Title: Comparative Performance of Two Whole Genome Capture Methodologies on Ancient DNA Illumina Libraries

Short title: Whole genome capture of aDNA Illumina libraries

Word count: 6819

María C. Ávila-Arcos[1,2*], Marcela Sandoval-Velasco[2], Hannes Schroeder[2,3], Meredith L. Carpenter[1], Anna-Sapfo Malaspinas[2], Nathan Wales[2], Fernando Peñaloza[2,4], Carlos D. Bustamante[1], M Thomas P Gilbert[2*]

[1] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[2]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark

[3]Faculty of Archaeology, Leiden University, PO Box 9515, 2300 Leiden, The Netherlands

[4]Undergraduate Program on Genomic Sciences, Universidad Nacional Autónoma de México, 62210, Cuernavaca, Morelos, Mexico

*Corresponding Authors

M Thomas P Gilbert

Natural History Museum of Denmark

Øster Voldgade 5-7

1350 København K

Phone: +45 23712519

Email: mtpgilbert@gmail.com

María C. Ávila-Arcos

Department of Genetics,

Stanford University

365 Lasuen St. Littlefield Center MC 2069

Stanford CA 94305

Phone: +1 650 723 2643

Email: maricugh@gmail.com

Abstract

1. The application of whole genome capture (WGC) methods to ancient DNA (aDNA) promises to increase the efficiency of ancient genome sequencing.
2. We compared the performance of two recently developed WGC methods in enriching human aDNA within Illumina libraries built using both double-stranded (DSL) and single-stranded (SSL) build protocols. Although both methods effectively enriched aDNA, one consistently produced marginally better results, giving us the opportunity to further explore the parameters influencing WGC experiments.
3. Our results suggest that bait length has an important influence on library enrichment. Moreover, we show that WGC biases against the shorter molecules that are enriched in SSL preparation protocols. Therefore application of WGC to such samples is not recommended without future optimization. Lastly, we document the effect of WGC on other features including clonality, GC composition and repetitive DNA content of captured libraries.
4. Our findings provide insights for researchers planning to perform WGC on aDNA, and suggest future tests and optimization to improve WGC efficiency.

INTRODUCTION

The introduction of next-generation sequencing (NGS) marked a dramatic turning point in aDNA investigation, enabling the study of genome scale datasets (e.g. Green *et al.* 2010, Meyer *et al.* 2012, Orlando *et al.* 2013, Rasmussen *et al.* 2014). Nevertheless, the DNA quality within most archaeological samples continues to hamper the field's development. Specifically, the fragmented and damaged nature of aDNA molecules, coupled with high levels of exogenous contaminant DNA, have required investment of significant resources in order to enable generation of meaningful levels of sequence data (Knapp & Hofreiter 2014). In response to this challenge, several key methodological improvements have been developed. These include optimization of DNA extraction and library preparation protocols to better retain and incorporate damaged endogenous DNA (e.g. Dabney *et al*. 2013, Gansauge & Meyer 2013, respectively), and the introduction of large-scale hybridization-capture based targeted enrichment techniques optimized for aDNA (e.g. Briggs *et al*., 2009, Burbano *et al.,* 2010, Maricic *et al.* 2010, Schuenemann *et al*, 2011, Ávila-Arcos *et al,* 2011, Fu *et al.* 2013).

Until recently, targeted enrichment approaches were limited to relatively small fractions of predefined regions of an already characterized genome, (e.g. Maricic *et al.* 2010, Schuenemann *et al.* 2013, Fu *et al.* 2013). However, a method introduced by Carpenter *et al.* (2013), termed Whole-genome In-Solution Capture (WISC), showed considerable genome-wide enrichment of human aDNA sequencing libraries with very low initial concentrations of endogenous DNA. In parallel, a

commercial method, MYbaits-WGE (MYbaits Whole Genome Enrichment, henceforth referred to as "MYbaits"), was developed by the company MYcroarray (Ann Arbor, MI, USA) following a similar methodological principle (Gnirke *et al.* 2009) but with some subtle differences in the protocol (for an application of the method see Enk *et al.,* 2014). In broad terms, the approach involves the construction of an RNA 'bait' library by transcribing modern genomic DNA that has been fragmented and ligated to T7 RNA polymerase promoters. The synthesis of RNA from the T7-ligated modern reference genomic DNA is carried out using biotinylated UTPs, and the RNA products are then hybridized in solution to aDNA libraries. The hybridized fragments can then be retrieved using streptavidin-coated magnetic beads, while the unbound molecules are washed away. Captured library molecules can subsequently be amplified and sequenced (Carpenter *et al.* 2013, Enk *et al.* 2014).

Although the fundamental molecular principle of both methods is similar, it is unclear if they performed equally well, and as such, whether there is a user benefit in employing one over the other. To explore this issue, and to better understand the parameters affecting the success of capture experiments, we enriched and sequenced a series of ancient human DNA libraries using both WISC and MYbaits. Using the data, we describe the effect that different bait lengths and hybridization times have on the resulting fold enrichment, when applying each protocol to aDNA libraries with variable levels of initial endogenous DNA content. To a lesser extent we also explore the potential role of differences in hybridization times. Furthermore, we investigate the performance of both whole genome capture (WGC) methods on single-stranded libraries (SSL), which have been previously

shown to contain higher levels of endogenous DNA than standard double-stranded libraries (DSL) (Meyer *et al.,* 2012, Gansauge & Meyer, 2013). We examined the effect on fold enrichment of applying WGC methods to this particular type of libraries.

MATERIALS AND METHODS

DNA extraction and aDNA library preparation

We generated sequence data from DNA extracted from eight archeological human skeletal samples originating from a range of different archaeological contexts and environmental conditions, dated to between 300 and 2500 years BP (Table 1). We deliberately chose samples from different contexts and with variable amounts of endogenous DNA as determined through shotgun sequencing (0.2-8.0%) to assess the performance of the two different WGC methods on samples of different quality. All DNA extraction and library preparation steps were performed in dedicated clean laboratories at the Centre for GeoGenetics in Copenhagen, Denmark, to prevent contamination with modern DNA. Before extraction, the tooth samples were cleaned with 10% bleach solution and then UV-irradiated for 2 min on each side to cross-link potentially contaminant DNA to the surface. Part of the tooth root was then excised and the inside of the tooth was drilled to produce approximately 200 mg of powder. The powder was digested in 5 ml of an EDTA-based digestion buffer containing 0.25 mg/mL Proteinase K. DNA was then isolated using a silica-based

123 extraction method (Rohland and Hofreiter 2007). Samples were eluted in 60 µl TET
124 buffer.
125
126 Following extraction, the DNA was divided into two aliquots of 30 µl. Each of these
127 were built into Illumina libraries, using a double- and single stranded protocol,
128 respectively. The single-stranded libraries were built following a previously
129 published protocol (Gansauge & Meyer, 2013) but without first removing
130 deoxyuracils. The double-stranded libraries were prepared using a blunt-end
131 library preparation kit from NEB (E6070) and blunt-end modified Illumina adapters
132 (Meyer and Kircher, 2010). The libraries were prepared according to the
133 manufacturer's instructions, although with minor modifications as outlined below.
134 The initial nebulization step was skipped because of the fragmented nature of aDNA.
135 End-repair was performed in 50 µl reactions with 30 µl of DNA extract. The end-
136 repair cocktail was incubated for 20 min at 12°C and 15 min at 37°C and purified
137 using Qiagen MinElute silica spin columns following manufacturer's instructions
138 and eluted in 30 µl. After end-repair, Illumina-specific adapters (Meyer and Kircher,
139 2010) were ligated to the end-repaired DNA in 50 µl reactions. The reaction was
140 incubated for 15 min at 20°C and purified using Qiagen QiaQuick columns before
141 being eluted in 30 µl EB. The adapter fill-in reaction was performed in a final volume
142 of 50 µl and incubated for 20 min at 37°C followed by 20 min at 80°C to inactivate
143 the *Bst* enzyme. Libraries were then amplified and indexed in a 50 µl PCR reaction,
144 using 15 µl of library template, 25 µl of a 2x KAPA U+ master mix, 5,5 µl H2O, 1,5 µl
145 DMSO, 1 µl BSA (20 mg/ml), and 1 µl each of a forward and reverse indexing primer

(10 µM). Thermocycling conditions were 5 min at 98°C, followed by 10-12 cycles of 15 sec at 98°C, 20 sec at 60°C, and 20 sec at 72°C, and a final 1 min elongation step at 72°C. The amplified libraries were then purified using Agencourt AMPure XP beads and eluted in 30 µl EB. Between 2-6 µl of the indexed DNA libraries were then quantified on an Agilent Bioanalyzer, pooled in equimolar amounts, and sequenced together with other samples on a HiSeq 2000 lane.

Whole Genome Capture

The remaining fraction of the libraries was subdivided, and re-amplified for 8-12 cycles using primers IS5 and IS6 (from Meyer & Kircher, 2010) and the same PCR conditions as above to obtain a minimum of 100 ng of library template. The libraries were then captured using the two different WGC methods. WISC, which makes use of homemade RNA probes, was carried out as described in Carpenter *et al.* (2013), with a bait-library hybridization time of 66 hours. For the MYbaits capture, a human whole genome enrichment kit MYbaits-HuWGE (MYcroarray, Ann Arbor), which relies on pre-made RNA probes and adapter blockers built using proprietary protocols, was used following manufacturer's instructions with 24 hours for bait-library hybridization. Captured libraries were re-amplified with IS5 and IS6 for 10-20 cycles (depending on library and capture method, see Table 2 for total number of PCR cycles per sample) and using the same PCR conditions as above. Subsequently, the libraries were purified, quantified and pooled as described above, and submitted for sequencing on the HiSeq 2000.

Fastq filtering and mapping

Libraries were sequenced on an Illumina HiSeq2000 in 100-cycle single-end runs. The base calls were performed using the Illumina software CASAVA 1.8.2 and the output fastq files were processed with the following software and scripts. First, reads were filtered with AdapterRemoval (Lindgreen 2012) by trimming adapter sequences as well as N's and low quality stretches and discarding fragments shorter than 30 bp. Reads passing filters were then mapped using BWA Version: 0.7.5a-r405 (Li & Durbin 2009) to the hg18 reference genome, setting the minimum mapping quality to 25. Next, Samtools rmdup (Li *et al.* 2009) was used to remove PCR clones while reads mapping to more than one unique region in the genome were also excluded by controlling for the XT, XA and X1 tags in the bam alignment. Fragmentation and damage patterns were calculated and plotted using MapDamage2 (Ginolhac *et al.* 2011, Jonsson *el at.* 2013).

Contamination estimates

To estimate the contamination fraction for each experiment, we mapped the data to the whole nuclear genome (hg18) as well as to a consensus mitochondria (mt) sequence that was generated separately for each sample [STM1, STM2, STM3 (Schroeder et al, unpublished data)]. We only retained reads that mapped to the consensus mt with a mapQ> 30, and excluded reads with potential alternative mapping coordinates to the nuclear genome by controlling for XT, XA and X0 tags in the bamfile. This has the effect of reducing the number of reads that map both to the mitochondrial genome and to nuclear copies of mitochondrial genes (numts). We

192 then used a method detailed in the Supplemental Information section 5 of (Fu *et al.*
193 2013) that generates a moment-based estimate of the error rate and a Bayesian-
194 based estimate of the posterior probability of the contamination fraction. We ran
195 three chains of 50,000 iterations for the Monte Carlo Markov Chain and discarded
196 the first 10,000, as was done in (Fu *et al.* 2013). We assessed convergence of the
197 chain by visualizing the potential scale reduction factor (PSRF) and verifying that
198 the median of PSRF is below 1.01 for all cases (Gelman & Rubin 1992, Plummer *et al.*
199 2006).
200
201 **Length distribution, repeat and GC content analysis**
202 Bam files were intersected with the hg18 UCSC 50mer mapability tracks
203 (downloaded from
204 ftp://hgdownload.cse.ucsc.edu/gbdb/hg18/bbi/crgMapabilityAlign50mer.bw)
205 using bedtools intersect (version 2.15.0, https://github.com/arq5x/bedtools) and
206 requiring at least 50% of the read overlapping with a region with a mapability value
207 of 0.5 or less. Unix scripts were used to calculate the GC content fractions and length
208 distribution. Plots were generated using Rstudio (http://www.rstudio.org/).
209
210 RESULTS
211
212 Shotgun sequencing of eight ancient human DNA libraries revealed pre-capture
213 endogenous content ranging between 0.2% and 8% (Table 1). Prior to sequencing,
214 human genomic DNA was enriched for two of the libraries using the MYbaits

215 protocol and for six of the libraries using both the WISC and MYbaits protocols.
216 Additionally, for three of the samples (samples STM1-3), SSL were built from the
217 same DNA extract used to build the DSL. These three SSL were shotgun sequenced
218 and also subjected to both capture schemes and sequenced on the HiSeq 2000
219 (Table 1). Sequenced reads that mapped to the human genome reference showed
220 the damage and fragmentation patterns characteristic of aDNA (Briggs *et al.* 2007)
221 and were consistent with the type of library build method used to generate them
222 (Supplementary figures S1-4).
223
224 **Enrichment rates on DSL**
225 To evaluate the performance of capture experiments, we computed the percentage
226 of reads in each of the sequenced libraries that matched the human genome
227 reference sequence, as well as the percentage of non-clonal, uniquely mapped reads
228 that matched the reference. To estimate the fold enrichment for each sample, we
229 compared the unique endogenous fraction of the captured libraries with that of the
230 shotgun libraries (Ávila-Arcos *et al.*, 2011).
231 Regardless of the capture method, an increase of the endogenous content
232 was consistently observed in all captured DSL libraries when compared to their pre-
233 capture counterpart (Table 3). Informative rates of enrichment ranged between 1.9
234 and 8.6 fold. Although the sequences generated from the WISC libraries initially
235 exhibited a higher fraction of human DNA, this derived from higher levels of clonal
236 reads (Table 2), and the actual informative (unique) rate of enrichment was
237 consistently higher for MYbaits experiments. Fold enrichment values ranged

238 between 2.9 and 8.6 (mean 5.2) for MYbaits libraries and between 1.9 and 5.6
239 (mean 3.6) for WISC (Table 3), consistent with reported values (2 to 13 fold) in
240 Carpenter *et al.* (2013).

241 In contrast to previous observations (Carpenter *et al.* 2013; Enk *et al*. 2013)
242 that lower initial endogenous DNA concentrations result in higher enrichment, we
243 did not observe a clear dependence of enrichment rates on initial endogenous
244 values. For example, among the pre-capture libraries, unique endogenous content
245 values ranged between 0.2% and 8.0 %, and for these extreme values the fold
246 enrichment was very similar (4.1 fold and 3.7 fold, respectively, Table 3). This
247 discrepancy could also be attributed to the small size of our dataset.

248

249 **Bait length influences the read length distribution of captured reads**

250 The endogenous DNA sequences generated from captured libraries were
251 consistently longer than those in shotgun libraries (Fig. 1a and 1b and
252 Supplementary Figs. S5-6), in agreement with observations in previous WGC studies
253 (Carpenter *et al.* 2014; Enk *et al.* 2014). Furthermore, libraries captured with WISC
254 showed stronger bias against short fragments, and consequently a higher
255 proportion of longer reads, than those enriched with MYbaits. To investigate if any
256 physical properties of the baits could explain this discrepancy, we plotted the
257 distribution of both bait libraries (see Materials and Methods) and observed the
258 distribution of WISC probes was wider than MYbaits, with a higher fraction of the
259 former exceeding 500 bp (Fig. 1c), whereas MYbaits had a narrower distribution
260 between roughly 100bp and 500bp.

Effect of hybridization times

Although one might intuitively expect a correlation between incubation time and hybridization success, we observed no such behavior. The time of hybridization in WISC experiments was 66 hours, versus 24 hours for MYbaits. Our results suggest that the increased time utilized in WISC did not have a clear positive effect in terms of informative enrichment; however, we did not directly compare hybridization times between experiments performed with the same protocol (i.e., WISC for 24 hrs. vs. 66 hrs.).

Clonality

The effect of capture on levels of sequence clonality was consistent with previous reports (Ávila-Arcos *et al.*, 2011). Firstly, sequence clonality (measured as the fraction of the total mapped reads that are clonal duplicates) in captured libraries was consistently higher than in pre-capture libraries. Secondly, the lower the endogenous content in the shotgun library, the higher the clonality in the post-capture ones. Furthermore, our results showed a logarithmic decrease of clonality with initial endogenous content (Table 3, Supplementary Fig. S7). In order to achieve the amounts of DNA required for WGC (both protocols requiring the same initial 100-500 ng), all libraries required two rounds of amplification prior to capture, and a further round of amplification post capture (see Materials and Methods). In fact, the amount of library DNA used as input for WISC was lower than

283 that used for MYbaits (Table 2), which made necessary additional PCR cycles for the
284 former in order to reach optimal concentrations for sequencing. As a result,
285 WISC libraries consistently exhibited higher levels of clonal reads than MYbaits
286 (Table 2).
287
288 **Whole Genome Capture on SSL**
289 An unexpected observation was that, despite the SSL having higher endogenous
290 DNA contents prior to capture than the DSL, both capture methods performed
291 poorly on the SSL, with endogenous contents increasing only marginally and even
292 decreasing relative to pre-capture values (Table 2). This poor performance of both
293 WGC methods on SSL can be explained by the fact that the biggest gain in SSL is due
294 to the recovery of short DNA fragments, which are then lost when WGC is performed
295 with the parameters assayed herein. Further optimization of these parameters (e.g.
296 hybridization temperature), could in principle overcome this limitation (see
297 Discussion).
298
299 **Contamination**
300 An important consideration when working with human aDNA is to control for
301 potential modern human contamination. Reads derived from such contaminating
302 DNA would intuitively be expected to have a longer average lengths than aDNA
303 reads; hence, we investigated whether the fact that capture biased toward longer
304 DNA molecules, resulted in higher contamination values in the captured libraries.
305 We estimated the mitochondria (mt) contamination values for the three samples for

306 which both DSL and SSL were subjected to WGC (see Material and Methods). We
307 observed less mt contamination in the post capture libraries, suggesting the longer
308 reads for those samples do not derive from a contaminant source. Although our
309 sample size is small, it is an encouraging result for the use of capture methods in
310 aDNA (Supplementary Table 1).
311
312 **GC content in WGC libraries**
313 Previous capture studies have reported differences in GC content between pre- and
314 post-capture libraries (Gnirke *et al.* 2009, Carpenter *et al.* 2013). We explored this
315 factor in our dataset by measuring the GC fraction in pre- and post-capture reads
316 and studying their distribution per experiment. We did not observe a specific
317 pattern of increase or decrease of average GC contents in post-capture libraries, but
318 instead a narrowing of the GC distribution in DSL libraries (Fig. 2a and
319 Supplementary Table 2). Average GC content ranged between 42-49% in pre-
320 capture DSL and between 44-46% and 43-45% in post-capture WISC and MYbaits
321 libraries, respectively (Supplementary Table 2). Interestingly we noticed that the
322 average GC of pre-capture SSL (35-38%) was lower that the than DSL, and that the
323 SSL GC distributions were shifted downwards, as shown in the histograms in Fig. 2c.
324 Remarkably, the GC content increased again when SSL were captured (Figs. 2b-c and
325 Supplementary Table 2). Post-capture SSL mean GC contents ranged between 42-
326 45% and 40-42% for WISC and MYbaits, respectively. We hypothesize that this bias
327 towards specific GC values in post-capture libraries, relates to the particular
328 selection of the hybridization temperatures and times in capture experiments.

329

330 **Repeat enrichment in WGC libraries**

331 Despite the inclusion of blocking agents (human Cot-1 and salmon sperm DNA) in
332 both WGC protocols to "mask" repetitive elements in the library and avoid bait
333 hybridization to these, a higher proportion of reads in post-capture libraries
334 mapped to repetitive regions in the genome (Fig. 3). In particular, we observed that
335 libraries captured with WISC displayed a higher fraction of repeats. To investigate if
336 this was tied to the biased selection of longer reads, we computed and plotted the
337 length distribution of reads within and outside repeats for each type of experiment
338 (Supplementary table 3). These distributions confirmed that reads within repeats
339 are on average longer in the pre-capture libraries, which probably drives these to
340 preferential capture over the non-repeat ones, as evidenced by an also longer
341 average length of reads within repeats in both post-capture library types
342 (Supplementary Table 3).

343

344 DISCUSSION

345

346 By comparing the performance of WISC and MYbaits in enriching for endogenous
347 human DNA in ancient DNA extracts, we have been able to pinpoint potential factors
348 influencing the dynamics of WGC experiments. The assessment of the subtle
349 differences between both approaches to in-WGC enables us to draw insights on two
350 variables that may the affect capture efficiency – bait length distribution and
351 hybridization time. Our data furthermore provides insights into the effect of

352 blocking agents, and first insights into the performance of whole-genome
353 enrichment methods on SSL.
354
355 Although the experimental design and parameters used in this study seem to
356 suggest an apparent benefit of one of the methods over the other, we strongly
357 caution that batch effects could be playing an important role under these settings,
358 hence discourage such interpretation from our results. Likewise, it is worth
359 considering that even though there is a certain convenience in using a pre-made kit
360 (MYbaits), our observations point to specific factors that can be optimized in the in-
361 house method (WISC) namely bait length distribution and hybridization
362 parameters. Knowing the relevance of such parameters in WGC, gives users the
363 flexibility of customizing their capture experiment to match the particularities of
364 each aDNA library (see below).
365
366 **Role of bait length distribution on the efficiency of WGC**
367 Bait length distributions (Fig. 1c) differ mainly in that WISC shows a wider range
368 and longer bait lengths. This in principle could account for the marked retrieval of
369 longer reads in the WISC compared with the MYbaits experiments (Figs. 1a-b).
370 Following this rationale, the higher success of the latter could be explained by its
371 ability to better access a fraction of the sequencing library, specifically that with the
372 smaller fragments, while this fraction remains inaccessible due to the higher
373 concentration of longer baits used in WISC. An important consequence of this
374 feature was the poor and even unsuccessful outcome of capturing SSL, which

include a higher fraction of short fragments, with either method. At the same time, this limitation reveals an important area for future development in the context of WGC experiments.

Although there was a small, yet consistent, benefit in the MYbaits over the WISC, it would be rash to conclude that the MYbaits method always outperforms WISC. These results were generated using a single batch of WISC bait versus a single batch of MYbaits. Given that (i) bait lengths will likely vary between batches as a result of initial template DNA fragmentation, and (ii) our hypothesis that bait length may play a key role in retaining shorter DNA fragments, we believe it is more than likely our results simply reflect the fact that in these batches tested the WISC bait were slightly longer than the MYbaits. Future studies that examine the role of bait length in capture success will be needed to further examine this hypothesis.

Preferential retrieval of repeats in WGC experiments

Despite the inclusion of blocking agents, both capture methods resulted in a higher proportion of reads mapping to repetitive regions in post-capture compared to the pre-capture libraries. This enrichment of repeats in capture experiments has also been observed previously in WGC experiments, despite the inclusion of organism-specific Cot-1 DNA in the capture reactions (Carpenter *et al.* 2013 and Enk *et al.* 2014). Some of the possible explanations, as described by Enk *et al.* (2014), include more rapid association rates for repetitive sequences and jumping PCR in post-capture amplification. Furthermore, it has been observed that, against intuition, in

some cases Cot-1 can enhance non-specific hybridization and hybridization to conserved repetitive elements (Newkirk *et al.* 2005). Also, batch-to-batch differences in the ability of commercial Cot-1 preparations to reduce hybridization to repetitive sequences have been reported (Carter *et al.* 2002). Our analyses show that another element to consider is the length of repetitive fragments, which we found to be longer than non-repetitive fragments and probably preferentially captured. In summary, considerations in this regard, along with further tests on the amount of this type of blockers, are needed in order to increase our understanding of their effects and to improve the capture efficiency.

Cost-benefit of WGC experiments

We have demonstrated that WGC, at least for the two currently available methods, is an effective way to enrich endogenous content in aDNA sequencing libraries. However, it is important to consider that despite the success of its application, WGC would only represent a practical and cost-effective advantage for whole genome sequencing from ancient samples when applied on libraries with endogenous contents above 1%. Assuming an average read length of 60 bp, approximately 200 lanes worth of sequencing on the HiSeq2000 would be required to reach 1x of the human genome from a library with 1% endogenous (assuming high complexity in the library); however, only 23 would be needed if the sample is enriched to 8.6 fold, which is the maximum enrichment value observed in this study. Although the cost of sequencing 23 lanes is still very high, depending on the research question, less than 1x genome coverage could be enough to provide valuable information about the

421 sample (e.g. Skoglund *et al*. 2012, Sánchez-Quinto *et al.*, 2012, Carpenter *et al.* 2013)
422 Therefore, samples with less than 1% endogenous could in principle also be
423 considered for WGC when the research question does not require information from
424 the complete genome. Ultimately, optimization of capture enrichment protocols
425 might enhance the efficiency and lower costs enough to make these samples with
426 very little endogenous DNA accessible to genomic scale investigations.
427
428 With these observations in hand, we believe that an *a priori* knowledge of the
429 endogenous content of a sample is crucial before deciding the enrichment strategy,
430 or if it is even worth the investment of a WGC reaction. An initial shotgun screening
431 can provide this information. Even with as few as 1 million reads (1/250 of a
432 HiSeq2000 lane), characteristics of the library such as complexity, read length
433 distribution and endogenous content can be retrieved and inform the selection of
434 the strategy to follow, whether it is WGC (for reads with ~1% endogenous, or even
435 less when just a fraction of the genome is informative, and a higher proportion of
436 long reads) or SSL (for very fragmented libraries). In fact, both strategies can be
437 complementary and applied in parallel (but not consecutively) to cover a wider
438 spectrum of read lengths. Additional optimization of WGC bait lengths and
439 hybridization parameters might eventually overcome this limitation and make
440 accessible the shorter fragments to WGC approaches.
441
442 In summary, we believe that our observations, though based on a small dataset and
443 more descriptive than exhaustive, can be of value for researchers planning

444 enrichment experiments. As new methods in aDNA are being developed at an
445 unprecedented pace, being able to discern the best approach for a given sample is of
446 the utmost relevance in order to take advantage of the methods and avoid wasting
447 resources and precious sample material. New extraction methods, library
448 preparation, sequencing and enrichment protocols are in their infancy in aDNA but
449 promise to unlock fascinating information from ancient samples at an unparalleled
450 scale. Consequently, it is important to have standards and guidelines to choose the
451 best approach for any given project.
452

453 REFERENCES

454

455 Ávila-Arcos, M.C., Cappellini, E., Romero-Navarro, J.A., Wales, N., Moreno-Mayar, J.V.,
456 Rasmussen, M., Fordyce, S.L., Montiel, R., Vielle-Calzada, J.P., Willerslev, E. & Gilbert,
457 M.T. (2011) Application and comparison of large-scale solution-based DNA capture-
458 enrichment methods on ancient DNA. Scientific reports, 1, 74.

459

460 Briggs, A.W., Good, J.M., Green, R.E., Krause, J., Maricic, T., Stenzel, U., Lalueza-Fox, C.,
461 Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Schmitz, R., Doronichev, V.B., Golovanova,
462 L.V., de la Rasilla, M., Fortea, J., Rosas, A. & Paabo, S. (2009) Targeted retrieval and
463 analysis of five Neandertal mtDNA genomes. Science, 325, 318-321.

464

465 Briggs, A.W., Stenzel, U., Johnson, P.L., Green, R.E., Kelso, J., Prufer, K., Meyer, M.,
466 Krause, J., Ronan, M.T., Lachmann, M. & Paabo, S. (2007) Patterns of damage in
467 genomic DNA sequences from a Neandertal. Proceedings of the National Academy of
468 Sciences of the United States of America, 104, 14616-14621.

469

470 Burbano, H.A., Hodges, E., Green, R.E., Briggs, A.W., Krause, J., Meyer, M., Good, J.M.,
471 Maricic, T., Johnson, P.L., Xuan, Z., Rooks, M., Bhattacharjee, A., Brizuela, L., Albert,
472 F.W., de la Rasilla, M., Fortea, J., Rosas, A., Lachmann, M., Hannon, G.J. & Paabo, S.
473 (2010) Targeted investigation of the Neandertal genome by array-based sequence
474 capture. Science, 328, 723-725.

475

476 Carpenter, M.L., Buenrostro, J.D., Valdiosera, C., Schroeder, H., Allentoft, M.E., Sikora,
477 M., Rasmussen, M., Gravel, S., Guillen, S., Nekhrizov, G., Leshtakov, K., Dimitrova, D.,
478 Theodossiev, N., Pettener, D., Luiselli, D., Sandoval, K., Moreno-Estrada, A., Li, Y.,
479 Wang, J., Gilbert, M.T., Willerslev, E., Greenleaf, W.J. & Bustamante, C.D. (2013)
480 Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient
481 DNA Sequencing Libraries. American journal of human genetics, 93, 852-864.
482

483 Carter, N.P., Fiegler, H. & Piper, J. (2002) Comparative analysis of comparative
484 genomic hybridization microarray technologies: report of a workshop sponsored by
485 the Wellcome Trust. Cytometry, 49, 43-48.
486

487 Dabney, J., Knapp, M., Glocke, I., Gansauge, M.T., Weihmann, A., Nickel, B., Valdiosera,
488 C., Garcia, N., Paabo, S., Arsuaga, J.L. & Meyer, M. (2013) Complete mitochondrial
489 genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort
490 DNA fragments. Proceedings of the National Academy of Sciences of the United
491 States of America, 110, 15758-15763.
492

493 Enk, J.M., Devault, A.M., Kuch, M., Murgha, Y.E., Rouillard, J.M. & Poinar, H.N. (2014)
494 Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. Molecular
495 biology and evolution.

496 Fu, Q., Meyer, M., Gao, X., Stenzel, U., Burbano, H.A., Kelso, J. & Paabo, S. (2013) DNA
497 analysis of an early modern human from Tianyuan Cave, China. Proceedings of the
498 National Academy of Sciences of the United States of America, 110, 2223-2227.
499
500 Gansauge, M.T. & Meyer, M. (2013) Single-stranded DNA library preparation for the
501 sequencing of ancient or damaged DNA. Nature protocols, 8, 737-748.
502
503 Gelman, A. & Rubin, D.B. (1992) Inference from iterative simulation using multiple
504 sequences. Statistical science, 457-472 %@ 0883-4237.
505
506 Ginolhac, A., Rasmussen, M., Gilbert, M.T., Willerslev, E. & Orlando, L. (2011)
507 mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics,
508 27, 2153-2155.
509
510 Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell,
511 T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D.B., Lander, E.S. & Nusbaum,
512 C. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively
513 parallel targeted sequencing. Nature biotechnology, 27, 182-189.
514
515 Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N.,
516 Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E.Y., Malaspinas, A.S., Jensen, J.D.,
517 Marques-Bonet, T., Alkan, C., Prufer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz,
518 R., Aximu-Petri, A., Butthof, A., Hober, B., Hoffner, B., Siegemund, M., Weihmann, A.,

519 Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C.,
520 Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V.,
521 Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.,
522 Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J.,
523 Lachmann, M., Reich, D. & Paabo, S. (2010) A draft sequence of the Neandertal
524 genome. Science, 328, 710-722.
525
526 Jonsson, H., Ginolhac, A., Schubert, M., Johnson, P.L. & Orlando, L. (2013)
527 mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage
528 parameters. Bioinformatics, 29, 1682-1684.
529
530 Knapp, M. & Hofreiter, M. (2010) Next Generation Sequencing of Ancient DNA:
531 Requirements, Strategies and Perspectives %M doi:10.3390/genes1020227 %U
532 http://www.mdpi.com/2073-4425/1/2/227. Genes %@ 2073-4425, 1, 227-243.
533
534 Li, H. & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-
535 Wheeler transform. Bioinformatics, 25, 1754-1760.
536
537 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
538 G. & Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools.
539 Bioinformatics, 25, 2078-2079.
540

541 Lindgreen, S. (2012) AdapterRemoval: Easy Cleaning of Next Generation Sequencing
542 Reads. BMC research notes, 5, 337.
543
544 Maricic, T., Whitten, M. & Paabo, S. (2010) Multiplexed DNA sequence capture of
545 mitochondrial genomes using PCR products. PloS one, 5, e14004.
546
547 Meyer, M. & Kircher, M. (2010) Illumina sequencing library preparation for highly
548 multiplexed target capture and sequencing. Cold Spring Harbor protocols, 2010, pdb
549 prot5448.
550
551 Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G.,
552 Jay, F., Prufer, K., de Filippo, C., Sudmant, P.H., Alkan, C., Fu, Q., Do, R., Rohland, N.,
553 Tandon, A., Siebauer, M., Green, R.E., Bryc, K., Briggs, A.W., Stenzel, U., Dabney, J.,
554 Shendure, J., Kitzman, J., Hammer, M.F., Shunkov, M.V., Derevianko, A.P., Patterson,
555 N., Andres, A.M., Eichler, E.E., Slatkin, M., Reich, D., Kelso, J. & Paabo, S. (2012) A
556 high-coverage genome sequence from an archaic Denisovan individual. Science, 338,
557 222-226.
558
559 Newkirk, H.L., Knoll, J.H. & Rogan, P.K. (2005) Distortion of quantitative genomic
560 and expression hybridization by Cot-1 DNA: mitigation of this effect. Nucleic acids
561 research, 33, e191.
562

563 Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert,
564 M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P.L., Fumagalli, M., Vilstrup, J.T.,
565 Raghavan, M., Korneliussen, T., Malaspinas, A.S., Vogt, J., Szklarczyk, D., Kelstrup,
566 C.D., Vinther, J., Dolocan, A., Stenderup, J., Velazquez, A.M., Cahill, J., Rasmussen, M.,
567 Wang, X., Min, J., Zazula, G.D., Seguin-Orlando, A., Mortensen, C., Magnussen, K.,
568 Thompson, J.F., Weinstock, J., Gregersen, K., Roed, K.H., Eisenmann, V., Rubin, C.J.,
569 Miller, D.C., Antczak, D.F., Bertelsen, M.F., Brunak, S., Al-Rasheid, K.A., Ryder, O.,
570 Andersson, L., Mundy, J., Krogh, A., Gilbert, M.T., Kjaer, K., Sicheritz-Ponten, T.,
571 Jensen, L.J., Olsen, J.V., Hofreiter, M., Nielsen, R., Shapiro, B., Wang, J. & Willerslev, E.
572 (2013) Recalibrating Equus evolution using the genome sequence of an early Middle
573 Pleistocene horse. Nature, 499, 74-78.
574
575 Plummer, M., Best, N., Cowles, K. & Vines, K. (2006) CODA: Convergence diagnosis
576 and output analysis for MCMC. R news, 6, 7-11.
577
578 Rasmussen, M., Anzick, S.L., Waters, M.R., Skoglund, P., DeGiorgio, M., Stafford, T.W.,
579 Jr., Rasmussen, S., Moltke, I., Albrechtsen, A., Doyle, S.M., Poznik, G.D.,
580 Gudmundsdottir, V., Yadav, R., Malaspinas, A.S., White, S.S.t., Allentoft, M.E., Cornejo,
581 O.E., Tambets, K., Eriksson, A., Heintzman, P.D., Karmin, M., Korneliussen, T.S.,
582 Meltzer, D.J., Pierre, T.L., Stenderup, J., Saag, L., Warmuth, V.M., Lopes, M.C., Malhi,
583 R.S., Brunak, S., Sicheritz-Ponten, T., Barnes, I., Collins, M., Orlando, L., Balloux, F.,
584 Manica, A., Gupta, R., Metspalu, M., Bustamante, C.D., Jakobsson, M., Nielsen, R. &

585 Willerslev, E. (2014) The genome of a Late Pleistocene human from a Clovis burial
586 site in western Montana. Nature, 506, 225-229.
587
588 Rohland, N. & Hofreiter, M. (2007) Ancient DNA extraction from bones and teeth.
589 Nature protocols, 2, 1756-1762.
590
591 Sánchez-Quinto, F., Schroeder, H., Ramirez, O., Avila-Arcos, M.C., Pybus, M., Olalde, I.,
592 Velazquez, A.M., Marcos, M.E., Encinas, J.M., Bertranpetit, J., Orlando, L., Gilbert, M.T.
593 & Lalueza-Fox, C. (2012) Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-
594 Gatherers. Current biology : CB.
595
596 Schuenemann, V.J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest,
597 S., Coombes, B.K., Wood, J.W., Earn, D.J., White, W., Krause, J. & Poinar, H.N. (2011)
598 Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia
599 pestis from victims of the Black Death. Proceedings of the National Academy of
600 Sciences of the United States of America, 108, E746-752.
601
602 Schuenemann, V.J., Singh, P., Mendum, T.A., Krause-Kyora, B., Jager, G., Bos, K.I.,
603 Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J.L., Kjellstrom, A.,
604 Wu, H., Stewart, G.R., Taylor, G.M., Bauer, P., Lee, O.Y., Wu, H.H., Minnikin, D.E., Besra,
605 G.S., Tucker, K., Roffey, S., Sow, S.O., Cole, S.T., Nieselt, K. & Krause, J. (2013) Genome-
606 wide comparison of medieval and modern Mycobacterium leprae. Science, 341, 179-
607 183.

608

609 Skoglund, P., Malmstrom, H., Raghavan, M., Stora, J., Hall, P., Willerslev, E., Gilbert,

610 M.T., Gotherstrom, A. & Jakobsson, M. (2012) Origins and genetic legacy of Neolithic

611 farmers and hunter-gatherers in Europe. Science, 336, 466-469.

612

613

ACKNOWLEDGEMENTS

The authors thank the staff at the Danish National High throughput Sequencing Centre for Technical Assistance, Jay Haviser and Corinne Hofman for providing samples, Jean-Marie Rouillard and Jake Enk for thoughtful input into the manuscript, and Alexandra Adams for technical support during WISC experiments. The authors acknowledge the following grants for generous financial support: Danish Research Foundation grant DNRF94, Danish Council for Independent Research grant 10-081390, Lundbeck Foundation grants R52-A5062 and Marie Curie Actions grant EUROTAST FP7-PEOPLE-2010, George Rosenkranz Prize for Health Care Research in Developing Countries, National Science Foundation award DMS-1201234, NIH NRSA postdoctoral fellowship (grant no. 5 F32 HG007342), Swiss National Science Foundation and NEXUS1492 (ERC Synergy Project Grant Agreement n° 319209)

CONFLICT OF INTEREST

The authors declare no conflict of interest.

636 TABLES

637

638 Table 1

639 Description of samples and WGC scheme used for each. Provenance and

640 approximate age of the samples are shown, along with the pre-capture endogenous

641 content of the DSL libraries built from them. The type of library (DSL: Double-

642 stranded library; SSL: Single-stranded library) built and the whole genome capture

643 method(s) used are depicted in the last two columns.

Sample	Region	Age (years BP)	% Endogenous	DSL-Capture?	SSL-Capture?
CDM25	Caribbean	400	1.8%	WISC + MYbaits	No
CDM95	Caribbean	400	3.1%	MYbaits only	No
ETR2	Europe	2500	0.3%	WISC + MYbaits	No
ETR5	Europe	2500	0.9%	MYbaits only	No
ETR9	Europe	2500	0.2%	WISC + MYbaits	No
STM1	Caribbean	300	2.7%	WISC + MYbaits	WISC + MYbaits
STM2	Caribbean	300	0.3%	WISC + MYbaits	WISC + MYbaits
STM3	Caribbean	300	8.0%	WISC + MYbaits	WISC + MYbaits

644
645

646 Table 2

647 Lower pre-capture endogenous content results in higher clonality (measured as the
648 fraction of the total mapped reads that are clonal (PCR) duplicates) in post-capture
649 libraries. The total number of cycles includes two pre-capture and one post-capture
650 amplification rounds. Higher clonality in WISC experiments is likely caused by the
651 higher number of total PCR cycles used in these experiments, which in turn is due to
652 less DNA used as input of the capture experiment. This correlation follows a
653 logarithmic curve as illustrated in Supplementary Fig. S7.

		Clonality (%)		Library DNA input (ng)		Total number of PCR cycles	
Sample	% Endogenous	MYbaits	WISC	Mybaits*	WISC **	MYbaits	WISC
CDM25	0.61%	37.98	82.76	520.2	226.8	30	42
CDM95	0.40%	81.43	-	460.8	-	30	-
ETR2	0.27%	56.56	91.73	539.1	213.3	32	42
ETR5	0.33%	24.26	-	408	-	40	-
ETR9	0.15%	62.46	93.33	420.3	175.5	32	42
STM1	2.71%	19.44	44.01	481.5	156.6	36	44
STM2	0.33%	44.71	82	396	175.5	34	44
STM3	8.04%	12.86	36.84	445.5	135	32	44
STM1-ssl	5.69%	29.95	89.34	485.4	83.7	30	38
STM2-ssl	1.67%	64.31	98.79	417.6	129.6	32	38
STM3-ssl	18.04%	12.42	45.76	392.4	194.4	30	38

*DNA in 4µL
**DNA in 27µL

654
655

656 Table 3

657 Fold enrichment of WGC by method and type of captured library. Numbers

658 represent the unique informative (non-clonal) enrichment of post-capture libraries

659 when compared to their pre-capture counterpart.

	DSL			SSL			
Sample	% End[a]	WISC	MYbaits	% End[b]	shotgun SSL[c]	WISC[c]	MYbaits[c]
CDM25	0.6%	3.04	4.83	-	-	-	-
CDM95	0.4%	-	5.906	-	-	-	-
ETR2	0.3%	2.70	5.10	-	-	-	-
ETR5	0.3%	-	2.90	-	-	-	-
ETR9	0.2%	1.93	4.13	-	-	-	-
STM1	2.7%	5.53	5.96	5.7%	2.10	0.42	2.22
STM2	0.3%	5.66	8.62	1.67%	5.14	0.12	1.52
STM3	8.0%	2.74	3.73	18.07%	2.24	1.15	1.68

[a]Endogenous DNA content of pre-capture DSL

[b]Endogenous DNA content of pre-capture SSL

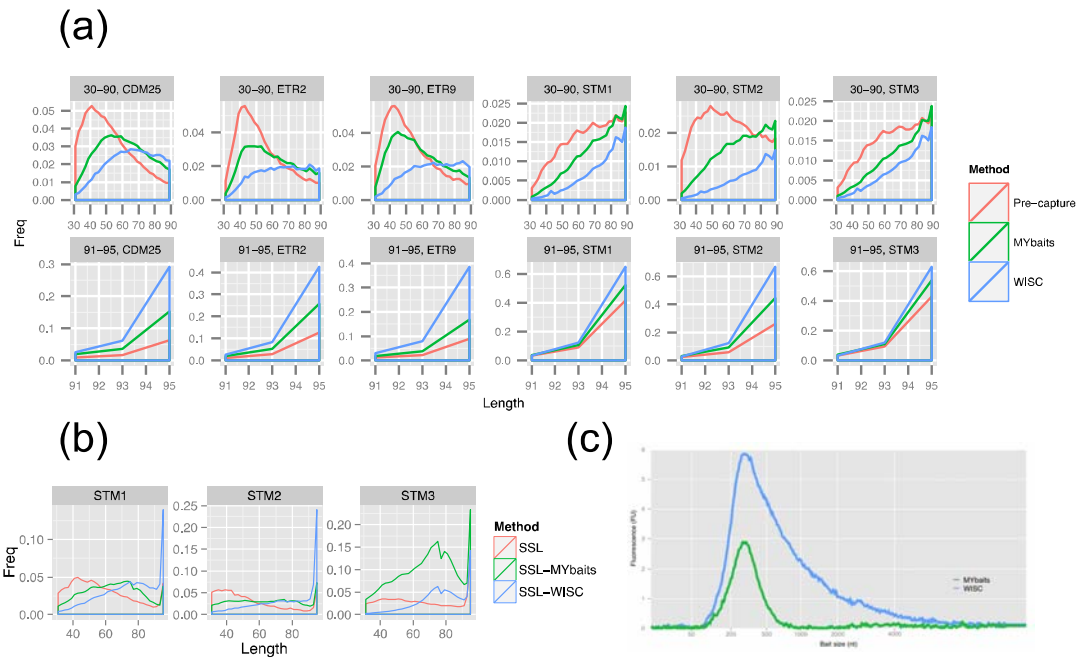
[c] Enrichment values are in relation to pre-capture DSL

660

661

662

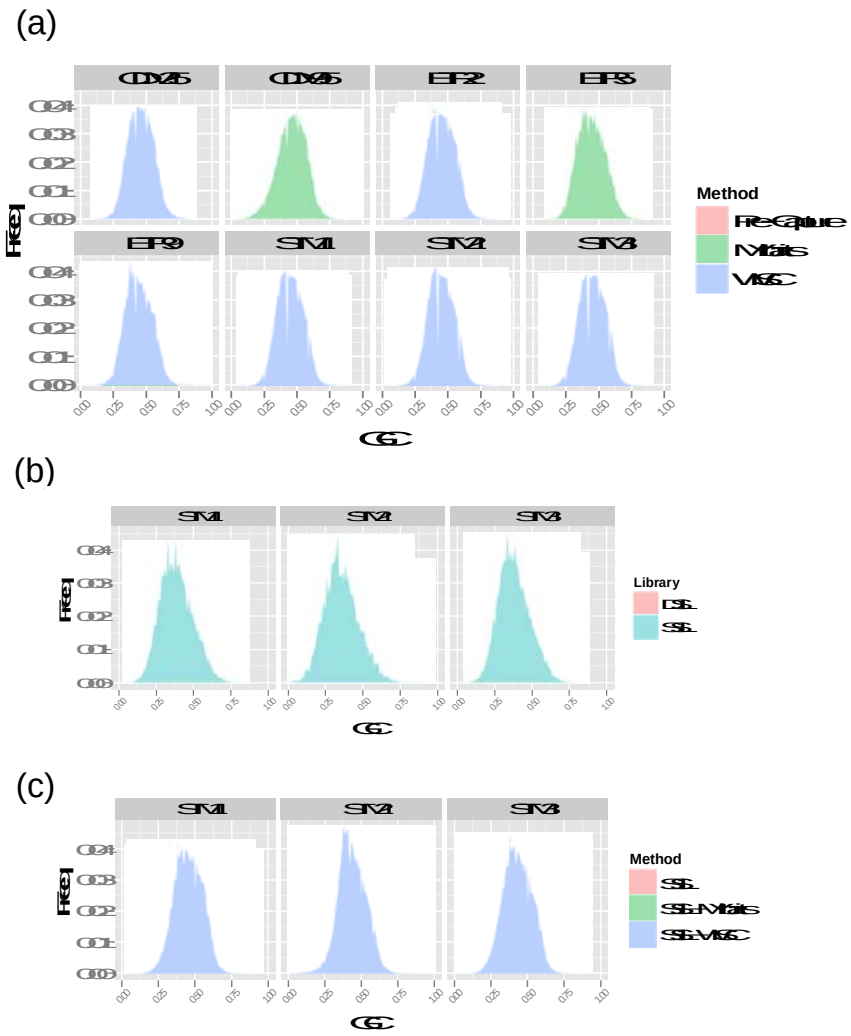
663 FIGURES



664

665 Fig. 1 WGC preferentially retrieves longer fragments in sequencing libraries. The
666 read length distribution of pre-capture and post-capture libraries is shown for (a)
667 double-stranded libraries (DSL) and (b) single-stranded-libraries. In (a) the x axis is
668 split in <90 bp and >90 bp to adjust the scale and better illustrate the higher
669 concentration of short reads in the pre-capture libraries (pink line) and the bias
670 observed against these in capture experiments (green and blue lines) where longer
671 fragments are preferentially retrieved. (b) Illustrates that the relative gain of
672 shorter fragments obtained by building a SSL, is lost by capturing these types of
673 libraries. The plot shown in (c) depicts the bioanalyzer profile of the bait libraries
674 revealing that for WISC a wider tail is observed for longer baits, which might explain
675 the stronger bias in favor of longer fragments by this particular method.

676



677

678 Fig. 2 GC content distribution of WGC libraries is narrower than pre-capture ones.

679 GC distributions in DSL libraries (a) show a subtle narrowing in the post-capture

680 experiments (See Supplementary table 2 for exact ranges and summary statistics).

681 Whereas the GC distribution in pre-capture SSL libraries displays a shift toward

682 lower GC values when compared to pre-capture DSL (b), these values increase in

683 post-capture SSL (c).

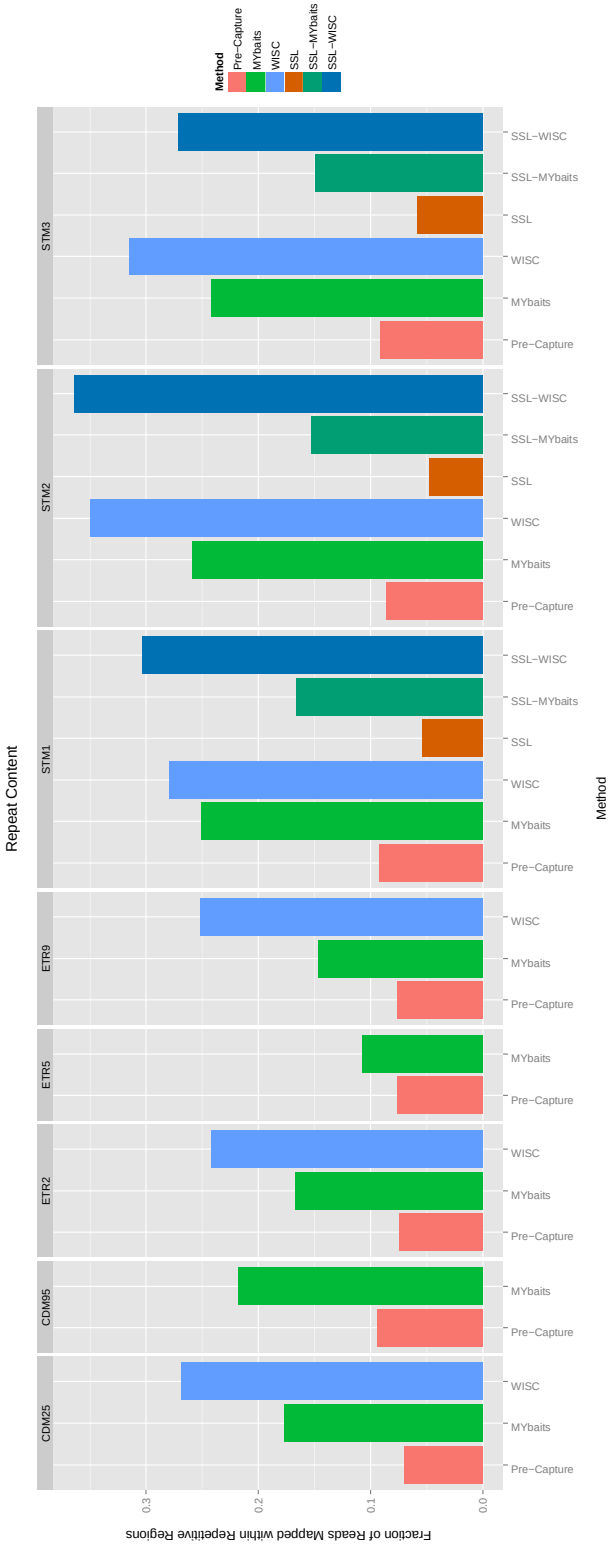


Fig. 3 Fraction of reads mapping to repeated regions in the human genome (see Materials and Methods) in pre- and post-capture libraries. Post-capture libraries display a higher fraction of reads within repeats than pre-capture. This pattern is likely tied to the preferential capture of longer reads, which are also enriched in repeats in pre-captured libraries (Supplementary Table 3)