Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx consequences on nucleotide misincorporation and methylation patterns

Seguin-Orlando, Andaine; Hoover, Cindi A.; Vasiliev, Sergei K.; Ovodov, Nikolai D.; Shapiro, Beth; Cooper, Alan; Rubin, Edward M.; Willerslev, Eske; Orlando, Ludovic

Published in:
Science and Technology of Archaeological Research

DOI:
10.1179/2054892315Y.0000000005

Publication date:
2015

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

STAR: Science & Technology of Archaeological Research

ISSN: (Print) 2054-8923 (Online) Journal homepage: https://www.tandfonline.com/loi/ysta20

Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns

Andaine Seguin-Orlando, Cindi A. Hoover, Sergei K. Vasiliev, Nikolai D. Ovodov, Beth Shapiro, Alan Cooper, Edward M. Rubin, Eske Willerslev & Ludovic Orlando

To cite this article: Andaine Seguin-Orlando, Cindi A. Hoover, Sergei K. Vasiliev, Nikolai D. Ovodov, Beth Shapiro, Alan Cooper, Edward M. Rubin, Eske Willerslev & Ludovic Orlando (2015) Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns, STAR: Science & Technology of Archaeological Research, 1:1, 1-9, DOI: 10.1179/2054892315Y.0000000005

To link to this article: https://doi.org/10.1179/2054892315Y.0000000005

Science and Technology of Archaeological Research

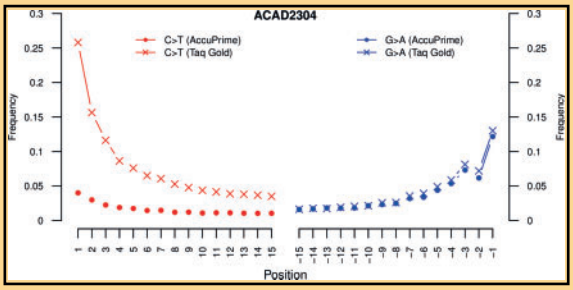
Short Report

Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns

Andaine Seguin-Orlando[1], Cindi A. Hoover[2], Sergei K. Vasiliev[3], Nikolai D. Ovodov[4], Beth Shapiro[5], Alan Cooper[6], Edward M. Rubin[2], Eske Willerslev[7], and Ludovic Orlando[7]

[1]Kobenhavns Universitet, GeoGenetics, Copenhagen, Denmark
[2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA, USA
[3]Institute of Archaeology and Ethnography, Russian Academy of Sciences, Novosibirsk, Russia
[4]Laboratory of Archaeology and Paleogeography of Central Siberia, Institute of Archaeology and Ethnography, Russian Academy of Sciences, Novosibirsk, Russia
[5]Department of Ecology and Evolutionary Biology, University of California Santa Cruz, CA, USA
[6]Australian Centre for Ancient DNA, School of Earth and Environmental Sciences, The University of Adelaide, South Australia, Australia
[7]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Denmark

Abstract Recent developments in high-throughput sequencing (HTS) technologies have enabled the reconstruction of complete genome sequence and nucleosome and methylation maps from ancient individuals. A diversity of degradation reactions affect DNA molecules after an organism dies, introducing specific nucleotide misincorporation patterns in HTS data that are typically exploited to distinguish modern contaminants and authentic DNA templates. In this study, we used two different DNA polymerases to amplify Illumina TruSeq DNA libraries built on DNA extracts from ancient equines. The AmpliTaq Gold, widely used in ancient DNA studies, generated amplified libraries showing typical misincorporation patterns. Such patterns were partially lost following AccuPrime *Pfx* amplification, for which a two-fold reduction of endogenous content was also observed. This is explained by the inability of the AccuPrime *Pfx* to bypass uracils, which represent the most common *post-mortem* base modification and derive from the *post-mortem* deamination of cytosines. Our study, therefore, reveals that amplification of TruSeq DNA libraries with AccuPrime *Pfx* increases the cost of whole-genome sequencing for samples showing substantial levels of DNA degradation and creates atypical nucleotide misincorporation patterns for data authentication. The method can, however, be exploited to identify ancient methylation marks, and potentially, nucleosome occupancy maps.



Statement of significance After an organism dies, DNA molecules are subjected to degradation resulting in fragmentation and bases modifications. After high- throughput sequencing (HTS), it is possible to detect the molecular signature of these damages, demonstrating that the data generated is authentic. DNA extracts from fossil remains indeed contain, besides endogenous ancient molecules, a variety of modern contaminants, but only the former generally exhibits molecular signatures typical of *post-mortem* damage. Our study reveals that these signatures depend on the protocol used for amplifying DNA samples. It thus helps defining better guidelines for authenticating ancient HTS data. Our results also reveal one molecular approach that limits our ability to exploit the whole complexity of ancient DNA molecules preserved in fossil specimens but can help track ancient methylation marks.

Keywords DNA library preparation, Ancient DNA, Next-generation sequencing, Palaeogenomics, Damage pattern, PCR bias

Cite this article Seguin-Orlando, A., Hoover, C.A., Vasiliev, S.K., Ovodov, N.D., Shapiro, B., Cooper, A., Rubin, E.M., Willerslev, E. and Orlando, L. Amplification of TruSeq ancient DNA libraries with AccuPrime Pfx: consequences on nucleotide misincorporation and methylation patterns. *STAR* 2015; 1(1), *STAR*2015112054892315Y.0000000005

Received 06 February 2015; **accepted** 29 March 2015

Data availability The authors confirm that all data underlying the findings are fully available without restriction. All relevant data are contained within the paper.

Corresponding author, email: ASOrlando@snm.ku.dk

DOI 10.1179/2054892315Y.0000000005

Introduction

Ancient DNA can survive as extremely fragmented molecules in fossil material for up to several hundreds of thousands of years (Dabney *et al.* 2013; Meyer *et al.* 2013; Orlando *et al.* 2013). Despite extensive *post-mortem* DNA degradation reactions (Höss *et al.* 1996), colonisation of fossil material by living microbes (Poinar *et al.* 2006) and possible contamination from different human DNA sources (Sampietro *et al.* 2006), the sensitivity of DNA extraction and library building methods and the capacities of high-throughput sequencing (HTS) platforms are now compatible with the reconstruction of complete ancient genomes (see Shapiro and Hofreiter 2014; Der Sarkissian *et al.* 2015; Ermini *et al.* 2015 for review). Additionally, not only genome-wide nucleosome and methylation maps (Pedersen *et al.* 2014; Gokhman *et al.* 2014) but also copy-number variants (Meyer *et al.* 2012; Prüfer *et al.* 2014) and structural rearrangements (Schuenemann *et al.* 2013) can be overlaid on the sequence of ancient genomes.

The detection of typical nucleotide misincorporation patterns in HTS data has proved essential for authenticating ancient DNA results (Krause *et al.* 2010), allowing discrimination between ancient sequences and modern contaminants (Orlando *et al.* 2013). The base composition at sequence termini and at neighbouring genomic coordinates revealed that *post-mortem* DNA degradation was mainly driven by depurination (Briggs *et al.* 2007; Orlando *et al.* 2011; Meyer *et al.* 2012) and also highlighted specific ligation bias in certain DNA library construction methods (Seguin-Orlando *et al.* 2013). The latter has been shown to affect in turn the nucleotide misincorporation pattern used for sequence authentication, which generally consists of decreasing C → T mutation rates from sequence starts and increasing G → A mutation rates towards read ends (Briggs *et al.* 2007).

The Illumina TruSeq DNA sample preparation approach is acknowledged to generate high-quality DNA libraries and is commonly used in genomic sciences (Holmqvist, Reimegaard and Wagner 2013; Ruark *et al.* 2013) and ancient DNA research (e.g. Cui *et al.* 2013; Star *et al.* 2014). How much this approach affects the nucleotide misincorporation pattern has not yet been investigated. The PCR conditions used for amplifying DNA libraries have also been shown to have a strong impact on the final molecular complexity of DNA libraries, with particular classes of DNA polymerases introducing various sorts of base-compositional and length-dependent biases (Dabney and Meyer 2012). Some DNA polymerases, such as AccuPrime *Pfx*, have been shown to virtually annihilate such biases when amplifying Illumina DNA libraries built using blunt-ended adaptors but their performance still remains to be tested on other DNA library types.

In this paper, we investigate whether TruSeq DNA sample preparation introduces significant bias in the base composition and nucleotide misincorporation profile of ancient DNA libraries. We also evaluate the impact of two DNA polymerases (AccuPrime *Pfx* and AmpliTaq Gold), used during DNA library amplification, on the endogenous content of ancient TruSeq libraries. We show that although the inability of AccuPrime *Pfx* to bypass uracils reduces the endogenous content, and thereby increases the overall sequencing costs, it can advantageously be used to reveal methylation patterns in ancient genomes.

Materials and Methods

Samples

The horse bone fragment CGG10027 was provided by the University of Kansas, Lawrence, USA, where it was registered under XA08-43, UP08.BS.265. It was excavated from the Lake Koehkuel caves (Taimyr, Russia) and radiocarbon dated to 24 306 ± 135 uncal. BP (UBA-16483) (Orlando *et al.* 2013). We used the original DNA extract from Orlando *et al.* (2013).

The sussemione sample ACAD2304, provided by the Australian Centre for Ancient DNA, Adelaïde, Australia, consisted of a bone fragment of an *Equus* (*Sussemionus*) *ovodovi* individual excavated in Proskuriakova Cave (Khakassia, Southwestern Siberia, Russia) from deposits carbon dated at ∼40 000 BP (Eisenmann 2010). We used the original DNA extract from Vilstrup *et al.* (2013).

DNA library preparation

All pre-PCR steps were performed in laboratory facilities dedicated to the analysis of fossil material and geographically separated from post-PCR laboratories. Sequencing libraries were built using a protocol adapted from the Illumina TruSeq DNA Sample Prep Kit (reference FC-121-2001, Illumina, San Diego, CA, USA). End repair of 35 μl of DNA extract was performed using the End-It DNA End-Repair Kit (reference ER0720, Epicentre, Madison, WI, USA) in a 75 μl reaction, using 3 μl End-Repair Enzyme Mix and incubating as follow: 4°C for 2 minutes, 25°C for 45 minutes and 4°C hold. The end-repaired DNA was purified using 1.4× of AMPure XP beads (reference A63880, Beckman Coulter, Brea, CA, USA) and resuspended in 19 μl of nuclease-free water. The supernatant was further processed for 3′ ends adenylation using the TruSeq DNA Sample Prep Kit. Adaptor Ligation was performed by incubating the previous at 30°C for 30 minutes, with 2.5 μl of Ligation Mix, 4.5 μl of nuclease-free water and 7.5 pM adapter: 5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T-3′ and 5′-GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC NNN NNN ATC TCG TAT GCC GTC TTC TGC TTG-3′, where Ns correspond to the library index. After addition of 5 μl of stop ligation buffer, total reaction volume was adjusted to 100 μl with nuclease-free water, purified using 1.8× AMPure XP beads and resuspended in 21 μl of nuclease-free water.

DNA libraries were amplified in a 25 μl volume reaction using 10 μl of DNA library, 5 U AmpliTaq Gold (Life Technologies, ThermoFisher Scientific, Waltham, MA, USA), 1× Gold buffer, 4 mM $MgCl_2$, 1 mg/ml BSA

(Life Technologies), 62.5 μM of each dNTP (Life Technologies), 0.3 μM of Primer 1.0 (5′-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC -3′) and 0.3 μM of Primer 2.0 (5′-CAA GCA GAA GAC GGC ATA CGA GAT-3′). PCR cycling conditions consisted of initial denaturation for 10 minutes at 92°C, followed by 15 cycles (30 seconds denaturation at 92°C, 30 seconds annealing at 65°C and 3 minutes elongation at 72°C) and a final 7 minutes elongation step at 72°C.

Similar conditions were used for PCR amplifications with AccuPrime *Pfx*, where 3.125 enzyme units and 1 mM of $MgSO_4$ were used in the reaction mix. PCR cycling conditions were 2 minutes at 95°C, 15 cycles of 15 seconds at 95°C, 30 seconds at 60°C and 40 seconds at 68°C, followed by a final elongation of 7 minutes at 68°C.

PCR products were purified on a MinElute (QIAGEN, Venlo, Limburg, Netherlands) column and eluted in 20 μl EB following 15 minutes incubation at 37°C. Adapter dimers were removed using a size selection on a 2% NuSieve 3:1 agarose gel (Lonza, Basel, Switzerland) at 80 mV for 1 h 30 minutes, along with a 50 bp ladder (GeneRuler, Thermo Fisher Scientific, Waltham, Massachusetts). Excised fragments were purified using the QIAquick Gel Extraction Kit (QIAGEN, reference 28704) and eluted in 25 μl EB following 15 minutes incubation at 37°C.

Sequencing

Size-selected libraries were checked for size and concentration using a 2100 Bioanalyzer, high-sensitivity DNA assay (Agilent Technologies, Santa Clara, CA, USA), pooled with other non related indexed libraries and sequenced over three lanes of 94 Single-Read (plus a six bases Index Read) on the HiSeq 2000 at the Danish National High-Throughput DNA Sequencing Centre, Copenhagen, Denmark.

Sequence analysis

To avoid bias due to sequencing depth differences, we randomly selected a subset of eight million sequencing reads for each amplified library. The sequence analysis procedure mostly follows a workflow described previously (Schubert *et al.* 2012) and implemented in the PALEOMIX pipeline (Schubert *et al.* 2014). Each set of reads was processed using AdapterRemoval (Lindgreen 2012) in order to remove residual adapter sequences, allowing a mismatch rate of 1/3, and to trim low quality bases at read termini. Reads shorter than 25 bp were not processed further, in order to reduce noise resulting from misalignments. Furthermore, reads were trimmed for palindromic artefacts using the Python script clip_inverted_repeats.py provided by Star *et al.* (2014) and aligned against reference genomes using BWA v0.5.9-r26-dev with standard parameters, except that seeding was disabled (Schubert *et al.* 2012). Only high-quality hits (Phred score $\geqslant$25) were retained. Lastly, aligned reads were filtered for PCR duplicates using MarkDuplicates from the Picard toolkit (http://picard.sourceforge.net/). The horse mitochondrial genome (Accession Nb. NC_001640, Xu et al. 1994) and nuclear genome (EquCab2, excluding mitochondrial DNA; Wade *et al.* 2009) were mapped separately in order to avoid possible numt misidentification. Given that mitochondrial genomes are circular, we followed Orlando *et al.* (2013) and copied the first 30 nucleotidic positions at the end of the sequences to connect sequence ends and enable read alignment across this region. *Post-mortem* damage patterns at read termini were retrieved from mapDamage2 (Jónsson *et al.* 2013) and the number of misincorporations observed in various dinucleotide contexts was determined using a Perl script parsing multifasta alignments between reads and EquCab2 as produced by mapDamage2. All sequence data used in this study are available at the Short Read Archive database, accession number SRP035380.

Results and Discussion

AccuPrime *Pfx* amplification of DNA libraries prepared on fresh extracts or ancient extracts where uracils were excised has previously been shown to limit the base compositional and insert size bias otherwise introduced following AmpliTaq Gold amplication (Dabney and Meyer 2012). We built DNA extracts from two ancient equine bone samples into TruSeq DNA libraries. Half of the DNA library was PCR amplified using AmpliTaq Gold, a standard polymerase in ancient DNA research, whereas the second half was amplified using AccuPrime *Pfx*. We measured endogenous DNA content and clonality levels using a random selection of eight million of Single-End sequence reads per amplified library (Table 1). Strikingly, AmpliTaq Gold amplified libraries showed approximately a two-fold increase in endogenous content compared to AccuPrime *Pfx* amplified libraries (3.05 vs 1.81% for CGG10027 and 7.87 vs 4.13% for sample ACAD2304), and comparable clonality levels (1.08–0.62 and 0.56–0.44%, respectively). This suggested that about half of the ancient templates present in the DNA library were refractory to amplification with AccuPrime *Pfx*.

TruSeq libraries were also found to be associated with significant proportions of palindromic artefacts, as previously reported (Star *et al.* 2014). Interestingly, this proportion was multiplied ~5-fold in the DNA library constructed on the ACAD2304 extract and amplified with AccuPrime *Pfx*, relative to that amplified with AmpliTaq Gold (1.0 vs 0.2%; Table 1). This potentially results from base compositional bias within palindromic sequences, which show a %GC content of 34.7 for the library amplified with AmpliTaq Gold versus 42.5% for that amplified with AccuPrime *Pfx*.

We next explored DNA misincorporation patterns for both amplification types (Fig. 1). DNA libraries amplified with AmpliTaq Gold showed typical misincorporation patterns, consisting of decreasing in C $\rightarrow$ T substitution rates from sequence starts parallelled by an increasing G $\rightarrow$ A substitution rates towards sequence ends. This

Table 1 Sample information and sequence data generated.

Sample	Origin	Date (uncal. BP)	Amplification	#Reads	#Trimming	#Palindromic	#mtDNA	#nuclear	%Endo.	%Clon.
CGG10027	Krasnoyarsk, Taimyr, Russia	24,306 ± 135	AmpliTaq Gold	8 000 000	795 674	18 891 (2.4%)	614	216 224	3.05	1.08
			AccuPrime *Pfx*	8 000 000	1 233 355	36 874 (3.0%)	242	122 279	1.81	0.62
ACAD2304	Proskuriakova Cave, Khakassia, Russia	Late Pleistocene	AmpliTaq Gold	8 000 000	3 686 477	8063 (0.2%)	197	332 183	7.87	0.56
			AccuPrime *Pfx*	8 000 000	3 699 821	38 261 (1.0%)	88	174 463	4.13	0.44

Eight million Single-End sequences were randomly selected from reads generated on a Illumina HiSeq 2000 platform from TruSeq ancient DNA libraries amplified either with AmpliTaq Gold or AccuPrime *Pfx*. Reads were first trimmed for residual adapter sequences and low quality ends (#Trimming) and then for palindromic artefacts (#Palindromic) using PALEOMIX (Schubert *et al.* 2014) and the Python script clip_inverted_repeats.py provided by Star *et al.* (2014), respectively. Trimmed reads were then processed through PALEOMIX (Schubert *et al.* 2014) to identify reads of equine origin. The percentage of endogenous reads (%Endo.) is estimated from the total number of unique hits mapping against the horse nuclear and mitochondrial reference and showing mapping qualities greater or equal to 25. Clonality estimates (%Clon.) are provided in percentage by the proportion of PCR duplicates identified among the total number of reads.

pattern is due to the preferential accumulation in overhanging regions of cytosines that have been deaminated *post-mortem* into uracils and that are read as thymines (Briggs *et al.* 2007).

At sequence starts, the C → T misincorporation patterns observed for DNA libraries amplified with AccuPrime *Pfx* were found to be atypical, showing a 6.4 to 8.4-fold reduction in amplitude at the first sequence position compared to the sequence data derived from AmpliTaq Gold amplified libraries. At sequence ends, the G → A misincorporation pattern observed post-AccuPrime *Pfx* amplification was virtually identical to that observed with AmpliTaq Gold amplified libraries (Fig. 1). It also appeared similar to the nucleotide misincorporation patterns described for other types of DNA library preparation procedures, such as those based on blunt-end ligation (Meyer and Kircher 2010), except that the second position from the ends showed lower G → A misincorporation rates than the third one. G → A misincorporation rates were found to be lower at the last read position post-trimming than C → T misincorporation rates at sequencing starts (Fig. 1). This is likely due to the fact that a significant fraction of inserts were not sequenced over their full length, as shown by their size distribution, which shows a significant fraction of reads of maximal size (94 bp, Supplementary Material 1). For those reads, the last sequence position available post-trimming is not necessarily part of an overhang and therefore does not show increased deamination rates.

We hypothesised that the ca. two-fold reduction in endogenous DNA content and the loss of the expected damage profile towards sequence starts originated from the incapacity to amplify one of the two original strands present in the ancient DNA insert. In particular, it has been demonstrated that the AccuPrime *Pfx* cannot amplify DNA templates containing uracil (Lasken, Schuster and Rashtchian 1996). As a result, DNA strands showing cytosine deamination would be excluded from the pool of molecules sequenced, in contrast to their complementary strand (Fig. 2A). Since most cytosine deamination reactions take place within overhanging ends (Briggs *et al.* 2007) and 3′-overhangs, albeit present (Orlando *et al.* 2011; Prüfer et al. 2014), are eliminated during library building, only cytosine deamination events located at 5′-overhangs are detectable in our protocol. These have been filled-in with complementary adenines during the end-repair step and following AccuPrime *Pfx* amplification, those will appear as G → A substitutions towards sequence ends. As AmpliTaq Gold can bypass uracils, those will appear both as C → T substitutions towards sequence starts and G → A substitutions towards sequence ends, leading to misincorporation patterns reminiscent of what observed for ancient DNA libraries built following blunt-end ligation (Meyer and Kircher 2010, Figure S2).

CpG sites offer a unique opportunity to validate our model, as they are the main target of cytosine

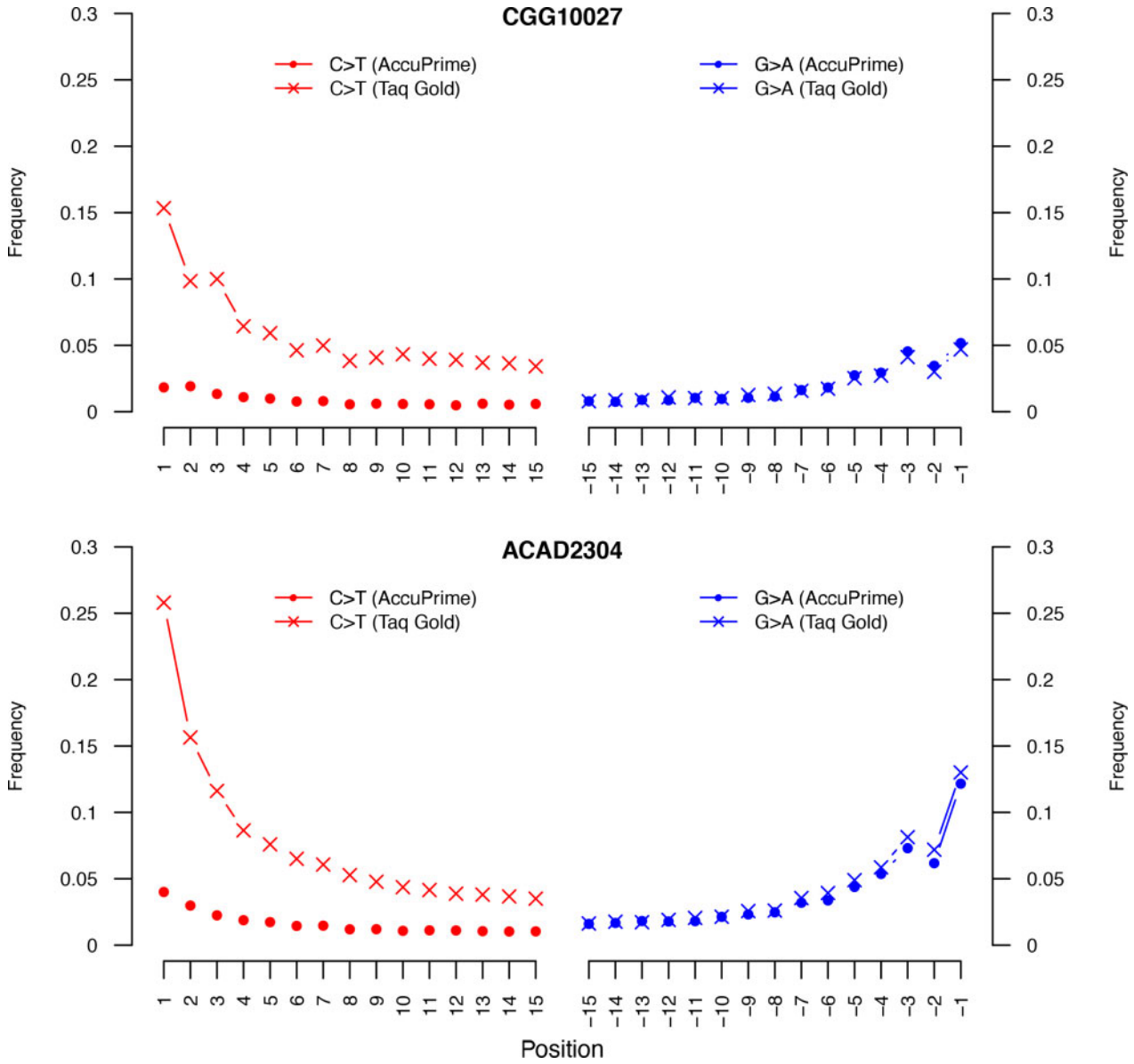


Figure 1 Nucleotide misincorporation patterns following AmpliTaq Gold and AccuPrime *Pfx* amplification of TruSeq ancient DNA libraries. The rate of C → T substitutions observed between the reference genome and sequencing reads is shown for the first 15 (left) and last 15 (right) positions sequenced. Those rates were calculated using mapDamage2 (Jonsson *et al.* 2013). *Top:* horse sample CGG10027. *Bottom:* sussemione sample ACAD2304.

methylation in mammalian genomes (Plongthongkum, Diep and Zhang 2014). Post deamination of cytosines, unmethylated CpG sites are transformed into UpG, whereas ^{5m}CpG are transformed into TpG (Pedersen *et al.* 2013; Gokhman *et al.* 2014; Smith *et al.* 2014). Following our hypothesis, only the latter would provide adequate templates for AccuPrime *Pfx* amplification, which would restore the expected C → T misincorporation patterns towards sequence starts (Fig. 2B). We therefore predicted that the C → T substitutions observed at sequence starts would occur preferentially in a CpG context. We counted in which dinucleotide context each C → T substitution observed in the first 10 positions along sequencing reads took place. For sequences generated following AmpliTaq Gold amplification, such substitutions occur in only a minority of CpG sites (Fig. 3), with CpG representing only 7.2–10.4% of the four possible CpN dinucleotide contexts. This situation appeared in striking contrast to that observed following AccuPrime *Pfx* amplification, where 34.2–41.6% of C → T substitutions were incorporated at CpG sites, in agreement with the prediction of our model. The C → T changes observed in other dinucleotide contexts likely reflected sequencing errors and/or sites where the ancient samples differed from the horse genome reference used for read alignment.

No significant difference was observed towards sequence ends, where CpG dinucleotides accounted for ca. 8.3–8.8% (CGG10027) and ca. 12.6–13.7% (ACAD 2304) of the G → A substitutions, regardless of which polymerase was used. This is also in agreement

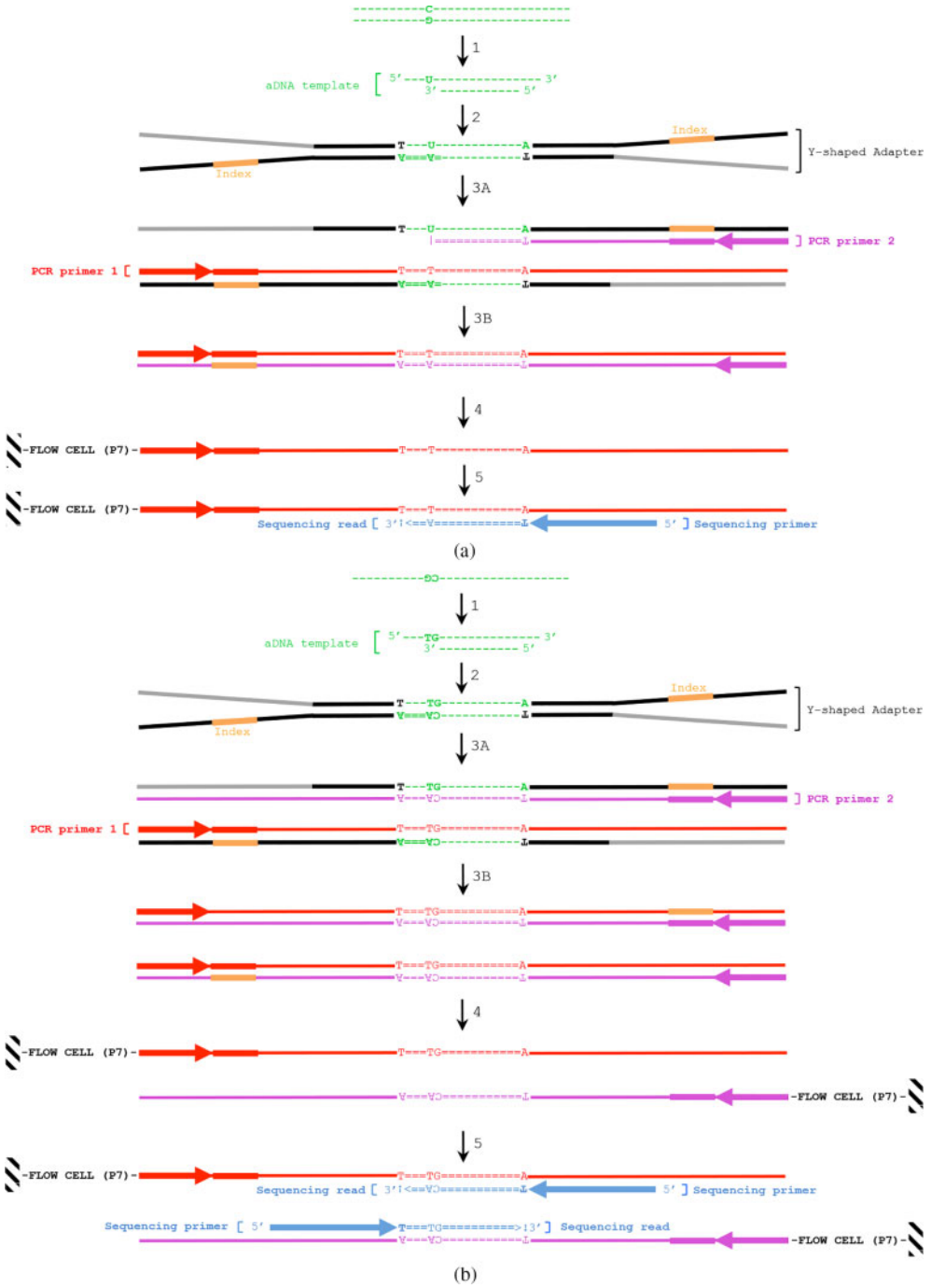


Figure 2 AccuPrime *Pfx* amplification of TruSeq ancient DNA libraries results in asymmetrical cytosine deamination misincorporation patterns. (A) *Step 1: post-mortem* DNA hydrolysis results in the formation of 5′-overhangs with deaminated cytosine residues (i.e. uracils) (Briggs *et al.* 2007). When present, 3′-overhangs are polished during library construction. *Step 2:* a TruSeq DNA library is prepared. Adenine residues, complementary to uracils, are incorporated during end repair. *Step 3A:* first PCR amplification cycle of the TruSeq DNA library. The AccuPrime *Pfx* DNA polymerase does not by-pass uracils. As a result, the damaged strand is not copied in contrast to its undamaged complementary strand, which incorporates a C → T substitution. *Step 3B:* second PCR amplification cycle. DNA templates identical to those described in step 3A are formed but not indicated for clarity. Elongation from the second PCR primer results in the copy of the strand with a C → T misincorporation. *Step 4:* Following cluster generation, the strand starting with the P7 sequence is linked to the flow-cell. *Step 5:* illumina sequencing, resulting in the detection of a G → A substitution at sequence ends. (B) Same as (A), except that cytosine deamination occurred at a methylated CpG site, which creates a TpG site in the ancient DNA template. As a result, both library strands are copied in the first amplification cycle and C → T and G → A misincorporations are observed both at sequence starts and sequence ends (except when library inserts have not been sequenced over their full length). Note that both sequencing reactions reported occur at independent clusters.

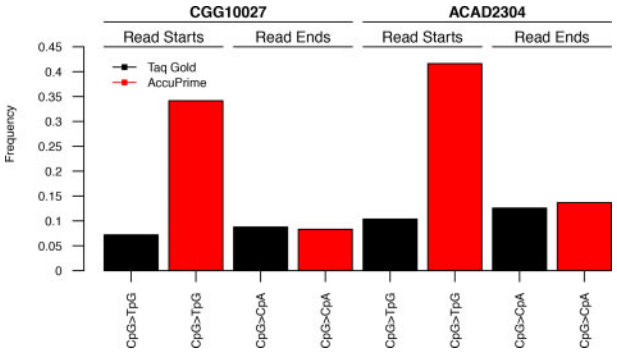


Figure 3 Comparison of C → T misincorporation rates across different dinucleotide contexts. The number of C → T substitutions observed between the reference genome and sequencing reads was counted for the first 10 positions sequenced and broken down by dinucleotide context (CpA, CpC, CpG and CpT). The same analysis was performed for complementary G → A substitutions at the last 10 positions sequenced and reverse complementary dinucleotide contexts. The relative importance of CpG → TpG misincorporations at read starts and CpG → CpA misincorporations at read ends are shown, as only those are related to CpG methylation.

with our model, as sequence ends record the presence of cytosine deamination located on the complementary strand and fixed as a G → A during end-repair (Fig. 2A,B).

Overall, we demonstrated that AccuPrime *Pfx* precludes amplifying DNA strands of TruSeq DNA libraries for which at least one cytosine has been deaminated *post-mortem*. This results in the loss of up to half of the endogenous templates originally ligated to adapters. This method will therefore substantially increase the sequencing costs underlying ancient genome characterisation and should therefore not be recommended in cases where significant deamination levels are observed. The amplification of ancient DNA libraries with AccuPrime *Pfx* could, however, still be used with no reduction in endogenous DNA content in situations where DNA damage by-products have been eliminated prior to DNA library building, using UNG and EndoVIII enzymatic treatments that cleave DNA templates 3′ of uracils (Briggs *et al.* 2010). AccuPrime *Pfx* will then limit over-amplifying short and %GC-rich DNA inserts and will therefore preserve the original molecular complexity of the DNA libraries more efficiently than a variety of other DNA polymerases (Dabney and Meyer 2012). We also note that amplifying single-stranded DNA libraries (Gansauge and Meyer 2013) with AccuPrime *Pfx* will not result in a significant drop in endogenous DNA content (Meyer *et al.* 2013), because the original DNA strand is first copied using *Bst* DNA polymerase, which results in a G → A misincorporation, before a double-stranded DNA library is built and PCR amplified. This is in line with the great performance observed for this procedure on range of ancient DNA extracts (Meyer *et al.* 2013; Prüfer *et al.* 2014; Bennett *et al.* 2014).

Sequencing ancient genomes following AccuPrime *Pfx* amplification of TruSeq DNA libraries could, however, remain cost-effective in cases where *post-mortem* deamination reactions were limited and/or where the microbial colonisation of the extracts was minimal. The sequence data generated then can also provide access to ancient methylation marks, following a principle similar to the work of Pedersen and colleagues (2013) and Gokhman and colleagues (2014). In particular, the procedure described by Pedersen and colleagues (2013) exploited C → T misincorporations at sequence starts to measure regional methylation levels in the sequence data underlying the genome of a 4,000 year-old paleo-Eskimo individual (Rasmussen *et al.* 2010). These sequences were generated following the original Illumina sample preparation protocol, which included amplification with *Phusion* DNA polymerase (Finnzymes Oy, Thermo Fisher Scientific, Waltham, Massachusetts), also found to be incapable of copying uracils (Rasmussen *et al.* 2010). Additionally, Pedersen and colleagues (2013) exploited lateral variation of depth-of-coverage to reconstruct whole-genome nucleosome maps. However, this required prior correction of the strong %GC-bias introduced during *Phusion* amplification. Such biases are known to be minimal with AccuPrime *Pfx* (Dabney and Meyer 2012). Therefore, the procedure described here, coupled or not with UNG and EndoVIII enzymatic treatments, holds the potential for uncovering ancient nucleosome maps, where depth-of-coverage peaks along the genome reflect nucleosome protection and occupancy.

Recently, sequencing libraries based on rapid ligation at AT-overhangs were found to show a strong base compositional bias, with preferential ligation of inserts starting at adenine *versus* thymine or its analogue uracil (Seguin-Orlando *et al.* 2013). This bias was observed both on fresh and ancient DNA extracts. TruSeq DNA libraries also rely on ligation at AT-overhangs but do not show the base compositional bias described following rapid ligation, as the highest C → T (G → A) misincorporation rates are observed at the first (last) position within sequencing reads (Fig 1). This suggests that the DNA ligase involved and/or the ligation conditions (30 minutes at 30°C vs 10 minutes at 25°C), rather than ligation at AT-overhangs *per se*, are the sources for the base compositional bias previously reported. Overall, the TruSeq protocol is less subjected to introduce a strong base compositional bias than the rapid ligation approach and therefore should be preferred while constructing Illumina DNA libraries based on ligation at AT-overhangs.

Finally, the presence of typical nucleotide misincorporation patterns has been used with great success for authenticating ancient DNA datasets (Krause *et al.* 2010; Orlando *et al.* 2011; Raghavan *et al.* 2013). With standard DNA library preparation methods adapted from the 454 procedure, where DNA templates are first end repaired, then ligated to adapters and finally filled-in (Meyer *et al.* 2010), this pattern consists of an increase of C → T substitutions towards

sequence starts and an increase of G → A substitutions towards sequence ends. However, other patterns are expected depending on the molecular tools used for preparing DNA libraries (Seguin-Orlando *et al.* 2013; Prüfer *et al.* 2014) and/or the sequencing platform considered (Orlando *et al.* 2011). In this study, we identified another of such deviations to the standard pattern, which should be considered as the expectation for authenticating ancient DNA datasets generated using similar molecular tools.

Conflicts of interest

Ludovic Orlando is Editor-in-Chief at *STAR*, Beth Shapiro and Alan Cooper are members of the Editorial Board at *STAR*.

Author biographies

Andaine Seguin-Orlando is a PhD student at the Centre of Excellence in GeoGenetics, University of Copenhagen. She works on developing new approaches to improve access to ancient genomes.

Dr. Cindi Hoover currently works in the Synthetic Biology Group at the DOE Joint Genome Institute, University of California. Her research interests focus on transcriptome profiling and functional genomics of environmental and marine organisms.

Dr. Sergei Vasiliev is paleontologist at the Russian Academy of Sciences.

Dr. Nikolai Ovodov is paleontologist at the Russian Academy of Sciences.

Dr. Beth Shapiro is associate professor in Paleogenomics Lab of the University of California, Santa Cruz. Her research aims to better understand how populations and species change through time, in particular in response to environmental and other changes to their habitat.

Pr. Alan Cooper is the director of the Australian Centre for Ancient DNA at the University of Adelaide. His work focuses on the use of ancient DNA to record and study evolutionary processes in real time, especially those associated with environmental change.

Dr. Edward Rubin is the director of the DOE Joint Genome Institute, University of California. His research has involved the development and application of computational and biological approaches for studying genomes.

Pr. Eske Willerslev is the director of Centre of Excellence in GeoGenetics, University of Copenhagen. His research interests include: DNA degradation, megafaunal extinctions, ancient and modern human populations genomics, microbial evolution, discovery of novel pathogens and environmental DNA.

Dr. Ludovic Orlando is associate professor at the Centre of Excellence in GeoGenetics, University of Copenhagen. His group develops integrative approaches in ancient DNA research, promoting the field of palaeomics by the merger of biochemistry, molecular biology, genomics and computational biology.

Acknowledgements

We thank the laboratory technicians of the Centre for GeoGenetics and the staff of the Danish National High-Throughput DNA Sequencing Centre for technical assistance.

This work was supported by the Danish Council for Independent Research, Natural Sciences (FNU, 4002–00152B); the Danish National Research Foundation (DNFR94) and a Marie Curie Career Integration Grant (FP7 CIG–293845).

References

Bennett, E. A., D. Massilani, G. Lizzo, J. Daligault, E. M. Geigl, and T. Grange. 2014. "Library Construction for Ancient Genomics: Single Strand or Double Strand?" *BioTechniques* 56: 289–300.

Briggs, A. W., U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, and S. Pääbo. 2007. "Patterns of Damage in Genomic DNA Sequences from a Neandertal." *Proceedings of the National Academy of Sciences of the United States of America* 104: 14616–14621.

Briggs, A. W., U. Stenzel, M. Meyer, J. Krause, M. Kircher, and S. Pääbo. 2010. "Removal of Deaminated Cytosines and Detection of in Vivo Methylation in Ancient DNA." *Nucleic Acids Res* 38: e87.

Cui, Y., J. Lindo, C. E. Hughes, J. W. Johnson, A. G. Hernandez, B. M. Kemp, J. Ma, R. Cunningham, B. Petzelt, J. Mitchell, D. Archer, J. S. Cybulski, and R. S. Malhi. 2013. "Ancient DNA Analysis of Mid-Holocene Individuals from the Northwest Coast of North America Reveals Different Evolutionary Paths for Mitogenomes." *PLoS ONE* 8 (7): e66948.

Dabney, J., and M. Meyer. 2012. "Length and GC-biases During Sequencing Library Amplification: A Comparison of Various Polymerase-buffer Systems with Ancient and Modern DNA Sequencing Libraries." *BioTechniques* 52: 87–94.

Dabney, J., M. Knapp, I. Glocke, M. T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. Garcia, S. Pääbo, J. L. Arsuaga, and M. Meyer. 2013. "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: 15758–15763.

Der Sarkissian, C., M. E. Allentoft, M. C. Ávila-Arcos, R. Barnett, P. F. Campos, E. Cappellini, L. Ermini, R. Fernández, R. da Fonseca, A. Ginolhac, A. J. Hansen, H. Jónsson, T. Korneliussen, A. Margaryan, M. D. Martin, J. V. Moreno-Mayar, M. Raghavan, M. Rasmussen, M. S. Velasco, H. Schroeder, M. Schubert, M. Schubert, A. Seguin-Orlando, N. Wales, M. T. P. Gilbert, E. Willerslev, and L. Orlando. 2015. "Ancient Genomics." *Philosophical Transactions of the Royal Society of London* 370 (1660): 20130387.

Eisenmann, V. 2010. "Sussemionus, a New Subgenus of Equus (Perissodactyla, Mammalia)." *Comptes Rendus Biologies* 333: 235–240.

Ermini, L., C. Der Sarkissian, E. Willerslev, and L. Orlando. 2015. "Major Transitions in Human Evolution Revisited: A Tribute to Ancient DNA." *Journal of Human Evolution* 79: 4–20.

Gansauge, M. T., and M. Meyer. 2013. "Single-stranded DNA Library Preparation for the Sequencing of Ancient or Damaged DNA." *Nature Protocols* 8: 737.

Gokhman, D., E. Lavi, K. Prüfer, M. F. Fraga, J. A. Riancho, J. Kelso, S. Pääbo, E. Meshorer, and L. Carmel. 2014. "Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan." *Science* 344: 523–527.

Holmqvist, E., J. Reimegaard, and E. G. H. Wagner. 2013. "Massive Functional Mapping of a 5'-UTR by Saturation Mutagenesis, Phenotypic Sorting and Deep Sequencing." *Nucleic Acids Research* 41: e122.

Höss, M., P. Jaruga, T. H. Zastawny, M. Dizdaroglu, and S. Pääbo. 1996. "DNA Damage and DNA Sequence Retrieval from Ancient Tissues." *Nucleic Acids Research* 24 (7): 1304–1307.

Jónsson, H., A. Ginolhac, M. Schubert, P. L. Johnson, and L. Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics* 29 (13): 1682–1684.

Krause, J., Q. Fu, J. M. Good, B. Viola, M. V. Shunkov, A. P. Derevianko, and S. Pääbo. 2010. "The Complete Mitochondrial DNA Genome of an Unknown Hominin from Southern Siberia." *Nature* 464: 894–897.

Lasken, R. S., D. M. Schuster, and A. Rashtchian. 1996. "Archaebacterial DNA Polymerases Tightly Bind Uracil-Containing DNA." *The Journal of Biological Chemistry* 271: 17692–17696.

Lindgreen, S. 2012. "AdapterRemoval: Easy Cleaning of Next-Generation Sequencing Reads." *BMC Research Notes* 5: 337.

Meyer, M., and M. Kircher. 2010. "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing." *Cold Spring Harbor Protocols* 2010 (6):, pdb.prot5448.

Meyer, M., M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. 2012. "A High-coverage Genome Sequence from an Archaic Denisovan Individual." *Science* 338: 222–226.

Meyer, M., Q. Fu, A. Aximu-Petri, I. Glocke, B. Nickel, J. L. Arsuaga, I. Martinez, A. Gracia, J. M. de Castro, E. Carbonell, and S. Pääbo. 2013. "A Mitochondrial Genome Sequence of a Hominin from Sima De Los Huesos." *Nature* 505 (7483): 403–406.

Orlando, L., A. Ginolhac, M. Raghavan, J. Vilstrup, M. Rasmussen, K. Magnussen, K. E. Steinmann, P. Kapranov, J. F. Thompson, G. Zazula, D. Froese, I. Moltke, B. Shapiro, M. Hofreiter, K. A. Al-Rasheid, M. T. Gilbert, and E. Willerslev. 2011. "True Single-Molecule DNA Sequencing of a Pleistocene Horse Bone." *Genome Research* 21: 1705–1719.

Orlando, L., A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubbert, E. Cappellini, B. Petersen, I. Moltke, P. L. Johnson, M. Fumagalli, J. T. Vilstrup, M. Raghavan, T. Korneliussen, A. S. Malaspinas, J. Vogt, D. Szklarczyk, C. D. Kelstrup, J. Vinther, A. Dolocan, J. Stenderup, A. M. Velazquez, J. Cahill, M. Rasmussen, X. Wang, J. Min, G. D. Zazula, A. Seguin-Orlando, C. Mortensen, K. Magnussen, J. F. Thompson, J. Weinstock, K. Gregersen, K. H. Røed, V. Eisenmann, C. J. Rubin, D. C. Miller, D. F. Antczak, M. F. Bertelsen, S. Brunak, K. A. Al-Rasheid, O. Ryder, L. Andersson, J. Mundy, A. Krogh, M. T. Gilbert, K. Kjær, T. Sicheritz-Ponten, L. J. Jensen, J. V. Olsen, M. Hofreiter, R. Nielsen, B. Shapiro, J. Wang, and E. Willerslev. 2013. "Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse." *Nature* 499: 74–78.

Pedersen, J. S., E. Valen, A. M. Vargas Velazquez, B. J. Parker, M. Rasmussen, S. Lindgreen, B. Lilje, D. J. Tobin, T. K. Kelly, S. Vang, R. Andersson, P. A. Jones, C. A. Hoover, A. Tikhonov, E. Prokhortchouk, E. M. Rubin, A. Sandelin, M. T. Gilbert, A. Krogh, E. Willerslev, and L. Orlando. 2013. "Genome-wide Nucleosome Map and Cytosine Methylation Levels of an Ancient Human Genome." *Genome Research* 24 (3): 454–466.

Plongthongkum, N., D. H. Diep, and K. Zhang. 2014. "Advances in the Profiling of DNA Modifications: Cytosine Methylation and Beyond." *Nature Reviews Genetics* 15 (10): 647–661.

Poinar, H. N., C. Schwarz, J. Qi, B. Shapiro, R. D. Macphee, B. Buigues, A. Tikhonov, D. H. Huson, L. P. Tomsho, A. Auch, M. Rampp, W. Miller, and S. C. Schuster. 2006. "Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA." *Science* 311 (5759): 392–394.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, and C. Theunert. 2014. "The Complete Genome Sequence of a Neanderthal from the Altai Mountains." *Nature* 505 (7481): 43–49.

Raghavan, M., P. Skoglund, K. E. Graf, M. Metspalu, A. Albrechtsen, I. Moltke, S. Rasmussen, T. W. Stafford, L. Orlando, E. Metspalu, M. Karmin, K. Tambets, S. Rootsi, R. Mägi, P. F. Campos, E. Balanovska, O. Balanovsky, E. Khusnutdinova, S. Litvinov, L. P. Osipova, S. A. Fedorova, M. I. Voevoda, M. DeGiorgio, T. Sicheritz-Ponten, S. Brunak, S. Demeshchenko, T. Kivisild, R. Villems, R. Nielsen, M. Jakobsson, and E. Willerslev. 2013. "Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans." *Nature* 505 (7481): 87–91.

Rasmussen, M., Y. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, and M. Bunce. 2010. "Ancient Human Genome Sequence of an Extinct Palaeo-Eskimo." *Nature* 463: 757–762.

Ruark, E., K. Snape, P. Humburg, C. Loveday, I. Bajrami, R. Brough, D. N. Rodrigues, A. Renwick, S. Seal, E. Ramsay, V. Duarte Sdel, M. A. Rivas, M. Warren-Perry, A. Zachariou, A. Campion-Flora, S. Hanks, A. Murray, N. Ansari Pour, J. Douglas, L. Gregory, A. Rimmer, N. M. Walker, T. P. Yang, J. W. Adlard, J. Barwell, J. Berg, A. F. Brady, C. Brewer, G. Brice, C. Chapman, J. Cook, R. Davidson, A. Donaldson, F. Douglas, D. Eccles, D. G. Evans, L. Greenhalgh, A. Henderson, L. Izatt, A. Kumar, F. Lalloo, Z. Miedzybrodzka, P. J. Morrison, J. Paterson, M. Porteous, M. T. Rogers, S. Shanley, L. Walker, M. Gore, R. Houlston, M. A. Brown, M. J. Caufield, P. Deloukas, M. I. McCarthy, J. A. Todd, Breast and Ovarian Cancer Susceptibility Collaboration; Wellcome Trust Case Control Consortium, C. Turnbull, J. S. Reis-Filho, A. Ashworth, A. C. Antoniou, C. J. Lord, P. Donnelly, and N. Rahman. 2013. "Mosaic PPM1D Mutations are Associated with Predisposition to Breast and Ovarian Cancer." *Nature* 493: 406–410.

Sampietro, M. L., M. T. P. Gilbert, O. Lao, D. Caramelli, M. Lari, J. Bertranpetit, and C. Lalueza-Fox. 2006. "Tracking Down Human Contamination in Ancient Human Teeth." *Molecular Biology and Evolution* 23: 1801–1807.

Schubert, M., A. Ginolhac, S. Lindgreen, J. F. Thompson, K. A. Al-Rasheid, E. Willerslev, A. Krogh, and L. Orlando. 2012. "Improving Ancient DNA Read Mapping Against Modern Reference Genomes." *BMC Genomics* 13: 178.

Schubert, M., L. Ermini, C. Der Sarkissian, H. Jónsson, A. Ginolhac, R. Schaefer, M. D. Martin, R. Fernández, M. Kircher, M. McCue, E. Willerslev, and L. Orlando. 2014. "Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX." *Nature Protocols* 9 (5): 1056–1082.

Schuenemann, V. J., P. Singh, T. A. Mendum, B. Krause-Kyora, G. Jäger, K. I. Bos, A. Herbig, C. Economou, A. Benjak, P. Busso, A. Nebel, J. L. Boldsen, A. Kjellström, H. Wu, G. R. Stewart, G. M. Taylor, P. Bauer, O. Y. Lee, H. H. Wu, D. E. Minnikin, G. S. Besra, K. Tucker, S. Roffey, S. O. Sow, S. T. Cole, K. Nieselt, and J. Krause. 2013. "Genome-wide Comparison of Medieval and Modern Mycobacterium Leprae." *Science* 341: 179–183.

Seguin-Orlando, A., M. Schubert, J. Clary, J. Stagegaard, M. T. Alberdi, J. L. Prado, A. Prieto, E. Willerslev, and L. Orlando. 2013. "Ligation Bias in Illumina Next-generation DNA Libraries: Implications for Sequencing Ancient Genomes." *PLoS One* 8: e78575.

Shapiro, B., and M. Hofreiter. 2014. "A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA." *Science* 343 (6169): 1236573.

Smith, O., A. J. Clapham, P. Rose, Y. Liu, J. Wang, and R. G. Allaby. 2014. "Genomic Methylation Patterns in Archaeological Barley Show De-Methylation as a Time-dependent Diagenetic Process." *Scientific Reports* 4: 5559.

Star, B., A. J. Nederbragt, M. H. Hansen, M. Skage, G. D. Gilfillan, I. R. Bradbury, C. Pampoulie, N. C. Stanseth, K. S. Jakobsen, and S. Jentoft. 2014. "Palindromic Sequence Artifacts Generated During Next Generation Sequencing Library Preparation from Historic and Ancient DNA." *PLoS One* 9 (3): e89676.

Vilstrup, J. T., A. Seguin-Orlando, M. Stiller, A. Ginolhac, M. Raghavan, S. C. A. Nielsen, J. Weinstock, D. Froese, S. K. Vasiliev, N. D. Ovodov, J. Clary, K. M. Helgen, R. C. Fleischer, A. Cooper, B. Shapiro, and L. Orlando. 2013. "Mitochondrial Phylogenomics of Modern and Ancient Equids." *PLoS One* 8: e55950.

Wade, C. M., E. Giulotto, S. Sigurdsson, M. Zoli, S. Gnerre, F. Imsland, T. L. Lear, D. L. Adelson, E. Bailey, R. R. Bellone, H. Blöcker, O. Distl, R. C. Edgar, M. Garber, T. Leeb, E. Mauceli, J. N. MacLeod, M. C. Penedo, J. M. Raison, T. Sharpe, J. Vogel, L. Andersson, D. F. Antczak, T. Biagi, M. M. Binns, B. P. Chowdhary, S. J. Coleman, G. Della Valle, S. Fryc, G. Guérin, T. Hasegawa, E. W. Hill, J. Jurka, A. Kiialainen, G. Lindgren, J. Liu, E. Magnani, J. R. Mickelson, J. Murray, S. G. Nergadze, R. Onofrio, S. Pedroni, M. F. Piras, T. Raudsepp, M. Rocchi, K. H. Røed, O. A. Ryder, S. Searle, L. Skow, J. E. Swinburne, A. C. Syvänen, T. Tozaki, S. J. Valberg, M. Vaudin, J. R. White, M. C. Zody, Broad Institute Genome Sequencing Platform; Broad Institute Whole Genome Assembly Team, E. S. Lander, and K. Lindblad-Toh. 2009. "Genome Sequence, Comparative Analysis, and Population Genetics of the Domestic Horse." *Science* 326: 865–867.

Xu, X., and U. Arnason. 1994. "The Complete Mitochondrial DNA Sequence of the Horse, *Equus Caballus*: Extensive Heteroplasmy of the Control Region." *Gene* 148: 357–362.