STAR: Science & Technology of Archaeological Research

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

Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples

Bastiaan Star, Marianne HS Hansen, Morten Skage, Ian R. Bradbury, Jane A. Godiksen, Olav S. Kjesbu & Sissel Jentoft

To cite this article: Bastiaan Star, Marianne HS Hansen, Morten Skage, Ian R. Bradbury, Jane A. Godiksen, Olav S. Kjesbu & Sissel Jentoft (2016) Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples, STAR: Science & Technology of Archaeological Research, 2:1, 36-45, DOI: 10.1080/20548923.2016.1160594

To link to this article: http://dx.doi.org/10.1080/20548923.2016.1160594

Science and Technology of Archaeological Research

Short report

Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples

Bastiaan Star[1]*, Marianne HS Hansen[1], Morten Skage[1], Ian R. Bradbury[2], Jane A. Godiksen[3], Olav S. Kjesbu[1,3], and Sissel Jentoft[1]

[1]Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biosciences, University of Oslo, PO Box 1066, Blindern, N-0316 Oslo, Norway.

[2]Fisheries and Oceans Canada, 80 East White Hills Road, St. John's, NL, Canada A1C 5X1.

[3]Institute of Marine Research (IMR) and Hjort Centre for Marine Ecosystem Dynamics, PO Box 1870, N-5817 Bergen, Norway.

Abstract Repetitive microsatellite DNA forms a universal component of eukaryote genomes and specific biochemical properties of such repeat regions may influence the outcome of laboratory protocols. The Atlantic cod (*Gadus morhua*) genome contains an order of magnitude more dinucleotide repeats than the majority of vertebrates, with over eight percent of its genome that can be classified as either AC or AG dinucleotide repeat. We find that the abundance of these repeats can be inflated in ancient DNA (aDNA) whole genome sequencing (WGS) data generated from this species, in particular in samples with a lower fragment length. This inflation is suppressed by a reduced number of amplification cycles and by the inclusion of manufactured dinucleotide repeat oligonucleotides during amplification. These data indicate that a biased amplification reaction leads to artificially high levels of AC and AG repeats. This process appears to be particularly efficient in Atlantic cod –likely due to its high genomic content of repeats with relatively simple sequence complexity. While the extend of such bias in other studies is unclear, we nonetheless urge caution when quantifying repeat content in aDNA WGS data, given that amplification bias can be difficult to detect if this process affects more complex repeat structures than dinucleotide repeats.

Keywords dinucleotide repeats; self-priming; repetitive DNA; amplification bias; aDNA; Atlantic cod

Received 23 June 2015; **accepted** 12 January 2016

Introduction

Microsatellite DNA or short tandem repeats (STRs) that iterate short motifs of less than 6 base pair (bp) form a universal component of eukaryote genomes (Tautz and Renz 1984, Ellegren 2004, Amos and Clarke 2008). Microsatellites occur in a range of different compositions, ranging from perfect stretches of simple mono- or dinucleotide repeats to complex compound combinations of imperfect repeats (Chambers and MacAvoy 2000). While compound microsatellites are found more frequently than expected by chance alone (Kofler et al. 2008), the majority of microsatellites in vertebrate genomes occur as dinucleotide repeats, with AC, AG and AT being the most common type, and with GC repeats being rare (Ellegren 2004). Their widespread occurrence and high level of individual variation have made microsatellites a popular genetic tool for an impressive range of biological applications (Tautz 1989, Chambers and MacAvoy 2000), even though microsatellite evolution itself is not fully understood (Buschiazzo and Gemmell 2006, Bhargava and Fuentes 2010).

The repetitive nature of microsatellites provides several challenges that make their genome wide analysis difficult. Most obviously, if sequencing reads do not span the repeat region, the algorithms used in genome assembly or read mapping cannot resolve the repeat pattern correctly (Gymrek et al. 2012). Moreover, their detection in genomic data is not straightforward and consequently a range of algorithms has been developed to address this issue (Merkel and Gemmell 2008). The bioinformatics issues associated with analyzing microsatellites repeats are therefore well recognized (e.g., Schaper et al. 2012). Nevertheless, the peculiar biochemical properties of microsatellites and in particular those of dinucleotide repeats are not often considered to affect whole genome sequencing (WGS) approaches.

Unusually high levels of AC and AG dinucleotide repeats have previously been demonstrated in ancient

*Corresponding author. email: bastiaan.star@ibv.uio.no

DOI: 10.1080/20548923.2016.1160594

DNA (aDNA) sequencing libraries generated from Atlantic cod (*Gadus morhua*) samples (Figure S5, Star et al. 2014). Such levels are problematic, because reads containing repeats cannot be reliably mapped towards a reference genome, reducing the proportion of endogenously classified reads. Given that sequencing aDNA samples is a relatively expensive endeavor, efficiency of library protocols is of prime interest when handling such samples.

It is well known that postmortem degradation and laboratory protocols introduce systematic sequence bias in data generated from aDNA samples. For instance, degradation leads to enhanced cytosine deamination in single stranded 5′-overhangs (Briggs et al. 2007, Green et al. 2009, Krause et al. 2010, Ginolhac et al. 2011) and to an increased proportion of purines at the positions immediately preceding the 5′ termini of DNA fragments (Briggs et al. 2007, Overballe-Petersen, Orlando, and Willerslev 2012, Meyer et al. 2012). Moreover, aDNA sequencing reads are biased in GC content depending on fragment length and type of Taq polymerase used (Green et al. 2008, Briggs et al. 2009, Meyer and Kircher 2010, Dabney and Meyer 2012) and are selected against starting with a thymine residue when using AT-overhang ligation protocols (Seguin-Orlando et al. 2013). Biases currently known to affect aDNA sequencing, however, do not explain the large proportion of AC and AG repeats detected in the aDNA sequencing data from Atlantic cod (Star et al. 2014).

Interestingly, considerably longer microsatellites have been reported in fish (Neff and Gross 2001) compared to other vertebrates, and the Atlantic cod genome in particular is rich in simple repeats (Jiang et al. 2014). We wondered whether the endogenously occurring dinucleotide repeats in the Atlantic cod genome could be causal to their overrepresentation in the aDNA sequencing data, perhaps due to a preferential preservation, ligation or amplification or a combination of these factors.

Here we characterize AC and AG repeats in a range of vertebrate genomes and in contemporary and aDNA sequencing data from Atlantic cod and several other species. First, we show that the Atlantic cod genome assembly contains an order of magnitude more AC and AG repeats than the majority of vertebrate genomes and that –depending on sample fragmentation– these values can be highly inflated sequencing data from historic samples compared to those from contemporary material in Atlantic cod. Second, we show that amplification conditions strongly influence the relative proportion of these types of repeats. Third, the proportion of these repeats can be altered through the inclusion of artificially made repetitive oligonucleotides during the amplification of WGS libraries generated from aDNA samples. We further investigate if similar patterns can be observed in publicly available contemporary and ancient DNA sequence data. Based on our results, we suggest two hypotheses that may allow the biased proliferation of repeats, leading to those levels observed in our sequencing data. These observations highlight the methodological complications that can be encountered when targeting species with divergent genomic compositions.

Material and Methods

Dinucleotide repeat estimation

We investigated variation in dinucleotide repeat content using several publicly available resources. First, complete vertebrate assemblies (n = 63) were obtained from Ensembl (release 76, see Supplementary Table 1). Primary assembly files were used for human and mouse while DNA toplevel fasta files were used for all other species. For salmon we used the ASM23337v1 assembly (http://www.icisb.org/atlantic-salmon-genome-sequence/). For rainbow trout we used the assembly described in Berthelot et al. (2014). Dinucleotide repeat content was estimated using Tandem Repeat Finder (TRF, version 407b) with settings '2 2 7 80 10 20 2 -ngs', which identifies repeats up to periodicity of two base pairs (bp) with a minimum length of 10 bp. Dinucleotide repeat estimates were summed for each respective repeat type, including their reverse complement (e.g. Total *AC* content = *AC* + *CA* + *GT* + *TG*). The total number of identified dinucleotide repeats was divided by the amount of bases in each assembly, excluding any undermined bases in gaps.

Second, we investigated publicly available contemporary Illumina Hiseq read data from six species (cat, dog, rabbit, rat, rainbow trout and human). Up to 30 individual fastq.gz files were randomly selected per species from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena). Specifically for dog, we excluded read data from tumor samples. Dinucleotide repeat estimates were obtained as described above using a subset of 1 million reads obtained with the 'sample' subcommand of seqtk (https://github.com/lh3/seqtk version of Oct 16, 2012, commit hash d43d3704d4). These estimates were compared to the respective repeat composition in each species' genome assembly.

Third, short-read sequence data (282 single-end fastq.gz files) were obtained from the ENA for three human aDNA studies (Rasmussen et al. 2010, Rasmussen et al. 2011, Gamba et al. 2014). These studies were specifically selected because they provide aDNA shotgun sequence data generated from samples with a high endogenous DNA content and use a similar library creation protocol, although various polymerases (Phusion, AmpliTaq Gold/Phusion and Accuprime) and amplification schemes were used. We trimmed adapter sequences using AdapterRemoval v1.5 (Lindgreen 2012) with –trimns, –minlength 25 as settings and estimated repeat content in subsets of 1 million truncated reads as described above.

Finally, we investigate the dinucleotide repeat content in fragmented, contemporary human sequencing

data obtained from Dabney and Meyer (2012). This study investigates the effects of different polymerase buffer systems and the data consist of identical sequencing libraries that have been amplified using a variety of conditions, allowing their direct comparison.

DNA extraction historic samples

DNA was extracted from Atlantic cod scales using the same protocol as described in Star et al. (2014) in a dedicated historic DNA facility at the Natural History Museum (NHM) in Oslo. Scale samples were obtained from Canada (n = 8, year 1940) and Norway (n = 24, year 1907, see Supplementary Table 2). Briefly, scales were incubated overnight in TNES buffer (10 mM Tris, pH 7.5, 400 mM, NaCl 100 mM EDTA 0.6% SDS) with 5 mM $CaCl_2$ and 10% proteinase K at 55°C. The extracts were concentrated (Amicon-30kDA Centrifugal Filter Units) and DNA was bound to Qiaquick Nucleotide Removal Kit spin columns according to manufacturer's instructions. DNA was eluted in 50 µl of EB buffer at 37° C for 15 minutes. Concentration was determined using the Qubit dsDNA HS Assay (Life Technologies).

Library creation and amplification

Approximately 150 ng of extracted DNA was used to create Illumina compatible sequencing libraries following the protocol from Meyer and Kircher (2010) using the NEBNext® DNA Library Prep Master Mix Set for 454™ (E6070S, New England Biolabs). A custom index of six bp was designed, with a minimum distance of three bp between each single index sequence. The ligated DNA was eluted in 30 µl EB after blunt-end repair, adapter ligation and adapter fill-in.

After recommendations (Dabney and Meyer 2012), ligated DNA extracts –unless stated otherwise– were amplified for a total of 12 cycles using the following conditions: The index PCR was performed with 15 µl of ligated DNA for eight cycles (2 min at 95°C, 8 cycles of 30s at 95°C, 30s at 60°C and 70s at 72°C, final extension step of 10 min at 72°C) with PfuTurbo Cx Hotstart DNA Polymerase (Agilent Technologies, 1x buffer, 0.2 mM per dNTP, 0.2 µM P5 index primer, 0.2 µM P7, 0.4 mg/ml BSA and 2.5 units polymerase). The indexed libraries were subsequently cleaned using MinElute spin columns (Qiagen), eluted in 32 µl and divided over four tubes. These aliquots were additionally amplified for four cycles (2 min at 95°C, 4 cycles of 20s at 95°C, 20s at 60°C and 40s at 72°C, final extension step of 3 min at 72°C) with Herculase II Fusion DNA Polymerase (Agilent Technologies, 1x buffer, 0.25 mM per dNTP, 0.25 µM P5, 0.25 µM P7, DMSO 1% and 1 unit polymerase) and pooled before cleanup with Agencourt AMPure XP (Beckman Coulter).

For a subset of six Canadian specimens, we divided the ligated DNA extract to investigate the effect of an increased number of amplification cycles. For these, 15 µl of ligated DNA from the same ligation reaction was also amplified for a total of 18 cycles by increasing the second round of PCR to 10 cycles. Finally, we performed an experiment using two Canadian specimens for each of which we created three ligated DNA extracts. For each specimen, one ligated DNA extract was amplified using the conditions described above (8 + 4 cycles), one extract was amplified with the addition of an artificial, single stranded dinucleotide AC_{15} repeat oligonucleotide of 30 bp and one extract was amplified with an AG_{15} oligonucleotide (Supplementary Table 2). These oligonucleotides were added to a final concentration of 0.2 µM before index amplification with PfuTurbo Cx Hotstart DNA Polymerase for 8 cycles, after which samples were treated as above. The quality and concentration of libraries was obtained using a Bioanalyzer 2100 (Agilent Technologies) with a high sensitivity DNA kit, after which they were pooled and sequenced on a Hiseq 2000 (Illumina).

DNA extraction and library creation of contemporary Atlantic cod samples

DNA was extracted from contemporary 24 Atlantic cod specimens from Norway using a DNeasy Blood & Tissue kit (Qiagen), and sheared to an approximate insert size of 350 bp. Over 2 µg of DNA per sample was used to create Illumina compatible sequencing libraries using a TruSeq DNA PCR-Free LT Library Preparation Kit. Samples were pooled in various combinations and sequenced on a Hiseq 2000 (Illumina).

Sequencing and analysis

Using Illumina RTA & CASAVA software (versions 1.18.61.0 & 1.8.4, respectively) paired-end sequencing reads were demultiplexed and assigned to individual samples based on their index sequence, allowing zero mismatch. Forward and reverse reads were collapsed and remaining adapter sequences were removed using the program AdapterRemoval v1.5 (Lindgreen 2012) with –mm 0.33, –collapse, –trimns, –minlength 25, requiring a minimum overlap of 11 bases. Collapsed reads were used for further analyses and, for the majority of individual datasets, dinucleotide repeat estimates were obtained as described above using a subset of 1 million collapsed reads obtained with seqtk (see Supplementary Table 1). We also investigated the relationship between read length and dinucleotide repeat content by dividing reads in 10 bp bins using PRINSEQ-lite (v0.20.4) (Schmieder and Edwards 2011). Following recommendations (Schubert et al. 2012), we aligned collapsed reads to the Atlantic cod reference genome (ATLCOD1C, Star et al. 2011) using the aln algorithm of BWA v.0.7.5a-r405 (Li and Durbin 2009) with seeding disabled and -o 1 and -n 0.03. Reads that align with a minimum mapping quality score (MapQ) of 25 were considered endogenous.

Results

The Atlantic cod genome contains the highest AC and AG content compared to any other vertebrate genome of which the majority contains less than one percent of

either type of repeat (Figure 1a, Supplementary Table 1). Interestingly, the AC and AG content of contemporary Atlantic cod read data is lower than that of its assembled genome (Figure 1b). Similar discrepancies are observed in the contemporary read data of six other species whereby one or both types of dinucleotide repeat are underrepresented in the majority of individual libraries (Supplementary Figure 1). Nonetheless, the AC and AG repeat content in Atlantic cod read data from historic sources is substantially higher and depends on sample location (Figure 1b). The Canadian samples in particular have an inflated proportion of AC and AG dinucleotide repeats –between 20% to 35% of all nucleotides can be classified as one of these types of repeats– while the Norwegian samples contain more moderate levels. For an identical number of PCR cycles (12 cycles), Norwegian samples contain more endogenous DNA (58%) than Canadian samples (42%, Supplementary Table 2). The high repeat content in Canadian samples is associated with a substantially lower average insert length, indicating that DNA is more fragmented in these samples compared to those from Norway. Moreover, the elevated levels of AC and AG repeats in the historic sequencing data are inversely related to read length with shorter reads containing more repeats (Supplementary Figure 2). In contrast, aDNA sequence data generated from human samples –with levels of endogenous DNA comparable to those of the Atlantic cod samples used here– do not have inflated proportions of AC or AG repeats (Supplementary Figure 3). While a portion of libraries from the Rasmussen et al. 2010 study contain more AC and AG repeats than the human reference genome and one outlier for AC repeats can be observed, these values are in a range that does not meaningfully affects the efficiency of sequencing.

Using 18 instead of 12 PCR cycles consistently increases the number of AC and AG repeats in Canadian Atlantic cod aDNA sequencing data, despite using the same ligated extract for each amplification condition (Figure 2). This increase in repeats coincides with a substantial lowered number of reads considered endogenous (i.e. those that can be reliably aligned to the Atlantic cod genome with MapQ values of over 25, Supplementary Table 2). For the libraries that were amplified for 18 cycles, an overestimation of fragment length and multimodal distribution of Bioanalyzer plots is indicative for the presence of heteroduplex constructs (see Supplementary Figure 4 for a typical example). The presence of these constructs is suggestive of reaching the PCR plateau phase and the depletion of Illumina compatible sequencing primers during amplification.

The addition of AC or AG single stranded repetitive oligonucleotides during amplification influences the proportion of dinucleotides in Atlantic cod aDNA sequencing data (Figure 3). Counterintuitively, *adding* AC or AG repeats *suppresses* the respective proportion of these repeats in the sequencing data. Moreover, the addition of AC or AG oligonucleotides respectively generates a different response regarding the proportion of endogenous reads and of the type of repeats that were *not* manipulated. When adding AC oligonucleotides, the reduction in AC repeats coincides with a relative increase in AG repeats (Figure 3a). This effect is absent when adding AG oligonucleotides (Figure 3b). It is unclear why this difference occurs, yet this relative increase in AG repeats –when adding AC oligonucleotides– likely negates any increase in the proportion of endogenous reads. In contrast, the proportion of endogenous reads does increase when adding AG oligonucleotides. Unfortunately, not enough DNA extract remains from the same specimens to perform an experiment whereby both repeats are added simultaneously, but we expect that both types would be suppressed, leading to even higher proportions of endogenous DNA.

An interaction between different polymerase-buffer systems and extended PCR cycling influences the proportion of dinucleotide repeats in the fragmented contemporary human data from Dabney and Meyer (2012) (Figure 4). Interestingly, Phusion polymerase-buffer systems show a minor but consistent bias, with AC repeats *more* and AG repeats *less* abundant with an increase in PCR cycle number. The reason for this divergent proliferation in AC and AG content during amplification cannot directly be explained by any known GC biases, because the GC content of both types of repeat is identical. Nevertheless, this divergent proliferation per repeat type does suggest that this polymerase bias is different from the one observed in the Atlantic cod data.

Discussion and conclusion

Here we report that the genome of Atlantic cod contains an unusually high abundance of AC and AG dinucleotide repeats and that sequencing data from historic samples are consistently biased towards these repeats in this species. This high level of repeats in these aDNA data occurs despite an apparent bias against such dinucleotide content in contemporary Illumina Hiseq sequencing reads. The reason for this bias in contemporary sequencing data remains unclear, although similar underrepresentation of dinucleotide repeats can also be observed in sequencing data from other vertebrates.

The high abundance of dinucleotide repeats in the sequencing data from historic samples of Atlantic cod appears to be introduced during the PCR amplification step of the library creation protocol. For instance, it is clear that an increased number of PCR cycles leads to a substantially higher repeat content in aDNA sequence data for a given ligated extract. Furthermore, the proportion of AC and AG repeats can be reduced through the addition of each respective oligonucleotide during library amplification. Thus, single stranded AC and AG oligonucleotides interact with the endogenous repetitive DNA present in the ligated

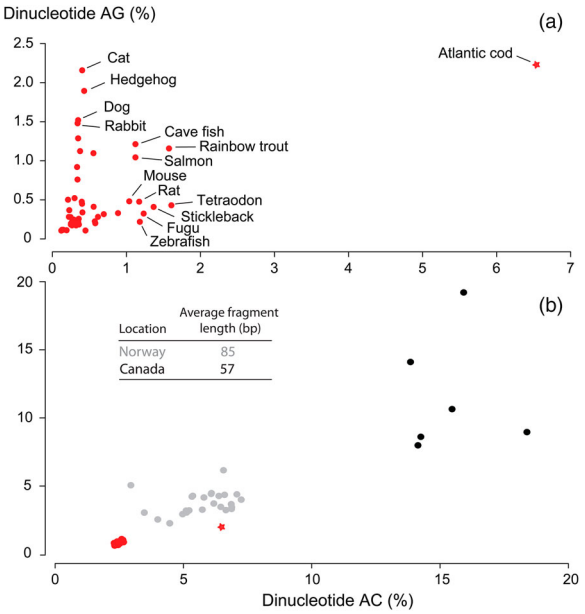

Figure 1 Dinucleotide repeat content in vertebrate assemblies and Atlantic cod sequencing data. (a) The amount of AC and AG dinucleotide repeats was obtained using Tandem Repeat Finder (TRF, version 407b) in 65 vertebrate genome assemblies (Ensembl, release 76), including salmon and rainbow trout. Estimates were divided by total assembly size, excluding undetermined bases. Assemblies with more extreme repeat content are indicated (see Supplementary Table 2 for a detailed species list). (b) Dinucleotide repeat estimates in Atlantic cod sequencing data from contemporary (red), historic Norwegian (grey) and historic Canadian (black) samples. Each dot represents a single library, generated from a different individual. The repeat content in the Atlantic cod assembly (red, star) is indicated for comparison. The average fragment length (bp) is shown for the historic Norwegian and Canadian samples. Notice the different scales on the x- and y- axes for each panel.

library. Based on the above observations, amplification rather than preferential preservation or ligation appear to be the main process responsible for the observed high abundance. While we find some evidence for polymerase-specific dinucleotide repeat bias in the fragmented human data from Dabney and Meyer (2012), these patterns do not fit those we see in the Atlantic cod data. Instead, we propose two hypotheses, a *fragmentation-length bias* hypothesis and a *self-priming* hypothesis that may lead to a proliferation of repeats during library amplification.

Fragmentation-length bias

It is possible that AC and AG dinucleotide repeats in Atlantic cod DNA are particularly prone to fragmentation during post mortem degradation relative to non-repetitive DNA, for instance by differential nucleosome packaging (Pedersen et al. 2014). This would lead to a DNA fragment composition whereby shorter fragments are more likely to consist of repetitive DNA for a given sequence context. Since some polymerases preferentially amplify shorter length fragments (Dabney and Meyer 2012), the amplification of a pool of DNA in which shorter fragments are associated with repeats could lead to an increase in repetitive DNA with increasing PCR cycles. While the observed association of shorter fragments with more repetitive DNA (Supplementary Figure 2) indeed agrees with such a mode of proliferation (but see also below), some consideration may question the efficiency of this process. We used polymerases (PfuTurbo Cx Hotstart and Herculase II Fusion) that should be among the least affected by amplification length bias for the rather moderate number of PCR cycles used here (Dabney and Meyer 2012). The observed repeat abundance appears therefore somewhat extreme to be solely due to fragment length bias during amplification.

Self-priming

Given the high endogenous AC and AG content in the Atlantic cod genome, we suggest that repeats may proliferate through a mechanism whereby repetitive DNA anneals to itself, i.e. self-priming. Self-priming explains typical PCR fragmentation patterns observed when using transcript-activator like effector (TALe) technology (Hommelsheim et al. 2014), which highlights the propensity of repetitive DNA to self-prime in a variety of protocols and conditions. The high repeat content in the Atlantic cod genome would lead to an ample supply of short repetitive fragments with high affinity for themselves after degradation. Self-priming could proceed according to the following model: First, single-stranded dinucleotide DNA anneals simultaneously with an amplification primer (Figure 5a). Elongation can proceed from the repetitive fragment and the primer. This will lead to one or two fragments, because the annealed repetitive region blocks elongation from the primer, given that the first polymerase used in these experiments (PfuTurbo Cx Hotstart) does not have strand-displacement capabilities. Since two fragments may be generated in a single cycle, this type of amplification would be particularly competitive relative to a normal amplification reaction, which only generates one fragment per annealed primer. In the presence of repetitive DNA and sequencing primers, these fragments can be amplified, leading to a population of repeat-associated primers with a specific affinity for repetitive DNA (Figure 5b). Finally, repetitive fragments associated with the Illumina P5 or P7 sequencing primer either hybridize with each other, or directly bind to fragments with the P5 and P7 primer pair, leading to constructs that can be sequenced (Figure 5c). The potential for these processes to occur will increase with the number of PCR cycles, through a continuous accumulation of repeat-associated primer pairs and depletion of the Illumina sequencing primers.

We expect that self-priming can also lead to association of shorter fragments with more repetitive DNA.

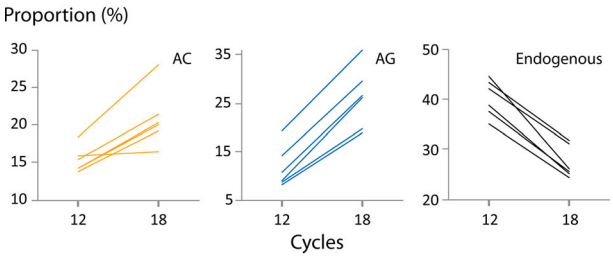

Figure 2 Effect of amplification conditions on repeat proliferation and endogenous DNA content. The proportion of AC (yellow), AG (blue) dinucleotide repeats was obtained using TRF. Reads aligning with a MapQ value of 25 or higher toward the Atlantic cod assembly are considered endogenous (black). A total of six ligated extracts from Canadian specimens were amplified either 12 or 18 PCR cycles, yielding 12 separate sequencing libraries (see text for details). Lines connect libraries that were created using the same extract. Notice the different scale on the y-axes for each panel. For libraries amplified using 18 cycles, an overestimation of fragment length and a multimodal distribution of BioAnalyzer plots indicated the presence of heteroduplexes, which is suggestive of reaching the plateau phase of PCR during library amplification.

First, the length of dinucleotide stretches in vertebrate genomes decays exponentially and higher fragmentation would lead to more fragments that consist entirely of repeats, which are likely to be more prone to self-priming than more complex, longer fragments. Moreover, shorter DNA fragments denature earlier for a given temperature to form single-stranded DNA (ssDNA, Pääbo et al. 2004). Since the ligation protocol targets dsDNA, any ssDNA present in the extract remains unligated in solution, providing a more abundant pool of shorter, rather than longer fragments available for self-priming. Hybridization of such shorter, repetitive fragments associated with the Illumina P5 or P7 sequencing primers would lead to lower average fragment length and its association with repetitive DNA.

While the fragment-length bias hypothesis may be less likely due to the choice of polymerases used here, we cannot exclude this process from contributing to the observed repeat proliferation altogether. An initial lower average fragment length before amplification may promote both types of repeat proliferation, and agrees with the higher repeat abundance observed in the Canadian samples. Regardless of the proliferation mechanism, given that AC and AG dinucleotide repeats are endogenously abundant in Atlantic cod DNA, their biased amplification will be difficult to prevent completely. Nevertheless, our results show that the abundance of repeats can be somewhat negated by using as little PCR cycles as possible. We would further recommend using polymerases with low length-bias during amplification. Finally, through the addition of repeat-specific oligonucleotides, the amount of dinucleotide repeats can be reduced, potentially increasing the proportion of non-repetitive endogenous DNA. These oligonucleotides are inexpensive and any gain in sequencing efficiency is rapidly economical.

Are other species affected?

Our finding represents an unexpected type of bias that is generated during library preparation of Atlantic cod aDNA samples. It is unclear however, to what extent other species are affected by this bias. In case of dinucleotide repeat proliferation, Atlantic cod has an unusual genomic composition, hence it is unlikely that other species have similar issues with regard to these types of repeats. Furthermore, we note that the samples used here are relatively recent, and – even though degradation processes have resulted in a DNA fragmentation comparable to substantially older samples (e.g. Rasmussen et al. 2010)– these

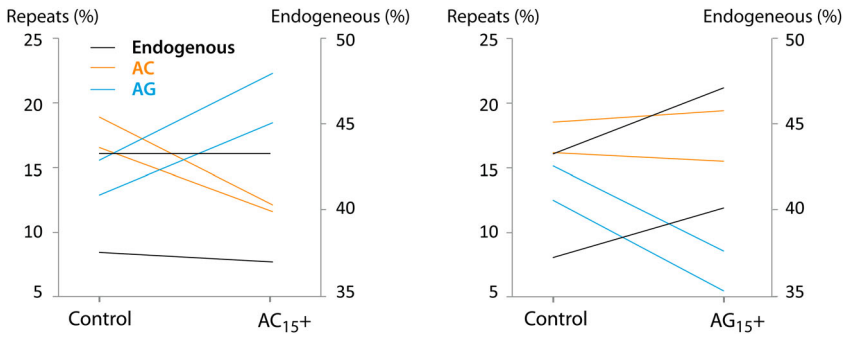

Figure 3 Effect of manufactured dinucleotides on repeat proliferation and endogenous DNA content. The proportion of AC (yellow) and AG (blue) dinucleotide repeats (both using left y-axis for scale) was obtained using TRF. Reads aligning with a MapQ value of 25 or higher toward the Atlantic cod assembly are considered endogenous (black, right y-axis). A total of 6 libraries were generated from two Canadian specimens (three ligated DNA extracts per specimen). Per individual, one extract was amplified using the standard protocol (C, data used in both panels), one extract was amplified including an AC repeat oligonucleotide of 30 base pair (AC_{15}+) and one extract was amplified including an AG repeat oligonucleotide (AG_{15}+, see text for details). Lines connect libraries created from the same specimen.

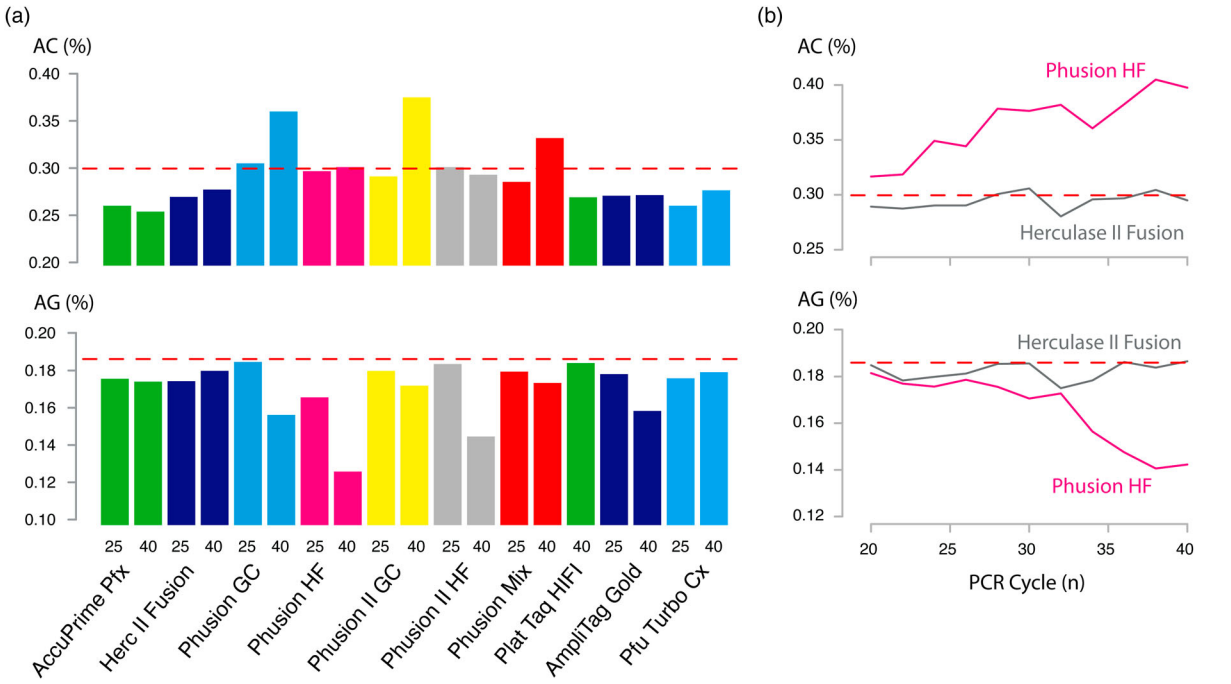

Figure 4 Dinucleotide repeat (AC and AG) bias in contemporary human sequencing data. (a) AC and AG repeat content after amplification with ten different polymerase-buffer systems compared to the content of the original, unamplified library (dashed red line). The number on the x-axis shows the number of PCR cycles. We only used sequencing data with a minimum of 100,000 reads per combination of polymerase type and cycle number. (b) A higher resolution of the AC and AG content of a sequencing library amplified with Herculase II Fusion (grey) and Phusion HF (purple) compared to the original, unamplified library (dashed red line). PCR reactions were performed at 2-cycle intervals from 20 to 40 cycles. The data in panel (a) and (b) were previously analyzed for GC and length bias in Dabney and Meyer (2012). Notice the different scale on the y-axes for each panel.

samples nonetheless contain large proportions of endogenous Atlantic cod DNA. These high levels of endogenous DNA in combination with its abundance of simple repeats create particularly favorable conditions for self-priming.

So far, most studies investigating aDNA shotgun data in vertebrates focus on human or horse, at least those that make their raw sequence data publicly available. These species typically have an order of magnitude lower AC and AG content compared to Atlantic cod and consequently have little potential for either fragment-length bias or self-priming of these repeats. It is therefore not surprising that we find little evidence for a proliferation of dinucleotide repeats in human aDNA data (Rasmussen et al. 2010, Rasmussen et al. 2011, Gamba et al. 2014). Nevertheless, there is a research effort focusing on dog aDNA samples, although so far only mitochondrial DNA has been used (e.g. Thalmann et al. 2013, Witt et al. 2014). Given that the dog genome contains a relatively high proportion of AG repeats, it will be interesting to see whether the same bias can be observed in aDNA shotgun libraries generated from ancient dog samples with high endogenous DNA content.

It is uncertain whether other types of repetitive DNA, such as Short Interspersed Elements (SINEs) or Long Interspersed Elements (LINEs) have the propensity to be amplified due to fragment-length bias or self-priming, given a high enough representation in a genome. Interestingly, human aDNA sequence data is enriched for SINEs relative to LINEs, and SINES are the more simple and most abundant type of repeat in the human genome. SINEs are also GC-rich however, and it has been hypothesized that aDNA data is biased towards such regions, either due to a denaturation of AT rich regions during library preparation (Green et al. 2008, Briggs et al. 2009, Meyer and Kircher 2010) or due to polymerase bias during amplification (Aird et al. 2011, Dabney and Meyer 2012). We note that the processes underlying amplification bias discussed here could also contribute towards such GC enrichment if fragments containing SINE sequences proliferate, and that these processes (fragment-length bias, self-priming, temperature dependent denaturation or polymerase bias) may act in concert.

Moreover, it can be difficult to detect the effects of biased amplification of more complex repeat regions if only a sole sequencing library –generated using a single amplification scheme– is analyzed and this phenomenon is not specifically investigated. First, the fragment-length bias and self-priming hypotheses proposed here do not necessarily lead to an increase in clonal reads, hence may be missed by altogether by algorithms detecting levels of clonality. Second, the sequence complexity of repetitive DNA rapidly inflates with an increase in the length of their periodicity, which reduces the

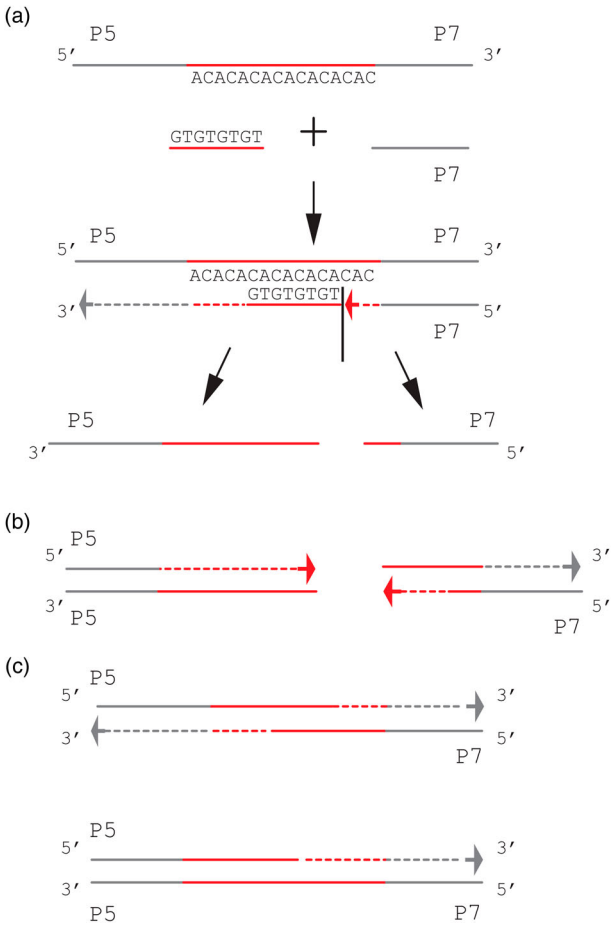

Figure 5 Hypothetical mechanism of self-priming of dinucleotide repeats during the amplification of whole genome sequencing libraries. (a) Single stranded repetitive DNA (red) anneals simultaneously with the library specific amplification primer (grey). Elongation from the primer is blocked by repetitive DNA, which is downstream annealed, leading to the formation of two fragments. (b) Annealing and elongation from amplification primer or repetitive DNA generates a population of repeat-associated primers with specific affinity for repetitive DNA. (c) Annealing and hybridization of complementary repetitive regions generates fragments with P5 and P7 primer pairs suitable for sequencing. Alternatively, repeat-associated primers with affinity for repetitive regions specifically amplify those regions. The size of primer and repeat regions is not shown to scale.

propensity of DNA to anneal to complementary regions during PCR. Self-priming may therefore not be particularly efficient for more complex repeats, resulting in a less discernable pattern than observed here in Atlantic cod. Given that the proposed models are amplification driven however, their effects can be quantified using different amplification conditions for the same ligation reaction and by subsequently assessing variation in repeat content. Such experiment would be of importance for reported enrichments of repeats in aDNA WGS data, like that of the endogenously abundant *Gypsy* retrotransposon in cotton (Palmer et al. 2012, Shapiro and Hofreiter 2014), confirming the evolutionary significance of such observations.

Microsatellite evolution and population genetic inference

Finally, the observed high endogenous AC and AG content in the Atlantic cod genome may have implications for our understanding of microsatellite evolution and their use in genetic studies. For instance, comparative knowledge of microsatellite distribution is essential to develop an understanding of their evolutionary properties (Bhargava and Fuentes 2010) and their unusual abundance in Atlantic cod is a clear indication that these properties are not fully understood for all taxa. One well-documented property however, is that microsatellites of longer length experience increased mutation rates leading to high allelic variation (Wierdl, Dominska, and Petes 1997, Kruglyak et al. 1998, Whittaker et al. 2003, Ellegren 2004). Interestingly, the high level of allelic variation that is typical of marine fish populations was suggested to reflect their relatively large (historic) population sizes rather than intrinsic mutation rates (DeWoody and Avise 2000). We highlight the possibility that, given the distinct microsatellite properties in fish (Neff and Gross 2001), and Atlantic cod specifically (Jiang et al. 2014, this study), mutation models suitable for other vertebrates may not apply (Buschiazzo and Gemmell 2006, Grover and Sharma 2011), affecting population genetic inference for these taxa (Jakobsson, Edge, and Rosenberg 2013, Putman and Carbone 2014).

Overall, we observe that the genome of Atlantic cod contains an unusual level of simple AC and AG dinucleotide repeats and that the abundance of these repeats is highly inflated WGS data generated from historical samples for this species. Our results indicate that it is amplification rather than preferential preservation or ligation that is responsible for the observed high abundance that leads to artificially high levels of AC and AG repeats. While the extent of similar repeat proliferation bias in other studies is unclear, we nonetheless urge caution when quantifying repeat content in aDNA WGS data, given that it may be challenging to detect the effects of repeat amplification if such processes affects more complex repeat structures than dinucleotide repeats.

Availability of supporting data

The data set(s) supporting the results of this article are available at the NorStore public data archive (https://archive.norstore.no) with DOI:10.11582/2015.00013.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Dr. Jesse Dabney and Dr. Matthias Meyer for providing access to their sequencing data. Sequencing was performed by the Norwegian Sequencing Centre, a national technology platform hosted by the University of Oslo (UIO) and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities (www.sequencing.uio.no).

Computational intensive analyses were done on the Abel Cluster, owned by the UIO and the Norwegian metacenter for High Performance Computing (NOTUR), and operated by the Department for Research Computing at USIT, the UIO IT-department (http://www.hpc.uio.no/). We thank Dr. Thomas H.A. Haverkamp, Dr. Alexander J. Nederbragt and Dr. Sanne Boessenkool for comments and suggestions. This research was supported by the Norwegian Research Council under projects "Fisheries induced evolution in Atlantic cod investigated by ancient and historic samples (#203850/E40)" and "The Aqua Genome Project (#221734/O30)". We have adhered to all local, national and international regulations and conventions, and we respected normal scientific ethical practices.

References

Aird, Daniel, Michael G Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B Jaffe, Chad Nusbaum, and Andreas Gnirke. 2011. "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries." *Genome biol* 12 (2):R18

Amos, William, and Andrew Clarke. 2008. "Body temperature predicts maximum microsatellite length in mammals." *Biology Letters* 4 (4):399–401

Berthelot, Camille, Frédéric Brunet, Domitille Chalopin, Amélie Juanchich, Maria Bernard, Benjamin Noël, Pascal Bento, Corinne Da Silva, Karine Labadie, Adriana Alberti, Jean-Marc Aury, Alexandra Louis, Patrice Dehais, Philippe Bardou, Jérôme Montfort, Christophe Klopp, Cédric Cabau, Christine Gaspin, Gary H. Thorgaard, Mekki Boussaha, Edwige Quillet, René Guyomard, Delphine Galiana, Julien Bobe, Jean-Nicolas Volff, Carine Genêt, Patrick Wincker, Olivier Jaillon, Hugues Roest Crollius, and Yann Guiguen. 2014. "The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates." *Nat Commun* 5

Bhargava, Atul, and F. F. Fuentes. 2010. "Mutational Dynamics of Microsatellites." *Molecular Biotechnology* 44 (3):250–266

Briggs, Adrian W., Jeffrey M. Good, Richard E. Green, Johannes Krause, Tomislav Maricic, Udo Stenzel, Carles Lalueza-Fox, Pavao Rudan, Dejana Brajkovic, Zeljko Kucan, Ivan Gusic, Ralf Schmitz, Vladimir B. Doronichev, Liubov V. Golovanova, Marco de la Rasilla, Javier Fortea, Antonio Rosas, and Svante Pääbo. 2009. "Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes." *Science* 325 (5938):318–321

Briggs, AW, U Stenzel, PLF Johnson, RE Green, J Kelso, K Prüfer, M Meyer, J Krause, MT Ronan, M Lachmann, and S Pääbo. 2007. "Patterns of damage in genomic DNA sequences from a Neandertal." *Proc Natl Acad Sci USA* 104:14616–14621

Buschiazzo, Emmanuel, and Neil J. Gemmell. 2006. "The rise, fall and renaissance of microsatellites in eukaryotic genomes." *BioEssays* 28 (10):1040–1050

Chambers, Geoffrey K., and Elizabeth S. MacAvoy. 2000. "Microsatellites: consensus and controversy." *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 126 (4):455–476

Dabney, Jesse, and Matthias 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 (2):87–94

DeWoody, J. A., and J. C. Avise. 2000. "Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals." *Journal of Fish Biology* 56 (3):461–473

Ellegren, Hans. 2004. "Microsatellites: simple sequences with complex evolution." *Nature Reviews Genetics* 5 (6):435–445

Gamba, Cristina, Eppie R. Jones, Matthew D. Teasdale, Russell L. McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, Alasdair Whittle, János Dani, Pál Raczky, Thomas F. G. Higham, Michael Hofreiter, Daniel G. Bradley, and Ron Pinhasi. 2014. "Genome flux and stasis in a five millennium transect of European prehistory." *Nat Commun* 5

Ginolhac, Aurelien, Morten Rasmussen, M. Thomas P. Gilbert, Eske Willerslev, and Ludovic Orlando. 2011. "mapDamage: testing for damage patterns in ancient DNA sequences." *Bioinformatics* 27 (15):2153–2155

Green, RE, AW Briggs, J Krause, K Prüfer, HA Burbano, M Siebauer, M Lachmann, and S Pääbo. 2009. "The Neandertal genome and ancient DNA authenticity." *EMBO J* 28:2494–2502

Green, Richard E., Anna-Sapfo Malaspinas, Johannes Krause, Adrian W. Briggs, Philip L. F. Johnson, Caroline Uhler, Matthias Meyer, Jeffrey M. Good, Tomislav Maricic, Udo Stenzel, Kay Prüfer, Michael Siebauer, Hernán A. Burbano, Michael Ronan, Jonathan M. Rothberg, Michael Egholm, Pavao Rudan, Dejana Brajković, Željko Kućan, Ivan Gusić, Mörten Wikström, Liisa Laakkonen, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. 2008. "A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing." *Cell* 134 (3):416–426

Grover, Atul, and PC Sharma. 2011. "Is spatial occurrence of microsatellites in the genome a determinant of their function and dynamics contributing to genome evolution." *Curr Sci* 100:859–869

Gymrek, Melissa, David Golan, Saharon Rosset, and Yaniv Erlich. 2012. "lobSTR: A short tandem repeat profiler for personal genomes." *Genome Research* 22 (6):1154–1162

Hommelsheim, Carl Maximilian, Lamprinos Frantzeskakis, Mengmeng Huang, and Bekir Ülker. 2014. "PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications." *Scientific reports* 4

Jakobsson, Mattias, Michael D Edge, and Noah A Rosenberg. 2013. "The relationship between FST and the frequency of the most frequent allele." *Genetics* 193 (2):515–528

Jiang, Q., Q. Li, H. Yu, and L. Kong. 2014. "Genome-wide analysis of simple sequence repeats in marine animals-a comparative approach." *Mar Biotechnol (NY)* 16 (5):604–19

Kofler, Robert, Christian Schlötterer, Evita Luschüzky, and Tamas Lelley. 2008. "Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites." *BMC Genomics* 9:612–612

Krause, Johannes, Qiaomei Fu, Jeffrey M. Good, Bence Viola, Michael V. Shunkov, Anatoli P. Derevianko, and Svante Pääbo. 2010. "The complete mitochondrial DNA genome of an unknown hominin from southern Siberia." *Nature* 464 (7290):894–897

Kruglyak, Semyon, Richard T. Durrett, Malcolm D. Schug, and Charles F. Aquadro. 1998. "Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations." *Proceedings of the National Academy of Sciences of the United States of America* 95 (18):10774–10778

Li, Heng, and Richard Durbin. 2009. "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25 (14):1754–1760

Lindgreen, Stinus. 2012. "AdapterRemoval: easy cleaning of next-generation sequencing reads." *BMC Research Notes* 5 (1):337

Merkel, Angelika, and Neil Gemmell. 2008. "Detecting short tandem repeats from genome data: opening the software black box." *Briefings in Bioinformatics* 9 (5):355–366

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

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F. Hammer, Michael V. Shunkov, Anatoli P. Derevianko, Nick Patterson, Aida M. Andrés, Evan E. Eichler, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. 2012. "A High-Coverage Genome

Sequence from an Archaic Denisovan Individual." *Science* 338 (6104):222–226

Neff, Bryan D., and Mart R. Gross. 2001. "Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats." *Evolution* 55 (9):1717–1733

Overballe-Petersen, S., L. Orlando, and E. Willerslev. 2012. "Next-generation sequencing offers new insights into DNA degradation." *Trends in Biotechnology* 30 (7):364–368

Pääbo, Svante, Hendrik Poinar, David Serre, Viviane Jaenicke-Després, Juliane Hebler, Nadin Rohland, Melanie Kuch, Johannes Krause, Linda Vigilant, and Michael Hofreiter. 2004. "Genetic analyses from ancient DNA." *Annual Review of Genetics* 38:645–679

Palmer, Sarah A., Alan J. Clapham, Pamela Rose, Fábio O. Freitas, Bruce D. Owen, David Beresford-Jones, Jonathan D. Moore, James L. Kitchen, and Robin G. Allaby. 2012. "Archaeogenomic Evidence of Punctuated Genome Evolution in Gossypium." *Molecular Biology and Evolution* 29 (8):2031–2038

Pedersen, Jakob Skou, Eivind Valen, Amhed M Vargas Velazquez, Brian J Parker, Morten Rasmussen, Stinus Lindgreen, Berit Lilje, Desmond J Tobin, Theresa K Kelly, and Søren Vang. 2014. "Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome." *Genome research* 24 (3):454–466

Putman, Alexander I., and Ignazio Carbone. 2014. "Challenges in analysis and interpretation of microsatellite data for population genetic studies." *Ecology and Evolution* 4 (22):4399–4428

Rasmussen, M, X Guo, Y Wang, KE Lohmueller, S Rasmussen, A Albrechtsen, L Skotte, S Lindgreen, M Metspalu, T Jombart, T Kivisild, W Zhai, A Eriksson, A Manica, L Orlando, FMDL Vega, S Tridico, E Metspalu, K Nielsen, MC Avila-Arcos, JV Moreno-Mayar, C Muller, J Dortch, MTP Gilbert, O Lund, A Wesolowska, M Karmin, LA Weinert, B Wang, J Li, S Tai, F Xiao, T Hanihara, G van Driem, AR Jha, F-X Ricaut, P de Knijff, AB Migliano, IG Romero, K Kristiansen, DM Lambert, S Brunak, P Forster, B Brinkmann, O Nehlich, M Bunce, M Richards, R Gupta, CD Bustamante, A Krogh, RA Foley, MM Lahr, F Balloux, T Sicheritz-Ponten, R Villems, R Nielsen, J Wang, and E Willerslev. 2011. "An Aboriginal Australian genome reveals separate human dispersals into Asia." *Science* 334:94–98

Rasmussen, M., Y. R. Li, S. Lindgreen, J. S. Pedersen, A. Albrechtsen, I. Moltke, M. Metspalu, E. Metspalu, T. Kivisild, R. Gupta, M. Bertalan, K. Nielsen, M. T. P. Gilbert, Y. Wang, M. Raghavan, P. F. Campos, H. M. Kamp, A. S. Wilson, A. Gledhill, S. Tridico, M. Bunce, E. D. Lorenzen, J. Binladen, X. S. Guo, J. Zhao, X. Q. Zhang, H. Zhang, Z. Li, M. F. Chen, L. Orlando, K. Kristiansen, M. Bak, N. Tommerup, C. Bendixen, T. L. Pierre, B. Gronnow, M. Meldgaard, C. Andreasen, S. A. Fedorova, L. P. Osipova, T. F. G. Higham, C. B. Ramsey, T. V. O. Hansen, F. C. Nielsen, M. H. Crawford, S. Brunak, T. Sicheritz-Ponten, R. Villems, R. Nielsen, A. Krogh, J. Wang, and E. Willerslev. 2010. "Ancient human genome sequence of an extinct Palaeo-Eskimo." *Nature* 463 (7282):757–762

Schaper, Elke, Andrey V. Kajava, Alain Hauser, and Maria Anisimova. 2012. "Repeat or not repeat? - Statistical validation of tandem repeat prediction in genomic sequences." *Nucleic Acids Research* 40 (20):10005–10017

Schmieder, Robert, and Robert Edwards. 2011. "Quality control and preprocessing of metagenomic datasets." *Bioinformatics* 27 (6):863–864

Schubert, Mikkel, Aurelien Ginolhac, Stinus Lindgreen, John Thompson, Khaled AL-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. 2012. "Improving ancient DNA read mapping against modern reference genomes." *BMC Genomics* 13 (1):178

Seguin-Orlando, Andaine, Mikkel Schubert, Joel Clary, Julia Stagegaard, Maria T. Alberdi, José Luis Prado, Alfredo Prieto, Eske Willerslev, and Ludovic Orlando. 2013. "Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes." *Plos One* 8 (10):e78575

Shapiro, B, and M Hofreiter. 2014. "A paleogenomic perspective on evolution and gene function: new insights from ancient DNA." *Science* 343 (6169):1236573

Star, Bastiaan, Alexander J. Nederbragt, Marianne H. S. Hansen, Morten Skage, Gregor D. Gilfillan, Ian R. Bradbury, Christophe Pampoulie, Nils Chr Stenseth, Kjetill S. Jakobsen, and Sissel Jentoft. 2014. "Palindromic Sequence Artifacts Generated during Next Generation Sequencing Library Preparation from Historic and Ancient DNA." *Plos One* 9 (3):e89676

Star, Bastiaan, Alexander J. Nederbragt, Sissel Jentoft, Unni Grimholt, Martin Malmstrom, Tone F. Gregers, Trine B. Rounge, Jonas Paulsen, Monica H. Solbakken, Animesh Sharma, Ola F. Wetten, Anders Lanzen, Roger Winer, James Knight, Jan-Hinnerk Vogel, Bronwen Aken, Oivind Andersen, Karin Lagesen, Ave Tooming-Klunderud, Rolf B. Edvardsen, Kirubakaran G. Tina, Mari Espelund, Chirag Nepal, Christopher Previti, Bard Ove Karlsen, Truls Moum, Morten Skage, Paul R. Berg, Tor Gjoen, Heiner Kuhl, Jim Thorsen, Ketil Malde, Richard Reinhardt, Lei Du, Steinar D. Johansen, Steve Searle, Sigbjorn Lien, Frank Nilsen, Inge Jonassen, Stig W. Omholt, Nils Chr Stenseth, and Kjetill S. Jakobsen. 2011. "The genome sequence of Atlantic cod reveals a unique immune system." *Nature* 477 (7363):207–210

Tautz, Diethard. 1989. "Hypervariabflity of simple sequences as a general source for polymorphic DNA markers." *Nucleic Acids Research* 17 (16):6463–6471

Tautz, Diethard, and Manfred Renz. 1984. "Simple sequences are ubiquitous repetitive components of eukaryotic genomes." *Nucleic Acids Research* 12 (10):4127–4138

Thalmann, O., B. Shapiro, P. Cui, V. J. Schuenemann, S. K. Sawyer, D. L. Greenfield, M. B. Germonpré, M. V. Sablin, F. López-Giráldez, X. Domingo-Roura, H. Napierala, H-P. Uerpmann, D. M. Loponte, A. A. Acosta, L. Giemsch, R. W. Schmitz, B. Worthington, J. E. Buikstra, A. Druzhkova, A. S. Graphodatsky, N. D. Ovodov, N. Wahlberg, A. H. Freedman, R. M. Schweizer, K.-P. Koepfli, J. A. Leonard, M. Meyer, J. Krause, S. Pääbo, R. E. Green, and R. K. Wayne. 2013. "Complete mitochondrial genomes of ancient canids suggest a european origin of domestic dogs." *Science* 342 (6160):871–874

Whittaker, John C., Roger M. Harbord, Nicola Boxall, Ian Mackay, Gary Dawson, and Richard M. Sibly. 2003. "Likelihood-Based Estimation of Microsatellite Mutation Rates." *Genetics* 164 (2):781–787

Wierdl, Monika, Margaret Dominska, and Thomas D. Petes. 1997. "Microsatellite Instability in Yeast: Dependence on the Length of the Microsatellite." *Genetics* 146 (3):769–779

Witt, Kelsey E., Kathleen Judd, Andrew Kitchen, Colin Grier, Timothy A. Kohler, Scott G. Ortman, Brian M. Kemp, and Ripan S. Malhi. 2014. "DNA analysis of ancient dogs of the Americas: Identifying possible founding haplotypes and reconstructing population histories." *Journal of Human Evolution* (79):105–118