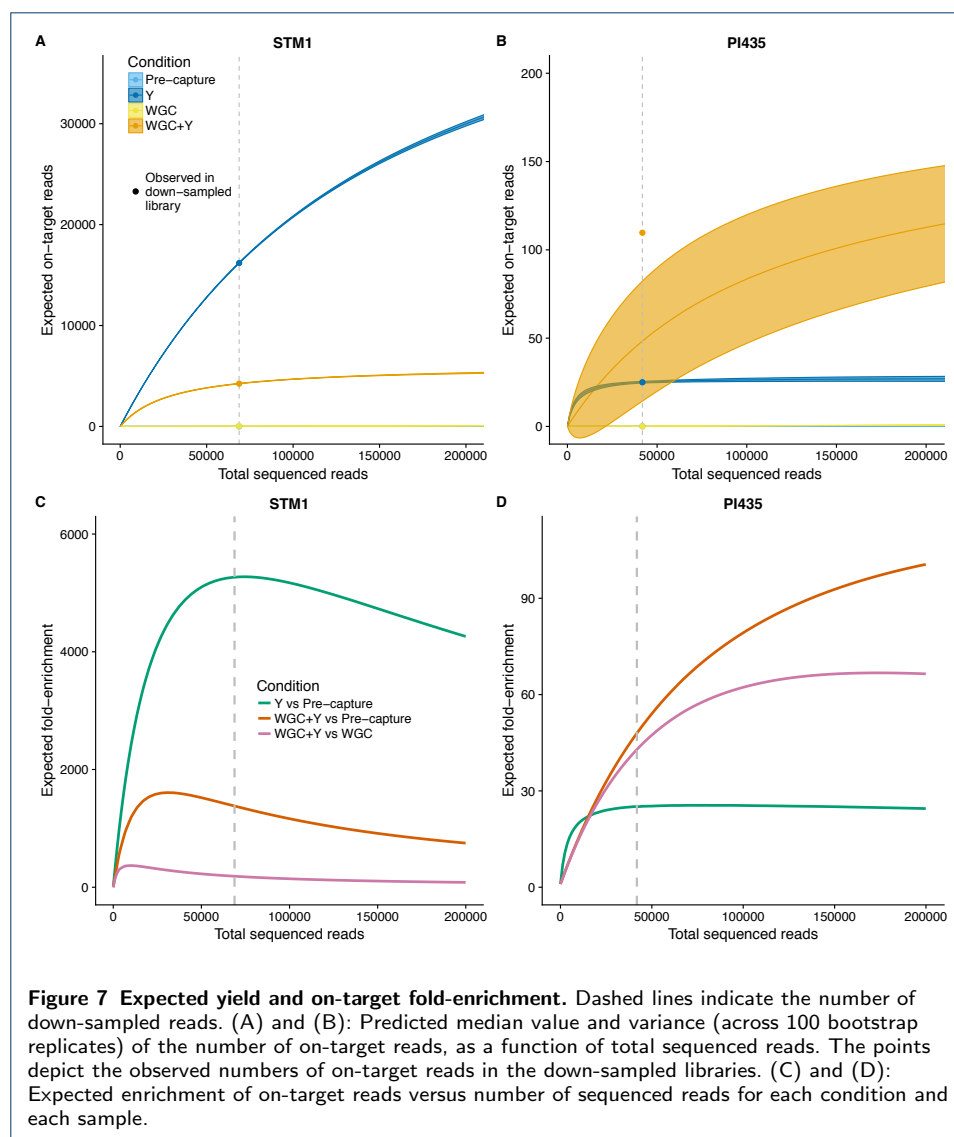
To explore which features of the pre-capture libraries may have influenced the differences in enrichment rates between the STM and PI groups, we contrasted the lengths and complexities of the individual libraries across experiments. We observed

a significant trend toward longer fragments after enrichment for all experiments (Figure 5, Supplementary figure 2) (paired samples t-test, p -value = 0.002), consistent with previous findings [7, 15, 38]. Reads from the STM samples were 91.7 to 92.1 base pairs (bp) long for the pre-capture condition, 87.4 to 94.5 bp after YCC, and 105.5 to 108.3 long after WGC+Y. Likewise, whereas the average length of PI reads in the pre-capture libraries ranged from 63.6 to 69.1 bp, average lengths increased to 76.3–102.3 and 69.3–82.2 and bp after YCC (paired samples t-test, p -value = 0.02) and WGC+Y (paired samples t-test, p -value = 0.03), respectively. On-target clonality levels (percentage of PCR duplicates) were considerable for all samples in the YCC and WGC+Y experiments. We observed 7.0% to 28.6% clonality for YCC of the STM samples at 68,795 and 30,629 down-sampled reads. The remaining WGC+Y (STM), and all WGC+Y and YCC (PI) libraries had greater clonality values, ranging from 65.4% to 94.1% (Figure 6). Although we did not observe on-target duplicates with which to calculate the clonality in any of the down-sampled pre-capture libraries (Supplementary table 2), using the whole data we estimated clonality values ranging from 0.8% to 3.5% (for 101,373,044 and 10,506,505 sequenced reads, Supplementary table 1).



On-target yield for aDNA libraries

The yield curves for both capture conditions (YCC and WGC+Y) corroborate the high clonality of the PI libraries (Figure 7 A, B). We observed that the YCC libraries of these samples plateaued at very shallow sequencing, saturating at 25,000 sequenced reads, compared to saturations at 50,000–100,000 sequenced reads for the WGC+Y libraries. However, the complexity curves indicate that after sequencing 100,000 reads of the WGC+Y libraries, we would not retrieve more than 150 different reads, regardless of the capture approach. On the other hand, although we sequenced fewer than 100,000 reads for each of the YCC experiments on the



STM individuals, the complexity curves suggest that these libraries could be further sequenced to increase the coverage of the targeted regions.

Enrichment rates based on yield estimates

We computed enrichment curves (Figure 7 C, D) to predict enrichment rates for deeper sequencing. Consistently, we observe that the YCC of STM2 recovers 1,790 more on-target reads than the pre-capture experiment at the down-sample point, but, as it has not reached saturation, this enrichment value can increase to at least 5,000-fold, similar to the maximum enrichment of STM1. In contrast, the projected enrichment of the PI WGC+Y libraries at 200,000 total reads is only 100- and 72-fold versus the pre-capture and WGC libraries, respectively. We also note that, for PI samples, the three enrichment rates calculated “decelerate” at 100,000 sequenced reads or fewer, once more reflecting the low complexity levels of the initial pool.

Y haplogroup calls

Finally, we combined all the sequence data generated for each sample to call Y-chromosome haplogroups. With the new data generated, we were able to call haplogroups for both STM individuals. The combined on-target depth of coverage was $0.03\times$ to $0.24\times$, with a sequencing depth of at least one for 0.36–1.96 Mb (Table 2). In the STM1 individual we observed derived alleles belonging to the R1b-M343 clade, consistent with its previously reported haplogroup (R1b1c-V88) [22]. In the same study, even after WGC, the haplogroup for STM2 could not be resolved [22]. Only after integrating the previously generated data with that produced in this study, were we able to assign the Y-haplogroup as E1b1a1a1-M80. Regarding the PI samples, we did not find any reads covering the SNPs in the database in the pre-capture or the WGC libraries. After YCC or WGC+Y, we observe between 5 and 74 variant sites per library, leading to the identification of haplogroups defining broader regions. For instance, the haplogroup found for PI383, P-M45, is the immediate ancestor of haplogroups R and Q. In Table 2, we show the haplogroups for all the individuals according to the most derived SNP identified in each condition.

Table 2 Numbers of Y-chromosome bases retrieved and haplogroups.

Sample	Condition	Positions recovered	Haplogroup retrieved
STM1	All	2,205,331	R1b-M343
	Pre-capture	93,725	R1b-M343
	YCC	1,372,731	R1b-M343
	WGC	354,255	R1b-M343
	WGC+Y	493,085	R1b-M343
STM2	All	428,846	E1b1a1a1-M180
	Pre-capture	19,994	E1b1a1-M2
	YCC	220,732	E1b1a1a1-M180
	WGC	109,411	E1b1a1-M2
	WGC+Y	219,152	CT-M168
PI174	All	3,224	A1-V168
	Pre-capture	129	-
	YCC	1993	A1-V168
	WGC	147	-
	WGC+Y	1,818	A1-V168
PI383	All	7,738	P-M45
	Pre-capture	146	-
	YCC	1809	P-M45
	WGC	473	-
	WGC+Y	6,890	P-M45
PI435	All	16,469	BT-M42
	Pre-capture	164	-
	YCC	1,918	BT-M42
	WGC	2,632	-
	WGC+Y	12,399	BT-M42
PI437	All	3,444	-
	Pre-capture	103	-
	YCC	938	-
	WGC	296	-
	WGC+Y	2,320	-

Discussion

In this paper, we have described the efficiency of Y-DNA recovery from in-solution Y-chromosome capture-enrichment experiments and from WGC followed by Y-chromosome enrichment on aDNA libraries obtained from the archaeological remains of six males excavated in the Caribbean and dating between 300 and 2,000 years old. The experimental design involved the targeted enrichment of 10.3 Mb of

the Y chromosome on both standard and WGC libraries. As several aspects of the library preparation and WGC enrichment protocols differed between sample groups, we evaluated the success of the enrichment using within-sample comparisons. Overall, both approaches succeed in increasing the proportion of on-target sequences as compared to pre-capture libraries. For both enrichment approaches and for all samples, we observed that most of the sequenced reads mapped to the targeted regions.

Factors influencing enrichment

Although we observed consistent enrichment of Y-DNA on aDNA libraries, the marked differences of the performance between samples (STM and PI) and library types (standard vs. WGC libraries) is mostly like due to differences in starting endogenous content, read-length distributions, and the complexities of the libraries. Although it is now possible to increase the endogenous content of poorly preserved tissues, it has been shown that in-solution capture enrichment techniques perform better on samples with starting endogenous content greater than 1% and with little clonality [15]. Thus, as expected, the PI sample enrichment levels were systematically lower than those in the STM samples. Low complexity levels in the starting libraries also hamper the success of capture experiments, as these protocols usually involve an amplification step, which further increases the clonality. In addition, the enriched libraries were subject to an increase in the fragment length most likely driven by the probe length (105 bp) (Figure 5). However, the shift was more pronounced for the PI samples, indicating that a substantial proportion of the shorter fragments in the pre-capture and WGC libraries was not retrieved in the YCC and WGC+Y. Together, these observations provide useful insights as to the features in standard libraries that should be considered when planning capture experiments, thereby opening avenues for investigating ways to optimize these protocols.

Subsequently, we noted that for the STM samples the YCC experiments yielded higher fold-enrichments of the targeted regions than did the WGC+Y captures. In contrast, for the PI samples, the WGC+Y performed better. While we believed the differences owe mostly to the endogenous content, it could also be in part due to the techniques employed to enrich the whole genomes of the STM and PI samples (MYbaits and WISC, respectively). Indeed, although based on the same molecular principle, those techniques have slight differences in their performances [15]. However, despite the slightly higher enrichment rates in WGC+Y libraries, as compared to YCC, for the PI libraries, we do not recommend implementing consecutive enrichments, as we show here that doing so considerably reduces the complexity of the libraries [39] and, consequently, the ability to identify informative SNPs for haplogroup assignment.

Implications of Y-haplogroups assignments for the samples

Finally, the haplogroup inference was effective only for the two STM samples, for which we recovered at least 0.36 Mb of the Y-chromosome (3.5% of on-target regions). The STM1 individual bore the M343 mutation characteristic of haplogroup R1b, consistent with the back-to-Africa R1b1c-V88 haplogroup reported in [22];

however, we did not observe any of the SNPs specific to the V88 branch. Whereas for STM2, we identified a common and widespread African haplogroup characteristic of the Bantu expansion [40], E1b1a1a1-M80, consistent with the results from the analyses of the autosomal chromosomes [22]. For the remaining individuals, we could not resolve haplogroups due to the low depth at which they were sequenced due to limited budget, although our results might be impacted by the paucity of Y-SNPs that define the tips of the Native American haplogroups versus haplogroups from other well-characterized populations in the database employed. For example, no Puerto Rican males in the 1000 Genomes dataset bore Native American Y haplogroups. Rather, all possessed European or African lineages, primarily belonging to R1b and E1b clades [4], reflecting sex-biased admixture patterns during European colonization of the island (and reproduced across the Americas) [41]. This fact highlights the need to assay ancient genetic variation among pre-contact Native American samples, for which methods such as ours can be of important advantage. It is often challenging to recover DNA for such samples and the enrichment method we discuss here would certainly help in those cases.

Conclusions

In the past decade, new technologies and protocol improvements have emerged to efficiently recover ancient DNA. However, the endogenous DNA fraction continues to be a limiting step in ancient genomics studies. The first efforts to overcome this limitation have focused on targeting the mtDNA, because it is relatively short (~16 kB), and it is present in multiple copies per cell, unlike the autosomes (two copies) and the Y chromosome (one copy). For the Y chromosome, targeted enrichment strategies are more problematic due to its richness in repetitive and palindromic sequences. For the same reasons, Y-chromosome content is relatively poor in WGC studies, although WGC is becoming a cost-effective alternative for ancient genomics. Therefore, we used previously reported high-quality regions to capture the most phylogenetically informative portion of the Y chromosome. We confirmed the effectiveness of the method by noting that, after capture, up to 99.1% of the reads mapping to the Y chromosome fall within the targeted regions. We observed that YCC and WGC+Y libraries outperformed pre-capture libraries with respect to Y-DNA content. Whether it is advantageous to first perform WGC before enriching for Y-DNA seems to depend on the level of endogenous DNA content, the complexities and fragment lengths of the starting libraries. In this study, libraries with endogenous DNA content greater than 0.12% yielded ~10-fold greater enrichment rates under YCC, as compared to WGC+Y libraries, whereas in the four samples with low endogenous DNA proportions (0.01% to 0.02%), we observed a greater enrichment for WGC+Y experiments, but at the cost of higher clonality. We thus stress the need to consider the initial complexity, endogenous DNA content, and read lengths when planning these experiments. We recommend a design that includes the estimation of predictive yield and enrichment curves, based on shallow sequencing, to inform the best sequencing strategy and avoid sequencing beyond saturation.

Finally, there is a vast potential to incorporate Y-chromosome information from aDNA samples into the study of human population history from regions beyond

Eurasia. In our work, we go beyond SNP capture and present the first instance of Y-chromosome capture on ancient samples, opening new avenues of research to improve the performance of these experiments and to extract Y-chromosome information from ancient samples.

Availability of data and material

YCC and WGC+Y reads of STM samples are available in the European Nucleotide Archive under the accession number PRJEB23498. Y-chromosome alignments for Paso del Indio samples are available in the NCBI Short Read Archive (SRA) under BioProject PRJNA419010.

Competing interests

The authors declare no competing interests.

Funding

MCAA's laboratory is supported by Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica - Universidad Nacional Autónoma de México grant IA206817. ASM and DICD would like to thank the SNFS and the ERC for funding. HS was funded by the European Research Council (FP7/2007-2013, grant no. 319209, Synergy project NEXUS1492). The work on the samples from Saint Martin was partially funded by the European Commission through the Marie Curie Actions (FP7/2007-2013, grant no. 290344, EUROTAST). Funding for work with the Puerto Rico samples was provided by the Arizona State University School of International Letters and Cultures Foster Latin American Studies Support Grant, the Arizona State University Graduate and Professional Student Association and Sigma-Xi.

Author's contributions

MCAA conceived the project with input from CDB, DP and AS. MNC, AS and HS performed laboratory work. ASM, DICD and MCAA designed the data analysis strategy. DICD performed most data analyses with input from DGP. DICD and MCAA wrote the manuscript with input from all the authors. All authors read and approved the final manuscript.

Acknowledgements

Calculations were performed on UBELIX (<http://www.id.unibe.ch/hpc>), the HPC cluster at the University of Bern. MNC would like to thank William J. Pestle, L. Antonio Curet and Edwin Crespo-Torres for providing the Paso del Indio samples and Meredith Carpenter, Morten Rasmussen and Rosa Fregel for assistance with computational analyses.

Author details

¹Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland. ²International Laboratory for Human Genome Research, National Autonomous University of Mexico, Boulevard Juriquilla 3001, 762330 Juriquilla, Mexico. ³Department of Computational Biology, University of Lausanne, Quartier Sorge, Batiment Génopode, 1015 Lausanne, Switzerland. ⁴School of Human Evolution and Social Change, Arizona State University, 900 S. Cady Mall, 85284 Tempe, United States of America. ⁵Department of Genetics, Stanford University, Mail Stop-5120, 94305 Stanford, United States of America. ⁶23andMe, Mountain View, 94043 California, United States of America. ⁷Centre for GeoGenetics, University of Copenhagen, Øster Voldgade, 1350 Copenhagen, Denmark. ⁸Faculty of Archaeology, Leiden University, Einsteinweg, 2333 Leiden, Netherlands. ⁹Institute of Human Origins, Arizona State University, 951 South Cady Mall, 85287 Tempe, United States of America. ¹⁰Department of Biomedical Data Science, Stanford University, Welch Road, 94305 Stanford, United States of America. ¹¹Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Génopode, 1015 Lausanne, Switzerland.

References

1. Jobling, M.A., Tyler-Smith, C.: Human Y-chromosome variation in the genome-sequencing era. *Nature Reviews Genetics* **18**(8), 485–497 (2017)
2. Poznik, G.D., Henn, B.M., Yee, M.-C., Sliwierska, E., Euskirchen, G.M., Lin, A.a., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.a., Bustamante, C.D.: Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science (New York, N.Y.)* **341**(6145), 562–5 (2013)
3. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., Fu, Q., Mittnik, A., Bánffy, E., Economou, C., Francken, M., Friederich, S., Pena, R.G., Hallgren, F., Khartanovich, V., Khokhlov, A., Kunst, M., Kuznetsov, P., Meller, H., Mochalov, O., Moiseyev, V., Nicklisch, N., Pichler, S.L., Risch, R., Rojo Guerra, M.a., Roth, C., Szécsényi-Nagy, A., Wahl, J., Meyer, M., Krause, J., Brown, D., Anthony, D., Cooper, A., Alt, K.W., Reich, D.: Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**(7555), 207–211 (2015). doi:10.1038/nature14317. 1502.02783
4. Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., Wilson Sayres, M.A., Ayub, Q., McCarthy, S.A., Narechania, A., Kashin, S., Chen, Y., Banerjee, R., Rodríguez-Flores, J.L., Cerezo, M., Shao, H., Gymrek, M., Malhotra, A., Louzada, S., Desalle, R., Ritchie, G.R.S., Cerveira, E., Fitzgerald, T.W., Garrison, E., Marcketta, A., Mittelman, D., Romanovitch, M., Zhang, C., Zheng-Bradley, X., Abecasis, G.R., McCarroll, S.A., Flicek, P., Underhill, P.A., Coin, L., Zerbino, D.R., Yang, F., Lee, C., Clarke, L., Auton, A., Erlich, Y., Handsaker, R.E., Bustamante, C.D., Tyler-Smith, C.: Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics* **12**(9), 809–809 (2016). doi:10.1038/ng.3559
5. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., Connell, S., Stewardson, K., Harney, E., Fu, Q., Gonzalez-Fortes, G., Jones, E.R.,

- Roodenberg, S.A., Lengyel, G., Bocquentin, F., Gasparian, B., Monge, J.M., Gregg, M., Eshed, V., Mizrahi, A.-S., Meiklejohn, C., Gerritsen, F., Bejenaru, L., Blüher, M., Campbell, A., Cavalleri, G., Comas, D., Froguel, P., Gilbert, E., Kerr, S.M., Kovacs, P., Krause, J., McGettigan, D., Merrigan, M., Merriwether, D.A., O'Reilly, S., Richards, M.B., Semino, O., Shamoon-Pour, M., Stefanescu, G., Stumvoll, M., Tönjes, A., Torroni, A., Wilson, J.F., Yengo, L., Hovhannysian, N.A., Patterson, N., Pinhasi, R., Reich, D.: Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**(7617), 419–424 (2016)
6. Ho, S.Y.W., Gilbert, M.T.P.: Ancient mitogenomics. *Mitochondrion* **10**(1), 1–11 (2010)
7. Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.M., Poinar, H.N.: Ancient whole genome enrichment using baits built from modern dna. *Molecular Biology and Evolution* **31**(5), 1292–1294 (2014). doi:10.1093/molbev/msu074
8. Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J., Pääbo, S.: DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences of the United States of America* **110**(6), 2223–7 (2013)
9. Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora, M., Rasmussen, M., Gravel, S., Guillén, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D., Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y., Wang, J., Gilbert, M.T.P., Willerslev, E., Greenleaf, W.J., Bustamante, C.D.: Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. *American journal of human genetics* **93**(5), 852–64 (2013). doi:10.1016/j.ajhg.2013.10.002
10. Maricic, T., Whitten, M., Pääbo, S.: Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**(11), 9–13 (2010). doi:10.1371/journal.pone.0014004
11. Burbano, H.A., Hodges, E., Green, R.E., Briggs, A.W., Krause, J., Meyer, M., Good, J.M., Maricic, T., Johnson, P.L.F., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert, F.W., Rasilla, M.D., Fortea, J., Rosas, A., Lachmann, M., Hannon, G.J., Pääbo, S.: Targeted Investigation of the Sequence Capture. *Science* **723**, 723–726 (2010). doi:10.1126/science.1188046
12. Pajmians, J.L.A., Fickel, J., Courtiol, A., Hofreiter, M., Förster, D.W.: Impact of enrichment conditions on cross-species capture of fresh and degraded DNA. *Molecular Ecology Resources* (2015). doi:10.1111/1755-0998.12420
13. Avila-Arcos, M.C.M.C., Cappellini, E., Romero-Navarro, J.A., Wales, N., Moreno-Mayar, J.V.V., Rasmussen, M., Fordyce, S.L., Montiel, R., Vielle-Calzada, J.-P., Willerslev, E., Gilbert, M.T.P.: Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.* **1**, 74 (2011). doi:10.1038/srep00074
14. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., Meyer, M.: Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **110**(39), 15758–63 (2013)
15. Ávila-Arcos, M.C., Sandoval-Velasco, M., Schroeder, H., Carpenter, M.L., Malaspina, A.-S., Wales, N., Peñalosa, F., Bustamante, C.D., Gilbert, M.T.P.: Comparative performance of two whole-genome capture methodologies on ancient DNA Illumina libraries. *Methods in Ecology and Evolution* **6**(6), 725–734 (2015). doi:10.1111/2041-210X.12353
16. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., Skoglund, P., Lazaridis, I., Sankararaman, S., Fu, Q., Rohland, N., Renaud, G., Erlich, Y., Willems, T., Gallo, C., Spence, J.P., Song, Y.S., Poletti, G., Balloux, F., van Driem, G., de Knijff, P., Romero, I.G., Jha, A.R., Behar, D.M., Bravi, C.M., Capelli, C., Hervig, T., Moreno-Estrada, A., Posukh, O.L., Balanovska, E., Balanovsky, O., Karachanak-Yankova, S., Sahakyan, H., Toncheva, D., Yepiskoposyan, L., Tyler-Smith, C., Xue, Y., Abdullah, M.S., Ruiz-Linares, A., Beall, C.M., Di Rienzo, A., Jeong, C., Starikovskaya, E.B., Metspalu, E., Parik, J., Vilems, R., Henn, B.M., Hodoglugil, U., Mahley, R., Sajantila, A., Stamatiyannopoulos, G., Wee, J.T.S., Khushainova, R., Khushnutdinova, E., Litvinov, S., Ayodo, G., Comas, D., Hammer, M.F., Kivisild, T., Klitz, W., Winkler, C.A., Labuda, D., Bamshad, M., Jorde, L.B., Tishkoff, S.A., Watkins, W.S., Metspalu, M., Dryomov, S., Sukernik, R., Singh, L., Thangaraj, K., Pääbo, S., Kelso, J., Patterson, N., Reich, D.: The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**(7624), 201–206 (2016)
17. Schlebusch, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Steyn, M., Soodyall, H., Lombard, M., Jakobsson, M.: Ancient genomes from southern Africa pushes modern human divergence beyond 260,000 years ago. *bioRxiv* (2017)
18. Kemp, B.M., Malhi, R.S., McDonough, J., Bolnick, D.A., Eshleman, J.A., Rickards, O., Martinez-Labarga, C., Johnson, J.R., Lorenz, J.G., Dixon, E.J., Fifield, T.E., Heaton, T.H., Worl, R., Smith, D.G.: Genetic analysis of early holocene skeletal remains from Alaska and its implications for the settlement of the Americas. *American Journal of Physical Anthropology* **132**(4), 605–621 (2007)
19. Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., Poznik, G.D., Gudmundsdottir, V., Yadav, R., Malaspina, A.-S., White, S.S., Allentoft, M.E., Cornejo, O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S., Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V.M., Lopes, M.C., Malhi, R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F., Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R., Willerslev, E., V, S.S.W., Allentoft, M.E., Cornejo, O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S., Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V.M., Lopes, M.C., Malhi, R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F., Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R., Willerslev, E.: The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**(7487), 225–229 (2014)
20. Rasmussen, M., Sikora, M., Albrechtsen, A., Korneliussen, T.S., Moreno-Mayar, J.V., Poznik, G.D., Zollikofer, C.P.E., Ponce de León, M.S., Allentoft, M.E., Moltke, I., Jónsson, H., Valdiosera, C., Malhi, R.S., Orlando, L., Bustamante, C.D., Stafford, T.W., Meltzer, D.J., Nielsen, R., Willerslev, E.: The ancestry and affiliations of Kennewick Man. *Nature*, 1–10 (2015)

21. Gallego Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D.G., Bhak, J., Pinhasi, R., Manica, A.: Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science (New York, N.Y.)* **350**(6262), 820–2 (2015)
22. Schroeder, H., Ávila-Arcos, M.C., Malaspina, A.-S., Poznik, G.D., Sandoval-Velasco, M., Carpenter, M.L., Moreno-Mayar, J.V., Sikora, M., Johnson, P.L.F., Allentoft, M.E., Samaniego, J.A., Havisser, J.B., Dee, M.W., Stafford, T.W., Salas, A., Orlando, L., Willerslev, E., Bustamante, C.D., Gilbert, M.T.P.: Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *Proceedings of the National Academy of Sciences* **112**(12), 201421784 (2015)
23. Pestle, W.J., Colvard, M.: Bone collagen preservation in the tropics: A case study from ancient Puerto Rico. *Journal of Archaeological Science* **39**(7), 2079–2090 (2012)
24. Pestle, W.J.: Diet and Society in Prehistoric Puerto Rico: An Isotopic Approach. Unpublished ph.d. dissertation, University of Illinois at Chicago (2010)
25. Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest, S., Coombes, B.K., Wood, J.W., Earn, D.J.D., White, W., Krause, J., Poinar, H.N.: Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proceedings of the National Academy of Sciences* **108**(38), 746–752 (2011)
26. Gilbert, M.T.P., Bandelt, H.J., Hofreiter, M., Barnes, I.: Assessing ancient DNA studies. *Trends in Ecology and Evolution* **20**(10), 541–544 (2005). doi:10.1016/j.tree.2005.07.005
27. Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., Lawlor, R.T., Scarpa, A.: DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples. *PLoS ONE* **8**(6) (2013)
28. Meyer, M., Kircher, M.: Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols* **2010**(6), 5448 (2010). doi:10.1101/pdb.prot5448
29. Kircher, M., Sawyer, S., Meyer, M.: Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* **40**(1), 1–8 (2012)
30. Schubert, M., Lindgreen, S., Orlando, L.: AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes* **9**(1), 88 (2016)
31. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009)
32. Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., Orlando, L.: Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**(5), 1056–1082 (2014)
33. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P.L.F., Orlando, L.: MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**(13), 1682–1684 (2013)
34. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016). R Foundation for Statistical Computing. <https://www.R-project.org/>
35. Korneliusson, T.S., Albrechtsen, A., Nielsen, R.: ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**(1), 356 (2014)
36. Skoglund, P., Storå, J., Götherström, A., Jakobsson, M.: Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**(12), 4477–4482 (2013)
37. Deng, C., Daley, T., Smith, A.D.: Applications of species accumulation curves in large-scale biological data analysis. *Quantitative biology (Beijing, China)* **3**(3), 135–144 (2015)
38. Cruz-Dávalos, D.I., Llamas, B., Gaunitz, C., Fages, A., Gamba, C., Soubrier, J., Librado, P., Seguin-Orlando, A., Pruvost, M., Alfarhan, A.H., Alquraishi, S.A., Al-Rasheid, K.A.S., Scheu, A., Beneke, N., Ludwig, A., Cooper, A., Willerslev, E., Orlando, L.: Experimental conditions improving in solution target enrichment for ancient DNA. *Molecular ecology resources* (2016). doi:10.1111/1755-0998.12595
39. Enk, J., Rouillard, J.M., Poinar, H.: Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. *BioTechniques* **55**(6), 300–309 (2013)
40. Rowold, D., Garcia-Bertrand, R., Calderon, S., Rivera, L., Benedico, D.P., Alfonso Sanchez, M.A., Chennakrishnaiah, S., Varela, M., Herrera, R.J.: At the southeast fringe of the Bantu expansion: Genetic diversity and phylogenetic relationships to other sub-Saharan tribes. *Meta Gene* **2**, 670–685 (2014)
41. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., Eng, C., Sandoval, K., Acevedo-Acevedo, S., Norman, P.J., Layrisse, Z., Parham, P., Martínez-Cruzado, J.C., Burchard, E.G., Cuccaro, M.L., Martin, E.R., Bustamante, C.D.: Reconstructing the Population Genetic History of the Caribbean. *PLoS Genetics* **9**(11), 1003925 (2013)

Additional Files

Supplementary table 1. Summary of sequenced and mapped reads of the complete dataset

Supplementary table 2. Mean and standard errors regarding reads from down-sampled libraries.

Supplementary figure 1. Depth of coverage across the Y-chromosome.

From top to bottom, rows depict the coverage levels for the pre-capture, YCC, WGC and WGC+Y conditions. Red boxes represent the targeted regions. Each blue point represents sequencing coverage within a 1,000-bp window, averaged across 10 subsampled replicates per sample per condition, explaining depths of coverage below 1. For visual purposes, we increased the opacity of the points in the PI samples.

Supplementary figure 2. Length distribution of mapped reads.

Length distributions of reads mapping to the whole genome. The length distribution was smoothed by fitting a polynomial curve to the observed frequencies; the ribbons correspond to 95% confidence intervals.

Supplementary figure 3. Expected yield and on-target fold-enrichment.

Dashed lines indicate the number of down-sampled reads. (A-F): Predicted median value and variance (across 100 bootstrap replicates) of the number of on-target reads, as a function of total sequenced reads. The points depict the observed numbers of on-target reads in the down-sampled libraries. (G-L): Expected enrichment of on-target reads versus number of sequenced reads for each condition and each sample.