Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA

Avila Arcos, Maria del Carmen; Cappellini, Enrico; Romero-Navarro, J. Alberto; Wales, Nathan; Moreno Mayar, José Victor; Rasmussen, Morten; Fordyce, Sarah Louise; Montiel, Rafael; Vielle-Calzada, Jean-Philippe; Willerslev, Eske; Gilbert, Tom

Published in:
Scientific Reports

DOI:
10.1038/srep00074

Publication date:
2011

Document version
Publisher's PDF, also known as Version of record

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
DNA
MOLECULAR BIOLOGY
NUCLEIC ACID SEQUENCING
PLANT SCIENCES

Received
3 June 2011

Accepted
9 August 2011

Published
24 August 2011

Correspondence and requests for materials should be addressed to M.T.P.G. (mtpgilbert@gmail.com)

Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA

María C. Ávila-Arcos[1], Enrico Cappellini[1], J. Alberto Romero-Navarro[1,2], Nathan Wales[1,3], J. Víctor Moreno-Mayar[1,2], Morten Rasmussen[1], Sarah L. Fordyce[1], Rafael Montiel[4], Jean-Philippe Vielle-Calzada[4], Eske Willerslev[1] & M. Thomas P. Gilbert[1]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark, [2]Undergraduate Program on Genomic Sciences, Universidad Nacional Autónoma de México, Av. Universidad s/n Col. Chamilpa 62210, Cuernavaca, Morelos, Mexico, [3]Department of Anthropology, University of Connecticut, 354 Mansfield Road, Storrs, Connecticut 06269, USA, [4]Laboratorio Nacional de Genómica para la Biodiversidad, CINVESTAV-IPN, Km 9.6 Libramiento Norte, Carretera Irapuato - León, CP 36821 Irapuato, Guanajuato, Mexico.

The development of second-generation sequencing technologies has greatly benefitted the field of ancient DNA (aDNA). Its application can be further exploited by the use of targeted capture-enrichment methods to overcome restrictions posed by low endogenous and contaminating DNA in ancient samples. We tested the performance of Agilent's SureSelect and Mycroarray's MySelect in-solution capture systems on Illumina sequencing libraries built from ancient maize to identify key factors influencing aDNA capture experiments. High levels of clonality as well as the presence of multiple-copy sequences in the capture targets led to biases in the data regardless of the capture method. Neither method consistently outperformed the other in terms of average target enrichment, and no obvious difference was observed either when two tiling designs were compared. In addition to demonstrating the plausibility of capturing aDNA from ancient plant material, our results also enable us to provide useful recommendations for those planning targeted-sequencing on aDNA.

The advent of second-generation sequencing has revolutionized the study of ancient DNA (aDNA), enabling it to move from the study of few, short fragments of DNA, into the so-called 'paleogenomic' era. Nevertheless, the limited amount of endogenous compared to environmental and contaminating DNA in ancient samples still poses a limitation in shotgun aDNA sequencing experiments. The application of DNA capture-enrichment methods is therefore particularly attractive for aDNA studies[1].

Capture-enrichment methods allow target-specific (user-defined) sequencing within organelle and complete genomes, by selectively enriching sequences prior to the sequencing run, yielding an increased depth of sequence over target regions, and overall lowering cost per target. Two main approaches exist to date: relatively large-scale capture using commercial on-array/in-solution methods that typically can target regions of as much as 50 MB sequence (Agilent SureSelect Human All Exon 50 Mb kit), and smaller scale methods (such as Primer Extension Capture (PEC)[2], and a recently published method that exploits biotinylated PCR amplicons as baits[3], that are generally limited to smaller targets (e.g. complete mtDNA genomes). For all methods, the principle relies on 'selection by hybridization' of the sequencing libraries to probes representing the regions of interest prior to sequencing. Probes can be either immobilized on glass slides (on-array capture)[4] or selectively recovered by affinity using magnetic beads (in-solution capture)[5].

Even though the effectiveness of the large-scale commercial methods has been extensively tested on modern samples (eg.[6–8]) their application to aDNA is limited to a single study[9], which reports successful capture of Neandertal nuDNA extracted from bone. Nevertheless, aDNA experiments are routinely complicated by a variety of factors, including type of sampled tissue, levels of contamination with exogenous sources of DNA, and DNA preservation. Specifically, aDNA tends to be highly fragmented and damaged, as well as embedded in a mixture of environmental DNA[10].Therefore, in aDNA capture experiments, the intrinsic factors and complexity of aDNA

handling are added to factors inherent to the capture method and design, and thus it is reasonable to assume that one cannot *a priori* expect such experiments to work in an identical manner to those on high quality DNA[1].

We take advantage of a dataset of unpublished ancient maize kernel and cob DNA sequences (700–1000 years old) (Supplementary Data online), which were sequenced on the Illumina GAIIx using both shotgun sequencing and in-solution capture, from two commercial providers (Agilent's SureSelect and MYcroarray's MySelect), in order to attempt to characterize key factors influencing the success of aDNA capture experiments. The target regions used in each capture method were identical, and incorporated a mix of single copy, multicopy and chloroplast DNA in each single reaction that span a total of 670 loci over ~0.7 Mb. Furthermore, the design of one of the capture kits, MYcroarray's MySelect, under two tiling regimes (11 and 24 bp) enabled us to explore whether tighter tiling plays a beneficial role. We associate our observations with GC content and type of targeted DNA (plastid, nuclear multicopy and nuclear single copy). To evaluate the actual benefit of aDNA capture over shotgun, we calculate the rate of enrichment across methods by comparing the amount of target sequence obtained after capture, to that obtained with shotgun sequencing of the same samples, and contrast the enrichment rates to the pre- and post-capture number of PCR cycles, purity and quality of the sample.

Ultimately, because the complexity and expense of working with aDNA makes generation of sufficient data to recover statistically supported observations challenging, we caution our results should be viewed as trends, as opposed to hard facts. However, we hope the findings presented will be helpful for those who are planning future targeted-sequencing studies on ancient DNA.

Results

On/off Target. Capture efficiency was first measured by estimating the proportion of reads that mapped to the 670 targets, the so-called 'on/off target' proportion (Table 1). The range for MySelect was 7–61% (distributed across 359–659 targets), whereas for Agilent it was 41–75%, covering 395–668 targets. For the two best samples (Arizona and Chile, for which shotgun sequencing indicates contained an initial endogenous maize DNA content of >90%), the proportion of on-target reads observed using the SureSelect resembles those reported by the manufacturer in modern samples and previous reports[7,8]. For all samples, the fraction of reads that mapped to the targets was higher when SureSelect, rather than MySelect, was used to capture.

The fraction of reads mapped to targets is, however, only an indirect estimation of the method's performance, since it needs to be considered in parallel with the coverage uniformity across targets. Furthermore, due to the PCR steps involved either pre- or post-capture, many reads that map to a target may derive from a single original template molecule (so-called 'clonal' reads). Therefore it is imperative to control for this effect in order to get an informative estimate of the effectiveness of the method.

Clonality. All reads with identical sequence mapping to the same start and end position in the target were conservatively considered to be clonal (or duplicates). These were collapsed, and reads not mapping to a unique position within the targets were filtered from the data. Just as SureSelect was the method yielding the highest fraction of reads mapping to targets, it also showed higher clonality values for all samples as illustrated in Table 2.

The number of reads remaining after removing duplicates was compared to the total number of mapped reads, giving an estimation of the clonality within the targets. Similarly, the number of unique reads in the whole library was calculated and compared to the total number of reads to provide an estimate of the clonality within the library. A very high amount of clonality was detected regardless of the sample and capture method (Table 2), particularly for sequences that mapped to targets (min 38 copies per read [cpr], max 597 cpr, median 118 cpr).

To further explore the distribution of clonality across targets, clonality content was broken down in three categories: nuclear single copy, nuclear multiple copy and chloroplast. We observed that targets falling into the multiple copy category contained the highest levels of clonality (avg. 20x more than the single-copy category) for almost all samples. Because the maize genome is highly repetitive with ~80% of the genome comprising common repeats[11], their capture is likely to be favored, and hence preferentially amplified in the post-capture PCR step.

This pattern was prevalent in all samples regardless of the chosen method; we therefore investigated whether the amount of detected clonality could be associated to one or a combination of the following:

1) Number of PCR cycles pre- and post-capture
2) Endogenous DNA content

Using the method described above, the degree of clonality was also assessed in the shotgun data, to enable a baseline for comparison and estimation of the effect of the number of post-capture PCR cycles on the final clonality present in capture experiments. The level of clonality in shotgun experiments was also observed to be higher in the mapped reads, supporting the notion that some reads called as "clonal", most likely correspond to repeated copies in the genome.

Table 1 | Proportion of sequence reads on target across experiments. As two tiling densities were used in the MySelect array designs, the tiling overlap is indicated in the relevant column headers. The method yielding the higher proportion of reads on target is highlighted in bold.

	Arizona				Chile			
Sample	MySelect 11	Myselect 24	SureSelect	Shotgun	MySelect 11	MySelect 24	SureSelect	Shotgun
Filtered reads	34,376,665	34,727,685	24,485,087	38,570,708	38,584,450	16,857,254	21,990,480	27,033,243
On target	18,403,448	21,168,407	17,926,390	1,307,768	17,548,482	8,018,092	16,444,488	950,692
%	53.53	60.96	**73.21**	3.391	45.48	47.56	**74.78**	3.517
	GMAG 10237				GMAG 10189			
Sample	MySelect 11	MySelect 24	SureSelect	Shotgun	MySelect 11	MySelect 24	SureSelect	Shotgun
Filtered reads	29,187,586	33,342,092	30,461,620	14,754,738	35,091,478	36,280,834	33,965,192	14,754,738
On target	6,141,411	2,208,139	12,471,652	30,725	9,747,308	8,166,851	14,171,670	30,725
%	21.04	6.62	**40.94**	0.208	27.78	22.51	**41.72**	0.208

Table 2 | Clonality values (copies per read "cpr") for each sample and capture method. MyS: MySelect, SuS: SureSelec, Sh: Shotgun

	Arizona				Chile			GMAG 10237-18			GMAG 10189			
	Sh	MyS 11	MyS 24	SuS	Sh	MyS 11	My S 24	SuS	Sh	MyS 11	My S24	SuS	Sh	MyS 11	MyS 24	SuS
Avg clonality total	1.26	8.53	10.37	2.75	1.24	16.88	9.00	2.76	1.84	11.09	9.33	8.22	1.05	8.70	8.29	1.90
Avg clonality in targets	13.17	106.20	129.93	54.34	13.48	355.38	144.30	50.74	3.68	402.40	162.32	597.47	2.66	86.89	73.56	42.53
Ratio	**10.46**	12.45	12.53	**19.73**	**10.87**	**21.05**	16.03	18.42	**2.00**	36.27	17.40	**72.66**	**2.53**	9.99	8.88	**22.38**
Chloroplast	1.87	73.90	95.20	30.21	1.48	160.20	78.80	24.10	1.97	283.30	120.24	266.42	1.13	41.04	39.02	20.13
With repeats	34.00	262.20	311.85	175.14	26.60	828.57	331.40	165.42	4.38	565.50	234.30	915.56	3.13	216.30	181.50	135.73
Without repeats	1.10	16.24	18.82	3.30	1.15	107.33	45.42	3.13	2.00	236.10	58.57	154.26	1.00	23.61	20.70	2.83

The ratio of clonality in the library to captured sequences implied an effect of the number of pre- and post-capture PCR cycles on the amount of clonality. The lowest values for this ratio were consistently found for shotgun experiments. The highest clonality ratios were found using the SureSelect method for three out of four samples, regardless of the PCR cycles.

A high number of PCR cycles (30 post capture) on samples with >90% maize DNA content did not seem to proportionally increase the amount of clonality. However the use of a high number of PCR cycles on samples that contained a low concentration of endogenous DNA considerably increased the amount of clones. Consequently, it seems preferable to obtain a sufficient yield of PCR amplified library through pooling of multiple independent amplification reactions on the library, each performed using a low number of PCR cycles (similar to how sample CMAG 10189 was prepared).

The effect of this can be observed in the two samples with the highest endogenous maize DNA content (as determined from the shotgun results), the Arizona and Chile kernels, both with >90% of maize DNA. Both had the same number of pre capture PCR cycles (20 for MySelect, 22 for SureSelect). However, for the Arizonan sample, the library was divided in seven different aliquots and parallel PCR reactions were carried out after capture with 18 cycles each for both MySelect experiments, and nine tubes with 22 cycles each for the SureSelect. For the Chilean sample, the same protocol was used for the SureSelect capture and the MySelect with the exception that three of the seven aliquots were subjected to a PCR with 30 cycles. This had a very subtle effect on the clonality ratios, but still perceptible in the MySelect experiments (increase of 4–8.5 cpr). The clonality ratios showed similar values for the two SureSelect experiments, which is consistent since the same pre- and post-PCR protocol was carried out for both.

On the other hand, samples CMAG 10237-18 and CMAG 10189 had a low proportion of endogenous DNA (24.7% and 11% respectively), in comparison to samples Chile and Arizona. For CMAG 10237-18, the MySelect-captured library was amplified for 22–27 cycles (three aliquots: 1×22 and 2×27) and with only 18 cycles for CMAG 10189 (5 aliquots). As a result CMAG 10237 had the highest values of clonality among all the samples and CMAG 10189 had the lowest, suggesting that the number of cycles does have an effect when the endogenous amount of DNA is below (at least) 25%.

Enrichment. We estimated the 'enrichment' rate by comparing the proportion of non-clonal reads that mapped uniquely to the targets out of the total uniquely mapped reads (genome wide) between capture and shotgun experiments (Supplementary Table S1, S3–S6 online). Therefore, this enrichment rate is independent of the amount of sequence generated since it takes proportions into consideration. The average enrichment, regardless of the method, ranged from 4 to 29 fold. However considerable differences were observed when the enrichment was estimated for each category (chloroplast, nuclear single-copy and nuclear multi-copy). Most of the enrichment was observed in the chloroplast and the single copy category, most likely due to the big loss of reads in the clonal removal step within the multi-copy category.

Neither method consistently outperformed the other. In samples with good amount of endogenous DNA (>90%), MySelect proved better for one of them (Arizona) and SureSelect for the other (Chile). For samples with low endogenous DNA content, the same pattern was observed with SureSelect proving better for one (CMAG 10237) out of the two samples. Regarding MySelect's tiling design, enrichment did not appear to be influenced by the overlap of the probes. Therefore at least for the densities explored here (11 and 24 bp) the adoption of lower tiling densities, that allows more target to be captured per reaction, does not appear to affect the final enrichment.

While for modern human DNA, solution-based capture enrichment has been reported to be ~400 fold[6], a report on aDNA capture

from Briggs *et al.*[2] indicated a raw target enrichment of 3640- to 80,400-fold when PEC and FLX sequencing were carried out to obtain the complete sequence of Neandertal mitochondria. Burbano *et al.*[9] on the other hand reported a ~190,000-fold target enrichment when using the solid-phase Microarray version of SureSelect to capture nuclear protein-coding positions also from Neandertal. We sought to explore why the enrichment we observed in our experiments was several orders of magnitude lower than those reported by Briggs *et al.*[2] and Burbano *et al.*[9].

The first discrepancy we observed was the method for enrichment estimation. Briggs *et al.*[2] and Burbano *et al.*[9] used a theoretical enrichment value that considered the target fraction of the genome and the proportion of endogenous DNA present in the sequencing library, as the baseline for comparison against the results from the enriched and clonal libraries. We directly calculated enrichment by mapping shotgun reads to the targets and then contrasted the fractions of uniquely mapped non-clonal reads between this and the enriched library. In our opinion this approach allows a straightforward calculation of effective enrichment and considers possible biases in the sequencing library that could influence the amount, quality and content of sequence reads, which must be taken into consideration in aDNA investigation. Additionally, because we do not include clones when calculating the enrichment, we obtain an "informative" rate of enrichment as opposed to enrichment due to clonal redundancy and PCR drift.

With regards to method, the main differences between the above experiments and ours reported here are the type of captured libraries and the number of capture reactions as well as the number of cycles used in the amplification steps. In both previous reports, FLX libraries were captured while we performed the capture on Illumina libraries. Specifically, in Burbano *et al.*[9] the original FLX libraries were amplified with two consecutive rounds of PCR, with 10 and 20 cycles respectively. In addition to this, two successive capture reactions were performed, each of which was followed by a 20 cycles amplification step. The captured libraries were then converted, by PCR, to Illumina format and sequenced on the Illumina GAII platform. Similarly in Briggs *et al.*[2] FLX libraries were also amplified in two consecutive rounds of PCR with first 8 and subsequently 8–10 cycles. The amplified libraries were then captured twice. Following the first capture, products were reamplified for 14 cycles, and amplified once more after the second capture for 8 cycles, prior to sequencing on a FLX platform. These differences may account for the discrepancies observed in the enrichment rates in our experiments. Because we carried out a single capture reaction per sample and no consecutive rounds of PCR, it is expected to have a lower enrichment. In addition to this, the number of PCR cycles we used per reaction (18–30, depending on the experiment) was higher than the one used by any of the above-mentioned studies, which also explains the low yield of non-clonal, unique sequences. Lastly, the fact that we captured Illumina, as opposed to FLX, libraries, is unlikely to influence the total yield of captured DNA.

Capturing repetitive elements. Our set of targets included 669 nuclear regions and the complete chloroplast. 653 of the nuclear regions were identical sequence regions (IDSRs) between Palomero and B73 genomes reported in a previous work[12] plus fifteen enolases and metabolic genes. 223 targets, including the chloroplast, contained common repeats (99 kb) in their sequence as well as regions of low complexity as detected by RepeatMasker[13]. Because ~85% of the maize genome is composed of hundreds of families of transposable elements[11] the capture of such regions poses a challenge, however we investigated the effect of capturing such regions on the total yield of informative sequences.

For shotgun experiments the proportion of reads mapping to annotated repeats, was ~90% for all samples (Supplementary Table S2 online) before removing clones. This is consistent with the reported proportion of repeats in the maize genome[11]. For the capture experiments this proportion varied depending on the method (49–78%, mean 65%), however it was the predominant category for all samples. After clones were removed it was clear that the chloroplast and the targets with common repeats carried most of the clonal reads, representing more than 98% of the lost reads. This indicates that the capture of repeat-rich regions causes a big limitation in the capture of non-repeated regions.

GC Content. Because an effect of GC content on the efficiency of capture has been previously reported for solution-based SureSelect in modern samples, we investigated whether the effect was similar in our ancient samples. Previous reports observe a Gaussian-like distribution of GC content versus normalized coverage, with a peak at 45% GC content and less successful capture at both ends of the distribution[6].

We evaluated the correlation of GC content average vs. average depth of coverage, independently for each category of target (Supplementary Figs. S1–S4 online). This was estimated per probe and not per target to control for the variance in lengths of the targets. For the two samples with the highest fraction of endogenous maize DNA (Arizona and Chile) the pattern observed in the single-copy results derived from SureSelect capture, highly resembled that reported by Tehewy and colleagues[6], except that the peak appears at ~55% GC. For chloroplast, a positive correlation is observed mainly due to its low GC content (avg. 38%), with the highest value for GC content, in any probe, being 63%.

Discussion

We present what to our knowledge is the first attempt to capture ancient DNA with two independent in-solution capture methods. Although based on a limited dataset, we believe that the data presented contains useful information to help guide future studies. Our data suggests that several of the variables tested appear to be of limited importance. For example, in general we did not observe any overwhelming advantage of one kit over the other. Although the SureSelect experiments contained higher proportions of on-target sequences, they also contained higher levels of clonal data, thus reducing the overall levels of useful sequence post filtering. Furthermore, while the use of different tiling designs in the MySelect experiments allowed a direct estimation of the advantage of using tighter designs (higher tiling density) at the expense of less target sequence, we observed the use of tighter designs was not consistently translated into better enrichment values.

In contrast, we believe that there are two major observations that will be important for future aDNA target-sequencing studies to consider. Firstly, a large proportion of the sequences derived from clonal amplification, and thus would be of limited use in down-stream analyses. Because ancient DNA capture-enrichment experiments are still in its earliest stage, we found difficult to compare our observations to the other few reports on the subject since no standardisation has been made to report enrichment rates, and it was not clear at first sight if these included clonality or not in their estimations. We therefore believe clonality should be unambiguously discussed in relevant publications, and such data should be excluded from enrichment calculations; at the very least we feel that our discipline should engage in a salient discussion on the matter to ultimately reach a standardized convention of enrichment rates.

Secondly, the inclusion of targets present at different levels in the genome in the same capture array appears to have a negative effect on the results. Specifically we speculate that repeat regions are preferentially captured and amplified. This highlights the relevance of designing probes to exclusively target single-copy loci as well as keeping the number of cycles to a minimum, which would also have a positive effect on the results by reducing the number of clones.

Although we demonstrate the plausibility of capturing aDNA from maize kernels, the efficiency in other ancient tissues remains to be explored. Even though the only additional report on nuclear aDNA capture-enrichment is limited to Neandertal bone, the future looks quite promising for aDNA investigation with targeted sequencing becoming more accessible and newer sequencing platforms entering the market.

Methods

Design and data generation. 120 bp length probes were designed (baits) with an 11 bp tiling for 670 targets in maize (covering ca. 700 Kb), which included the complete chloroplast, few enolases and genomic regions reported to be associated with maize domestication. Such regions included the different types of DNA: Nuclear single copy, nuclear multi-copy, and plastid DNA. The ability of each method to capture each one of these types was also evaluated. In addition, a different tiling design of 24 bp was tested for MySelect.

DNA was extracted from desiccated Chilean and Arizonan maize kernels, dated to 750–550 YBP, and 723±23 14C YBP, respectively; and two desiccated Mexican cobs, CMAG 10189 and CMAG 10237 dated to 1410±25 14C YBP. DNA extracts were converted into Illumina GAIIx libraries using two NEBNext kits following manufacturer's instructions (see Supplementary Methods online).

Capture-enrichment experiments were carried out following manufacturer's instructions, with the modifications shown in Supplementary Methods online. For each sample, three capture-enrichment reactions (Agilent's SureSelect, MySelect tilling 11 bp and MySelect tilling 24 bp) were performed. The number of cycles in the post-capture PCR amplification step was different among samples; these are described in Supplementary Tables S3–S6 online.

Non-captured libraries were also shotgun sequenced to measure the enrichment factor of each method. A total of four lanes per sample were required (three captured, one shotgun), yielding a total of 24 lanes on the Illumina GAIIx. Illumina's pipeline software (RTA1.8/SCS2.8) was used for base-calling.

Data analysis. Fastq files, representing a total of ~48 Gb (628,147,291 reads), produced by the base-calling program were analysed with a script to compute base frequency and quality per cycle. The quality of the reads and the removal of low quality bases and adapter sequences was carried out using R (version 2.13.0)[14] and the Bioconductor package ShortRead[15]. Adapter sequences were removed from the reads, and reads consisting of only adapter were excluded from the downstream analysis. The algorithm ensured that the adapter was detected regardless of its position in the sequence. All bases following a B quality base (Phred of 3) were removed. Finally, only sequences longer than 18 bp, after removing the adapter and trimming low quality stretches, were considered and mapped to the reference targets using BWA. This threshold was arbitrarily set to allow short sequences to be included in the analysis but not too short to produce spurious random matches.

Bam files, generated by BWA version 0.5.9-r16[16], were further manipulated using R and perl scripts, along with samtools version: 0.1.14[17], picard version 1.34[18] and GATK version 1.0.4641M[19]. Duplicates and ambiguous hits (reads with more than one hit) were removed using samtools. Average coverage and depth of coverage was calculated with GATK's DepthOfCoverage analysis.

Common repeats in the maize genome were detected using RepeatMasker[13].

1. Knapp, M. & Hofreiter, M. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives %M doi:10.3390/genes1020227 %U http://www.mdpi.com/2073-4425/1/2/227 *Genes %@ 2073–4425* **1**, 227–243 (2010).
2. Briggs, A. W. *et al*. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321, doi:10.1126/science.1174462 (2009).
3. Maricic, T., Whitten, M. & Paabo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS One* **5**, e14004, doi:10.1371/journal.pone.0014004 (2010).
4. Hodges, E. *et al*. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**, 1522–1527, doi:10.1038/ng.2007.42 (2007).
5. Gnirke, A. *et al*. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**, 182–189, doi:10.1038/nbt.1523 (2009).
6. Tewhey, R. *et al*. Enrichment of sequencing targets from the human genome by solution hybridization. *Genome Biol.* **10**, R116, doi:10.1186/gb-2009-10-10-r116 (2009).
7. Teer, J. K. *et al*. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* **20**, 1420–1431, doi:10.1101/gr.106716.110 (2010).
8. Kiialainen, A. *et al*. Performance of microarray and liquid based capture methods for target enrichment for massively parallel sequencing and SNP discovery. *PLoS One* **6**, e16486, doi:10.1371/journal.pone.0016486 (2011).
9. Burbano, H. A. *et al*. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723–725, doi:10.1126/science.1188046 (2010).
10. Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. Ancient DNA. *Nat Rev Genet* **2**, 353–359, doi:10.1038/35072071 (2001).
11. Schnable, P. S. *et al*. The B73 maize genome: complexity, diversity, and dynamics. *Science* **326**, 1112–1115, doi:10.1126/science.1178534 (2009).
12. Vielle-Calzada, J. P. *et al*. The Palomero genome suggests metal effects on domestication. *Science* **326**, 1078, doi:10.1126/science.1178437 (2009).
13. Smit, A., Hubley, R & Green, P. *RepeatMasker Open-3.0.*, <http://www.repeatmasker.org> (1996–2010).
14. Team, R. D. C. *R: A language and environment for statistical computing.* 409 (R Foundation for Statistical Computing, 2009).
15. Morgan, M. *et al*. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608, doi:10.1093/bioinformatics/btp450 (2009).
16. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, doi:10.1093/bioinformatics/btp324 (2009).
17. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, doi:10.1093/bioinformatics/btp352 (2009).
18. *Picard*, <http://picard.sourceforge.net/>
19. McKenna, A. *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303, doi:10.1101/gr.107524.110 (2010).

Acknowledgments

The authors acknowledge The American-Scandinavian Foundation (NW), the Danish Council for Independent Research 'Skou/Sapere Aude' (MTPG) and 'PalaeoRNA' (EW) grants, The Danish Basic Research Foundation 'GeoGenetics' grant (MTPG, EW), and the Marie Curie Actions 'TIMECAPSULE' (EC) for financial support. The authors also thank archaeologist José Luis Punzo-Díaz for facilitating access to the ancient cobs and Kim Magnussen for technical support.

Author Contributions

MCAA, EC, NW, ARN and MTPG conceived the study. JVC and RM provided material and archaeological context information, and provided the sequencing targets. EC, NW and SLF undertook the laboratory experiments. MCAA, ARN, JVMM and MR designed the data analysis and analysed the data. EW provided the sequencing resources. MCAA and MTPG wrote the manuscript, with critical input from all other authors.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/scientificreports

Competing Financial Interests The authors declare no competing financial interests.

How to cite this article: Ávila-Arcos, M.C. *et al*. Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. *Sci. Rep.* **1**, 74; DOI:10.1038/srep00074 (2011).