Investigation and Application of Methods for Ancient DNA Research

Stephen Malone Richards

A dissertation submitted to the University of Adelaide
in accordance with the requirements of
the degree of PhD
in the Faculty of Science,
School of Earth & Environmental Sciences

September 2014

Dedication

This thesis is dedicated to Ignatius J. Reilly.

Table of Contents

List of Figures

Chapter VI: Elucidating Bovid Evolution with Genotyping Technologies

List of Tables

List of Appendixes

Abstract

The introduction of high throughput sequencing (HTS) in 2005 caused a revolution in the field of ancient DNA (aDNA). Using the large sequencing capacity of HTS, researchers have overcome the abundant environmental contamination present in most aDNA extractions to reconstruct the genomes of long extinct organisms, such as an archaic horse that perished >500,000 years ago. The proliferation of genomes engendered by HTS has also led to the development of potential ancillary technologies for aDNA research such as genotyping microarrays. In this thesis, HTS and genotyping techniques were developed or refined to improve the application of aDNA to larger biological questions in evolution. This thesis successfully: *a) describes an in-house hybridization capture system that uses RNA probes generated from long-range PCR amplicons, b) demonstrates that recombinase polymerase amplification is a less biased alternative to PCR in hybridization capture of aDNA, c) develops an analytical approach that improves phylogenies generated with data from the Illumina BovineSNP50 BeadChip (a commercially available genotyping microarray).* In contrast, an attempt to determine the identity of modified nucleotides in aDNA with Pacific Bioscience's Single Molecule Real-Time (SMRT) sequencing prove to be unsuccessful and genotyping of ancient bison aDNA with the BovineSNP50 BeadChip generated inconsistent results. Furthermore, a hybridization capture probe design was tested and found to be unsuitable for aDNA enrichment. For the larger biological aspect of this thesis, several of the methods developed were used to study bison, because these animals are ideal models of megafauna evolution. Using the in-house hybridization capture system, whole mitochondrial genomes were enriched from aDNA and used to help identify a new extinct species of bison. Furthermore, the new analytical approach for BovineSNP50 BeadChip data was used to demonstrate a

significant genetic split between American woods and plains bison, which supports separating these animals at least at the subspecies level. This genetic split suggests that woods and plains bison should be conserved as separate species, which has considerable economic and political implications.

Thesis Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Stephen M. Richards

September 13, 2014

Introduction

1.0 History of Ancient DNA

Ancient DNA (aDNA) is loosely defined as DNA isolated from biological material that has not been specifically treated for later recovery of nucleic acids. The first study of aDNA was published in 1984 and examined genetic material extracted from muscle tissue of a museum specimen of a quagga, an extinct member of the horse family [1]. Since then, aDNA has been extracted and sequenced from a number of sources including bone [2], teeth [3], hair [4], egg shell [5], dental calculus [6], sediment [7], ice cores [8], and seeds [9]. aDNA is a critical resource for many disciplines of research as the material contains a real-time record of past genetic diversity [10], which can be used to address numerous questions in the evolution of a particular species or family. In biology, aDNA is commonly used to study the evolution of extinct species [11], whilst in anthropology aDNA is employed to track the migration and demographics of past human populations [12] and identify locations of agricultural crop domestication [13].

From the initial experiments on aDNA it became apparent that contamination would be a serious complication for most samples [14,15]. Nearly all aDNA is highly contaminated with exogenous DNA and endogenous molecules may only represent a small fraction of the total. Most of the exogenous contamination found in aDNA is bacterial and fungal molecules from environmental sources. Contamination with human DNA may also occur during the excavation and handling of a specimen. Contamination can also occur in a museum where the specimen may come in contact with modern material or animal-based glue or varnish. Lastly, there are multiple

sources of possible contamination in the laboratory including trace DNA in reagents or amplicons from previous experiments [16].

Early aDNA experiments [15], also identified *post-mortem* damage as a further serious complication for studying aDNA. After death, the cellular machinery that maintained the integrity of DNA ceases to function and DNA starts to acquire damage through decay processes. Initially, DNA is fragmented by degradative enzymes released from the break down of cellular compartments *post-mortem* and from organisms feeding on macro-molecules [17]. DNA that survives this initial biological attack is then subjected to a slower chemical decay primarily oxidative and hydrolytic damage [18]. The end result of these decay processes is that aDNA is fragmented into small molecules generally < 150 base pairs in length and contains abasic sites, intra- and inter-strand crosslinks, and modified bases [10,15].

Preservation of aDNA is dependent on environmental factors including moisture, salinity, and low temperature [19,20]. The significance of low temperature in preserving aDNA is evident as the oldest authenticated molecules are from samples found in permafrost conditions, such as ≈ 800 thousand year (kyr) old plant and invertebrate sequences from ice cores [8,21] and the genome of a > 500 kyr old archaic horse [22]. Permafrost samples often contain large aDNA molecules with fragments size reaching > 500 base pairs [23,24], again demonstrating the high levels of preservation produced by low-temperature conditions. However, even under cold conditions the preservation of nucleic acids can be variable as factors in the microenvironment such as moisture and salinity will cause variation in the survival of aDNA [18]. Nonetheless, aDNA is unlikely to survive over geological timescales and

by extrapolating from *in vitro* experiments it has been estimated that aDNA will not survive more than a million years [18,25].

Recently, the field of aDNA has made considerable progress in obtaining sequence data from samples collected from less than ideal conditions. Using extraction and library construction protocols that were optimized for recovery of short DNA fragments, mitochondrial sequences from a Middle Pleistocene (≈ 300 kyr) cave bear [26] and hominin [27] have been obtained from non-permafrost samples. Continued technical advances will push the limits of age and environmental conditions from which aDNA can be recovered.

1.1 DNA Sequencing and Polymerase Chain Reaction (PCR)

In early studies, aDNA was amplified with bacterial cloning and then sequenced. In this procedure, aDNA was extracted from a specimen and then ligated into a plasmid that was subsequently used to transfect bacteria. The bacteria were grown to produce many copies of the aDNA molecule, which was then sequenced with Sanger sequencing (chain-termination sequencing), the first widely used method for sequencing DNA. For Sanger sequencing, aDNA is denatured and a primer is annealed to one of the strands before replication is performed using a DNA polymerase with a mixture of deoxynucleotide triphosphates (dNTPs) and dideoxynucleotide triphosphates (ddNTPs). ddNTPs lack the 3' OH moiety, which is essential for the addition of the subsequent nucleotide during DNA synthesis. Replication with a small fraction of ddNTPs produces DNA fragments of varying lengths and the size of these molecules can be used to determine the sequence of the original DNA template [28,29]. Although bacterial cloning demonstrated that aDNA

could be recovered from biological specimens the technique has severe limitations. First, cloning recovers random DNA molecules so results can not be easily replicated [30] and second, cloning preferentially captures high copy number loci such as mitochondrial DNA [19].

Several years after the first aDNA study, the polymerase chain reaction (PCR) was described [31] and quickly replaced bacterial cloning as the primary method of amplification [15]. With PCR, Sanger sequencing remained the principal method for sequencing amplified products. The exponential power of PCR allowed the amplification and study of rare aDNA molecules and engendered an expansion and diversification of aDNA studies including investigations of systematics [32] and human evolution [33]. However, the exponential amplification produced by PCR makes contamination with even trace amounts of modern DNA a serious concern for aDNA studies, and a number of guidelines have been established to help ensure the authenticity of aDNA results [34,35]. Despite the advances represented by PCR and Sanger sequencing these technologies have a relatively low throughput that make it economically unfeasible to perform large scale studies such as genome sequencing with aDNA [36].

1.2 High Throughput Sequencing (HTS)

In 2005, high throughput sequencing (HTS) was introduced [37] and these technologies were quickly adopted by evolutionary science and caused a revolution in the field, particularly for aDNA studies [38]. HTS encompasses a group of technologies that sequence thousands of DNA molecules in parallel and drastically reduces the cost and time of generating data [38].

HTS can be roughly divided chronologically into an initial and a later wave of sequencing advances. Initial HTS technologies (e.g. Illumina/Solexa sequencing by synthesis and Life Sciences/Roche 454 pyrosequencing) became available in 2005 [37] and have been extensively adopted in aDNA research. Generally, these sequencers require study DNA to be converted into a sequencing library through enzymatic ligation of oligonucleotide adapters, which provide binding sites for primers that amplify the entire sequencing library. This first wave of HTS sequencers produce data from populations of clonal amplicons with each population generated from a single strand of DNA. These early HTS sequencers are also light based technologies that use sequencing by synthesis to generate data [39].

An example of an initial HTS technology is Illumina's single end sequencing, which is performed in a specialized glass chamber called a flow cell. Attached to the bottom of the flow cell are two populations of oligonucleotides and each of the populations is complimentary to one of the two adapters used to convert the study DNA into a sequencing library. Additionally, one of the oligonucleotide populations is attached to the flow cell with a cleavable diol linker. Prior to sequencing, a sequencing library is taken through a process called cluster generation in the flow cell. First, the sequencing library is denatured into Strand A and Strand B molecules and loaded in the flow cell. For illustrative purposes this description will follow a single Strand A through the cluster generation (Figure 1) and sequencing (Figure 2) processes. After loading the library, an adapter sequence in the Strand A molecule anneals with a complimentary oligonucleotide that lacks the diol linker. The oligonucleotide is extended by a DNA polymerase to produce a Strand B molecule that is attached to the flow cell. The original Strand A is removed from the flow cell through denaturation and washing.

Bridge amplification (amplification on a solid support) is then performed with DNA polymerase to produce a cluster of Strand A and B molecules and all Strand A molecules will be attached to the flow cell by oligonucleotides that contain a diol linker. The flow cell is treated with periodate to cleave the diol linkers, which allows the Strand A molecules to be removed from the cluster by washing. Sequencing is then initiated by annealing a primer and a DNA polymerase to the Strand B molecules in the cluster and subsequently adding reversible terminator nucleotides, which are the canonical nucleotides that contain a unique fluorescent label for each base and a reversible terminator attached to the 3' OH group of the pentose ring that prevents the incorporation of additional residues during DNA synthesis [40]. Both the fluorescent labels and the reversible terminators can be chemically cleaved from the nucleotide. During sequencing, the DNA polymerase incorporates a single nucleotide and the reversible terminator blocks further extension of the primer. A camera records which nucleotide was incorporated by the signal produced from the fluorescent label. The fluorescent label and reversible terminator are then chemically removed and the process of extending the primer by a single nucleotide is repeated.

These early HTS technologies have evolved into platforms with robust sequencing capacities. In a single run, the Illumina HiSeq 2500 platform can produce 600 gigabases (Gb) from 3 billion reads whilst the Roche 454 GS FLX+ sequencer can generate 0.7 Gb from a million reads. Importantly, the read lengths for these two platforms are quite different the HiSeq 2500 produces 2 x 100 base pair reads in comparison to the 700 base pair average for the 454 GS FLX+ [41].

In contrast, only a few of the more recent HTS technologies (e.g. Ion Torrent and Pacific Biosciences) have become commercially available and these sequencers have not been widely adopted in aDNA research. These later sequencers are also a much more diverse group of technologies with no common method for preparing DNA or methodological principle used for sequencing [42]. An example of these more recent HTS technologies is nanopore sequencing in which DNA is denatured and one of the strands is threaded through a small pore. As a nucleotide passes through the aperture a metric such as current change is used to record the identity of the residue [43].

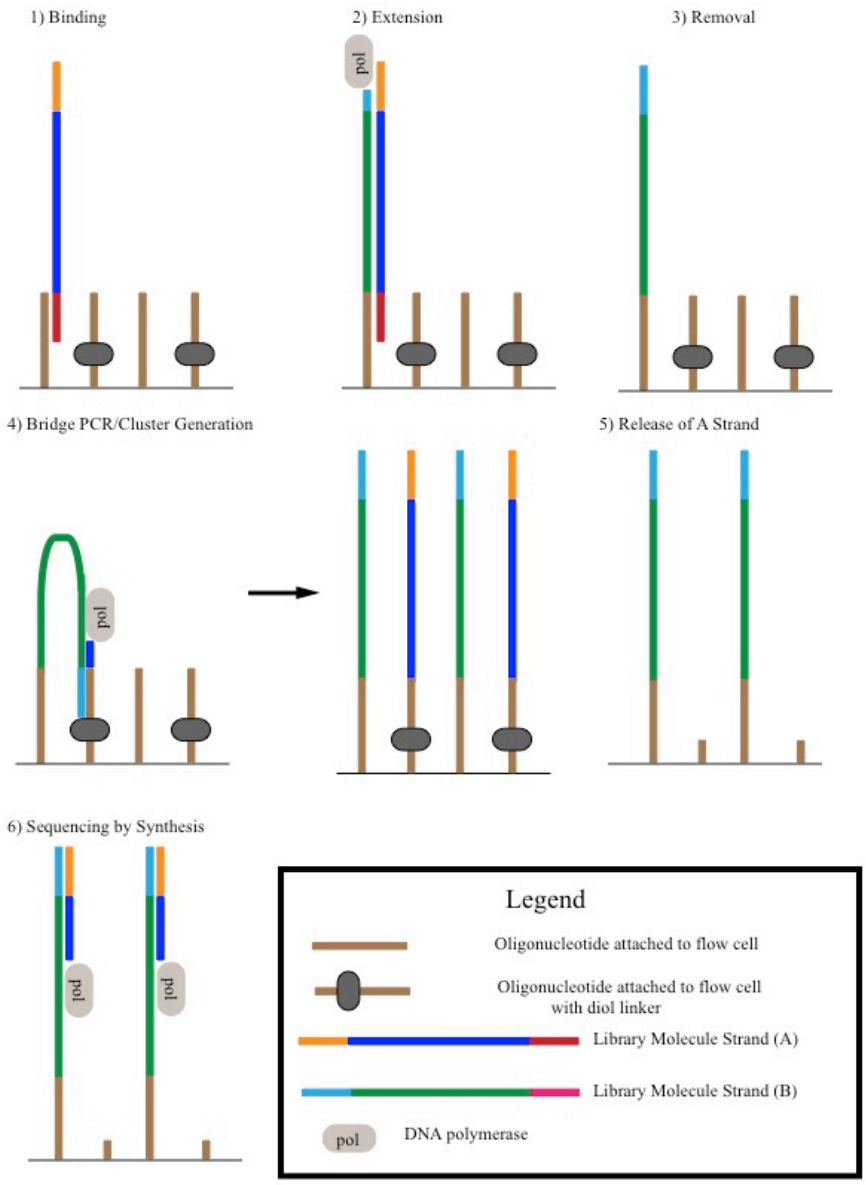


Figure 1. Illustration of cluster generation for Illumina's single end sequencing
Cluster generation and sequencing are performed in a glass chamber called a flow cell. Attached to the bottom of the flow cell are two populations of oligonucleotides that are complimentary to portions of the adapters of the sequencing library molecules. One of the oligonucleotide populations is also attached to the flow cell by a cleavable diol linker. A sequencing library is denatured in to Strand A and B molecules and loaded in to a flow cell. 1) In this example, a Strand A anneals to an oligonucleotide that lacks the diol linker. 2) A DNA polymerase then extends the oligonucleotide to produce a new Strand B molecule attached to the flow cell. 3) The original Strand A molecule is removed through denaturation and washing. 4) Bridge amplification is used to produce a cluster of Strand A and B molecules. 5) The flow cell is treated with periodate to cleave the diol linkers and Strand A molecules are released and washed from the cluster. 6) Sequencing by synthesis is performed using reversible terminator nucleotides (Figure 2).

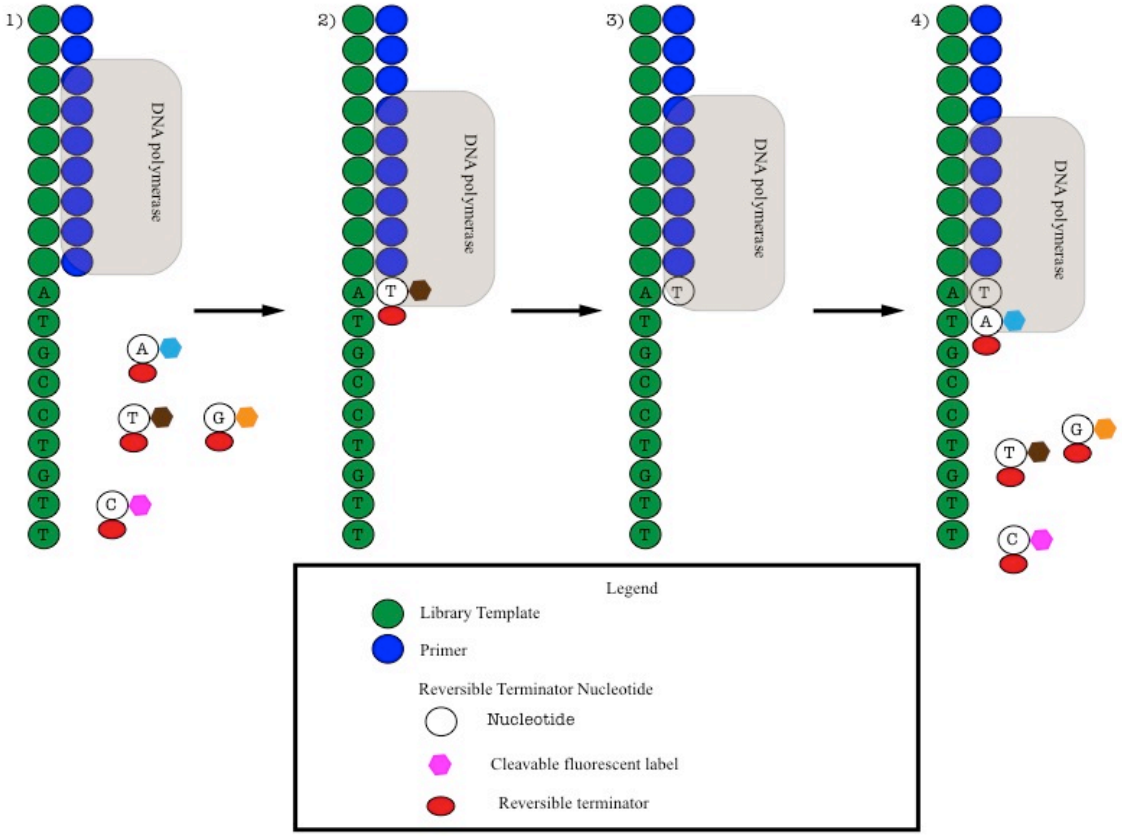


Figure 2. Illumina's sequencing by synthesis using reversible terminator nucleotides
Reversible terminator nucleotides contain a unique fluorescent label for each base and a reversible terminator attached to the 3' OH group of the pentose ring to prevent the incorporation of additional residues. Both the fluorescent label and the reversible terminator can be chemically cleaved from the nucleotide. 1) A primer and DNA polymerase are bound to a template attached to a flow cell and the reversible terminator nucleotides are then added. 2) The DNA polymerase incorporates a single nucleotide and the reversible terminator prevents the addition of further nucleotides. A camera records which nucleotide was added by the signal produced from the fluorescent label. 3) The fluorescent label and reversible terminator are then chemically cleaved from the incorporated nucleotide. 4) The process of incorporating nucleotides is repeated.

HTS has facilitated a number of advances in aDNA research that would not have been feasible using Sanger sequencing. The large sequencing capacity of HTS provides a brute force solution to the high levels of contaminating environmental DNA typically found in an ancient extract. Ancient shotgun libraries generally have a low concentration of endogenous DNA and hence HTS has enabled the analysis of genomes from a number of organisms including early modern humans [4,44,45],

archaic humans [46,47], and an archaic horse [22] by sheer volume of data. The large throughput of HTS has also allowed for detailed metagenomics studies such as determining the bacterial communities present in the oral cavity of past human populations [6,48] and the biodiversity of past environments [7,8].

Since the introduction of HTS, advances in sequencing equipment and chemistry have greatly expanded throughput capacity, whilst also increasing read length. These improvements have led to a dramatic reduction in sequencing cost, which has made a wide range of studies possible for even moderately funded laboratories. However, advances for some of the ancillary technologies needed for performing studies have not kept pace with those of HTS, notably the capability of computers and data storage devices (Figure 3). Many sequencing datasets produced with HTS are too large and complex to be processed by desktop computers, so obtaining the capability to analyze and store sequencing data has become a serious concern for many research groups. To adapt to the massive amounts of data produced by HTS, groups must establish adequate computing infrastructure to handle the volume of information generated by these sequencing technologies [49]. Furthermore, data analysis programs are constantly being improved to take advantage of new computational methodologies [50,51].

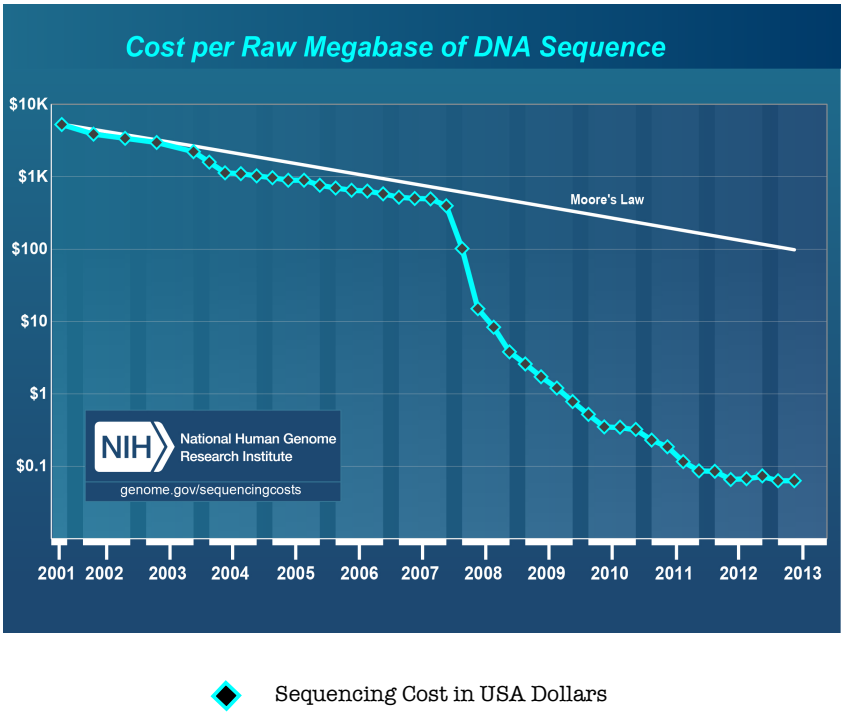


Figure 3. Cost of DNA Sequencing and Moore's Law
Since 2008, DNA sequencing cost has dropped dramatically because of the increase in the capacity of high throughput sequencing (HTS). Moore's law is a prediction made by Gordon E. Moore that the number of transistors on integrated circuits would double approximately every two years. The number of transistors on integrated circuits determines the capability of digital devices, such as computers. The capabilities of HTS sequencers have far out paced computers making data storage and analysis problematic. Figure used with permission from: Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed: January 12, 2014.

The molecular methods for isolating and preparing DNA for HTS are also constantly being refined and this is especially true for aDNA because of the inherent contamination and *post-mortem* damage. Laboratories are constantly investigating new methods to improve aDNA extraction and sequencing library construction. Changes to extraction methods have been reported to increase the recovery of endogenous sequences [52] and retrieval of short aDNA fragments [26], while the use of certain DNA polymerases can improve sequencing library fidelity [53]. Blunt–end ligation to attach adapters has also been reported to produce a library with less sequence bias than AT overhang ligation [54]. Despite these improvements and others

made over the last decade there is still a need to increase the quality of HTS data generated with aDNA. For example, many analysis of aDNA are still dependent on generating deep coverage by HTS, such as identifying the nucleotides in a heterozygous SNP, and increasing data quality will reduce the reliance on sequencing volume.

1.3 Microarray Technologies

The proliferation of sequenced genomes engendered by HTS has also led to the growth of microarray technologies [55]. Microarrays are an array of oligonucleotide probes attached to a solid support, usually a glass slide, and are produced with an assortment of technologies [56]. One type of microarray that has proven to be useful for evolutionary science is genotyping microarrays that assay single nucleotide polymorphisms (SNPs) throughout the genome of a specific organism. For genotyping, a sample is incubated on a SNP microarray allowing targets in the DNA to anneal to complimentary probes. Unbound material is then washed away and a fluorescent reporter system is then used to identify the SNPs in the DNA attached to the probes. SNP microarrays are a quick and inexpensive method to produce a large amount of data and commercially available microarrays can currently interrogate >900,000 SNPs. SNP microarrays are commonly used in a variety of studies involving modern DNA, including medical research [57] and genotyping agriculturally important plants and animals [58,59]. Genotyping microarrays have also been applied to a variety of evolutionary studies using modern DNA [60,61], however there is currently only one study that reports the use of genotyping microarrays to assay SNPs in aDNA [62]. The degraded nature of aDNA will make

genotyping ancient samples with SNP microarrays a challenging task and such studies will require a considerable amount of optimization.

Another use of the technology used to produce microarrays is the fabrication of probes for hybridization capture, which is a method to enrich loci of interest from a sequencing library. Hybridization capture is a technique that makes use of RNA or DNA complimentary oligonucleotide probes to bind and immobilize target loci in a sequencing library. Once targets are immobilized, unwanted library is washed away and the molecules bound to probes are released for sequencing by HTS. Hybridization capture can be performed with the probes attached to a solid support such as a microarray or in solution after the oligonucleotides have been stripped off a solid support [56,63]. Hybridization capture is particularly useful in aDNA studies because unwanted molecules, such as environmental contamination, are removed from a library prior to sequencing. With lower levels of contamination significant coverage of targets can be achieved with less sequencing and reduced cost. Because of this cost saving hybridization capture is becoming common in aDNA studies [12,64]. Although commercially bought probes will save on sequencing costs, the oligonucleotides themselves are expensive and many laboratories are developing in house methods to synthesize these reagents [12,65].

2.0 Ancient Bison: An Ideal Model for Megafaunal Evolution

The primary goal of this thesis was to develop or improve-on techniques for studying aDNA with HTS and genotyping microarrays and apply those methods to larger biological questions of evolution. Most of the effort of this thesis was directed at developing molecular methods for aDNA study but analytical tools were also

examined. For the broader biological aspect of this thesis, bison were chosen as a study organism because in many ways these animals are an ideal model of megafaunal evolution. Bison are large even-toed ungulates in the genus *Bison* of the family Bovidae. There are six recognized species of bison, four of which are extinct and two are extant. Three of the extinct bison species are *Bison antiquus*, *Bison latifrons*, and *Bison occidentalis,* which inhabited North America. The last extinct bison, *Bison priscus*, occupied the entire Holarctic, a region that spanned northern Mexico through Eurasia. The extant species are the European bison (*Bison bonasus*) and the modern American bison (*Bison bison*) [66,67].

The *Bison* genus originated in Asia and split from the lineage that produced domestic cattle between 2 to 5 million years (Myr) ago [68]. Bison fossils exhibit a large morphological variability, which may represent a wide range of species and sub-species. During the Late Pleistocene (126-11 kyr) separate bison species were present in the New and Old World and at least one identified species, *Bison priscus*, inhabited both regions. These past bison populations have left a large and diverse fossil record that will allow population level phylogenetic studies of the taxa to be performed across time and space [69].

Bison fossils form a record of past events that would have placed strong adaptive stresses on megafauna such as climate change, population fragmentation, and human settlement [16]. During the late Pleistocene, there were a series of extreme climate oscillations that produced drastic environmental alterations, and as a consequence, bison fossils contain a record of the impact of these past climate events through changes in morphology and aDNA [70,71]. Adaptive pressure during the late

Pleistocene is likely to have driven changes in nucleotide sequence and epigenetic modification of bison aDNA. Groups of the bison taxa have sometimes been separated by glacial ice sheets or isolated in different refugia by climate change [16], and aDNA will allow the investigation of these population fragmentations across a time-series. Similar to other New World megafauna such as mastodons and woolly mammoths, bison were present in North America prior to the arrival of humans and hence bison fossils are likely to record the effects of human colonization and habitat modification [69].

The well-preserved state of many bison fossils offers the opportunity to perform phylogenetic studies with high quality aDNA. Much of the extensive bison fossil deposits can be found in permafrost areas such as the Yukon Territory (Canada) and Alaska (United States), which provide excellent conditions for the preservation of aDNA for genetic studies and collagen for carbon dating. In North America, many bison remains are exposed during surface gold mining operations and consequently the fossils have limited surface exposure. Previously it has been possible to amplify aDNA sequences of 2,000 base pairs in length from North American bison fossils, demonstrating the high levels of preservation in these specimens [16]. Such well-preserved aDNA will likely facilitate the generation of larger data sets such as the sequencing of whole exomes (protein coding regions of a genome) or genomes. Furthermore, the extant bison species (European and American bison) can serve as genetic, genomic, and functional references for evolutionary studies of the bison taxa.

3.0 Outline of Thesis Chapters

This thesis contains five experimental chapters that examined two themes: methods development and bison evolution. Each of the experimental chapters is written as a manuscript draft for submission to a peer-reviewed journal. The order of the chapters does not represent the chronology in which experiments were performed. Some of the methods described in this thesis were not successfully performed until late in the PhD candidature and therefore could not be applied to the studies of bison evolution. Chapters II to IV pertain to library construction and sequencing accuracy, whilst V and VI examine bison evolution. A brief summary of each chapter and the journal to which it is formatted are below:

Chapter II (target journal: PLOS ONE)

Title: Optimizing Ancient DNA Sequencing Libraries

The two goals of this study were to investigate the effects of certain enzymatic treatments in converting bison aDNA into sequencing libraries and evaluate those enzymatic treatments in a novel hybridization capture system. Library construction is a critical step during HTS and can have a marked impact on the final sequence data [54,72]. For the first goal of this study, four different enzymatic treatments for library construction were evaluated to determine which maximized endogenous sequencing data from bison aDNA. In the second goal, the effect of the enzymatic treatments were examined in a novel hybridization capture system that produced RNA probes from long-range PCR amplicons.

Chapter III (target journal: PLOS ONE)

Title: Isothermal Amplification in Hybridization Capture of Ancient DNA

PCR is typically used to amplify aDNA in hybridization capture studies. However, PCR does not amplify with perfect fidelity and the technique will change the sequence composition of a sequencing library [53]. Isothermal amplification methodologies [73] may provide alternatives to PCR in hybridization capture. Isothermal methods amplify at a single relatively low temperature and denature DNA with enzymatic activity instead of heat. In this study, two isothermal amplification techniques were compared to PCR in the hybridization capture of a nuclear locus from bison aDNA sequencing libraries.

Chapter IV (target journal: PLOS ONE)

Title: Detection of Altered Bases in Ancient DNA Using SMRT Sequencing

Various altered bases have been described in aDNA molecules including nucleotides containing damage [15] and residues modified through epigenetic pathways [74]. In aDNA molecules, nucleotides are damaged in aDNA through various biological and chemical decay processes [18]. Epigenetic modifications are biochemical changes to DNA that participate in gene regulation but do not alter the underlying nucleotide sequence [75]. Little is known about the modified bases present in aDNA because many sequencing technologies require amplification, which if done with PCR or cloning will erase information on nucleotide modification from sequence data [10,76]. Single Molecule Real-Time (SMRT) is a HTS technology that can generate data from a single strand of unamplified DNA and can detect modified bases through changes in the kinetics of DNA polymerases [77]. For this study, unamplified steppe bison

aDNA was sequenced with SMRT sequencing in an attempt to identify the types and abundance of modified bases present.

Chapter V (target journal: Molecular Ecology)

Title: Paleoclimatic Impacts on European Bovid Megafauna in the Late Pleistocene

In this manuscript a new species of extinct bison, called *Bison X*, was described using molecular and morphological data. *Bison X* inhabited Eurasia from before the Oxygen Isotope Stage 3 (55 kyr) but disappeared from the paleontological record sometime after the Last Glacial Maximum. Data generated in this study gave strong support for *Bison X* being a separate sister species to the extant European bison. Phylogeographic data indicated that there were several replacement events between *Bison X* and the contemporaneous steppe bison, which were strongly correlated with changes in climate.

Chapter VI (target journal: PLOS ONE)

Title: Elucidating Bovid Evolution with Genotyping Technologies

The Illumina BovineSNP50 BeadChip is a SNP microarray that is designed to genotype taurine cattle breeds. Because the BovineSNP50 produces a considerable amount of data quickly and at a low cost, it is tempting to use the microarray to generate phylogenetic data on taxa related to cattle. The BovineSNP50 was not design for evolutionary studies and there are challenges that need to be addressed for using this microarray for phylogenetic analysis. This manuscript examined the data analysis issues of using the BovineSNP50 for phylogenetic studies in general and technical problems specific to aDNA.

Hybridization capture is another technology that may be applied to genotyping of aDNA. Probes can be designed to enrich molecules that contain known SNPs, which are subsequently sequenced with HTS to produce genotype data. In this study an alternative probe design was evaluated in producing genotyping data from aDNA. Typically in hybridization capture of aDNA several overlapping probes are designed to enrich a locus of interest. In order to maximize the number of SNPs that could be enriched with a commercially produced hybridization capture kit the current study evaluated the efficiency of enriching each variation with a single probe.

References

1. Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312: 282-284.
2. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft Sequence of the Neandertal Genome. Science 328: 710-722.
3. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, et al. (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. Nature 478: 506-510.
4. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757-762.
5. Oskam CL, Haile J, McLay E, Rigby P, Allentoft ME, et al. (2010) Fossil avian eggshell preserves ancient DNA. Proceedings of the The Royal Society Biological sciences 277: 1991-2000.
6. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, et al. (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nature Genetics 45: 450-455.
7. Boessenkool S, McGlynn G, Epp LS, Taylor D, Pimentel M, et al. (2013) Use of Ancient Sedimentary DNA as a Novel Conservation Tool for High-Altitude Tropical Biodiversity. Conservation Biology 28: 446-455.
8. Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, et al. (2007) Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science 317: 111-114.
9. Avila-Arcos MC, Cappellini E, Romero-Navarro JA, Wales N, Moreno-Mayar JV, et al. (2011) Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA. Scientific Reports 1: 1-5.
10. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, et al. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35: 5717-5728.
11. Krause J, Dear PH, Pollack JL, Slatkin M, Spriggs H, et al. (2006) Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. Nature 439: 724-727.
12. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
13. Kistler L, Montenegro Á, Smith BD, Gifford JA, Green RE, et al. (2014) Transoceanic drift and the domestication of African bottle gourds in the Americas. Proceedings of the National Academy of Sciences 111: 2937-2941.
14. Pääbo S (1985) Molecular cloning of Ancient Egyptian mummy DNA. Nature 314: 644-645.
15. Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 1939-1943.
16. Shapiro B, Cooper A (2003) Beringia as an Ice Age genetic museum. Quaternary Research 60: 94-100.

17. Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, et al. (2004) Genetic analyses from ancient DNA. Annual Review of Genetics 38: 645-679.
18. Dabney J, Meyer M, Pääbo S (2013) Ancient DNA Damage. Cold Spring Harbor Perspectives in Biology 5: 1-6.
19. Willerslev E, Cooper A (2005) Ancient DNA. Proceedings of the Royal Society B-Biological Sciences 272: 3-16.
20. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S (2012) Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLoS ONE 7: 1-7.
21. Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MTP, et al. (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300: 791-795.
22. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, et al. (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499: 74-78.
23. Lambert DM, Ritchie PA, Millar CD, Holland B, Drummond AJ, et al. (2002) Rates of evolution in ancient DNA from Adelie penguins. Science 295: 2270-2273.
24. Lydolph MC, Jacobsen J, Arctander P, Gilbert MTP, Gilichinsky DA, et al. (2005) Beringian Paleoecology Inferred from Permafrost-Preserved Fungal DNA. Applied and Environmental Microbiology 71: 1012-1017.
25. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.
26. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, et al. (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences 110: 15758-15763.
27. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, et al. (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505: 403-406.
28. Adams J (2008) DNA sequencing technologies. Nature Education 1: 193.
29. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America 74: 5463-5467.
30. Krause J (2010) From Genes to Genomes: What is New in Ancient DNA? Mitteilungen der Gesellschaft für Urgeschichte. Blaubeuren / Tübingen: Society for Prehistory and Friends of the Prehistory Museum Blaubeuren. pp. 11-34.
31. Mullis KB, Faloona FA (1987) Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology 155: 335-350.
32. Cooper A, Mourer-Chauviré C, Chambers GK, von Haeseler A, Wilson AC, et al. (1992) Independent origins of New Zealand moas and kiwis. Proceedings of the National Academy of Sciences 89: 8741-8744.
33. Lalueza C, Pérez-Pérez A, Prats E, Cornudella L, Turbón D (1997) Lack of Founding Amerindian Mitochondrial DNA Lineages in Extinct Aborigines from Tierra del Fuego-Patagonia. Human Molecular Genetics 6: 41-46.
34. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
35. Hofreiter M, Serre D, Poinar HN, Kuch M, Pääbo S (2001) Ancient DNA. Nature Reviews Genetics 2: 353-359.

36. Rizzi E, Lari M, Gigli E, De Bellis G, Caramelli D (2012) Ancient DNA studies: new perspectives on old samples. Genetics Selection Evolution 44: 1-19.
37. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.
38. Shapiro B, Hofreiter M (2014) A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. Science 343.
39. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, et al. (2009) The challenges of sequencing by synthesis. Nature Biotechnology 27: 1013-1023.
40. Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.
41. Lee C-Y, Chiu Y-C, Wang L-B, Kuo Y-L, Chuang EY, et al. (2013) Common applications of next-generation sequencing technologies in genomic research. Translational Cancer Research 2: 33-45.
42. Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759-769.
43. Maitra RD, Kim J, Dunbar WB (2012) Recent advances in nanopore sequencing. Electrophoresis 33: 3418-3428.
44. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, et al. (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505: 87-91.
45. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, et al. (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature 506: 225-229.
46. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.
47. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.
48. Warinner C, Rodrigues JF, Vyas R, Trachsel C, Shved N, et al. (2014) Pathogens and host immunity in the ancient human oral cavity. Nature Genetics 46: 336-344.
49. Lampa S, Dahlo M, Olason PI, Hagberg J, Spjuth O (2013) Lessons learned from implementing a national infrastructure in Sweden for storage and analysis of next-generation sequencing data. Gigascience 2: 9.
50. Huang K, Yellapantula V, Baier L, Dinu V (2013) NGSPE: A pipeline for end-to-end analysis of DNA sequencing data and comparison between different platforms. Computers in Biology and Medicine 43: 1171-1176.
51. Li B, Zhan X, Wing MK, Anderson P, Kang HM, et al. (2013) QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data. BioMed Research International 2013: 1-4.
52. Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, et al. (2011) True single-molecule DNA sequencing of a pleistocene horse bone. Genome Research 21: 1705-1719.
53. Dabney J, Meyer M (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.

54. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, et al. (2013) Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. PLoS ONE 8: 1-11.
55. Mu J, Seydel KB, Bates A, Su XZ (2010) Recent Progress in Functional Genomic Research in Plasmodium falciparum. Current Genomics 11: 279-286.
56. Murgha YE, Rouillard J-M, Gulari E (2014) Methods for the Preparation of Large Quantities of Complex Single-Stranded Oligonucleotide Libraries. PLoS ONE 9: 1-10.
57. Mao X, Young BD, Lu YJ (2007) The application of single nucleotide polymorphism microarrays in cancer research. Current Genomics 8: 219-228.
58. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, et al. (2011) A Large Maize SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. PLoS ONE 6: 1-15.
59. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, et al. (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: 1-13.
60. Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, et al. (2009) Fine-scaled human genetic structure revealed by SNP microarrays. Genome Research 19: 815-825.
61. vonHoldt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, et al. (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Research 21: 1294-1305.
62. Decker JE, Pires JC, Conant GC, McKay SD, Heaton MP, et al. (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proceedings of the National Academy of Sciences 106: 18644-18649.
63. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111-118.
64. Castellano S, Parra G, Sánchez-Quinto FA, Racimo F, Kuhlwilm M, et al. (2014) Patterns of coding variation in the complete exomes of three Neandertals. Proceedings of the National Academy of Sciences 111: 6666-6671.
65. Fu QM, Meyer M, Gao X, Stenzel U, Burbano HA, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences of the United States of America 110: 2223-2227.
66. Guthrie RD (1970) Bison Evolution and Zoogeography in North America During the Pleistocene. The Quarterly Review of Biology 45: 1-15.
67. Prusak B, Grzybowski G, Zieba G (2004) Taxonomic position of *Bison bison* (Linnaeus 1758) and *Bison bonasus* (Linnaeus 1758) as determined by means of cytb gene sequence. Animal Science Papers and Reports 22: 27-35.
68. McDonald JN (1981) North American Bison, Their classification and Evolution Berkeley, Los Angeles, London: University of California Press.
69. Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, et al. (2004) Rise and fall of the Beringian steppe bison. Science 306: 1561-1565.
70. Wolff EW, Chappellaz J, Blunier T, Rasmussen SO, Svensson A (2010) Millennial-scale variability during the last glacial: The ice core record. Quaternary Science Reviews 29: 2828-2838.

71. Martin PS (1984) Prehistoric overkill: the global model; Martin PS, Klein RG, editors. 354-403 p.
72. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biology 14: 1-20.
73. Gill P, Ghaemi A (2008) Nucleic acid isothermal amplification technologies - A review. Nucleosides Nucleotides & Nucleic Acids 27: 224-243.
74. Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, et al. (2012) High-Resolution Analysis of Cytosine Methylation in Ancient DNA. PLoS ONE 7: 1-6.
75. Laird PW (2010) Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics 11: 191-203.
76. Bart A, van Passel MW, van Amsterdam K, van der Ende A (2005) Direct detection of methylation in genomic DNA. Nucleic Acids Research 33: 1-6.
77. Clark TA, Spittle KE, Turner SW, Korlach J (2012) Direct detection and sequencing of damaged DNA bases. Genome Integrity 2: 1-9.

Statement of Authorship

Title of Paper	Optimizing Ancient DNA Sequencing Libraries
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards
Contribution to the Paper	Helped conceive study design, performed all experiments, helped analyze data, wrote paper
Signature	Date: June 18,2015

Name of Co-Author	Bastien Llamas
Contribution to the Paper	Helped analyze data, helped edit paper
Signature	Date: 18/06/2015

Name of Co-Author	James Breen
Contribution to the Paper	Helped analyze data, helped edit paper
Signature	Date: 18/06/2015

Name of Co-Author	Alan Cooper
Contribution to the Paper	Helped conceive study design, helped edit paper
Signature	Date: 24/06/2015

Optimizing Ancient DNA Sequencing Libraries

Stephen M. Richards*, Bastien Llamas, James Breen, and Alan Cooper

Australian Centre for Ancient DNA, University of Adelaide, South Australia, Australia

*Corresponding author
E-mail: steve.richards@adelaide.edu.au

Abstract
Sequencing libraries created from DNA are "immortal' resources because the molecules within the libraries can be amplified nearly unlimitedly. This is extremely valuable for ancient DNA (aDNA) studies as it allows large quantities of DNA to be generated from limited material and low starting concentrations. However, aDNA extracts are a complex mixture of exogenous and endogenous DNA molecules that are generally degraded, which complicates downstream analysis and raises costs by increasing the sequencing depth needed to achieve significant results for target loci. Consequently, it is desirable to optimize library construction to reduce the impact of nucleotide damage and exogenous contamination, while simultaneously maximizing the production of data from endogenous aDNA. In this study, four different enzymatic treatments were compared for their ability to increase the proportion of endogenous reads obtained via shotgun sequencing of aDNA libraries. The most effective treatment was found to be using Phusion DNA polymerase in combination with the USER enzyme cocktail mix, which removes uracil derived deaminated cytosine and subsequently cleaves the DNA molecule at the resulting abasic site. Additionally, a novel hybridization capture procedure was tested for library optimization by enriching enzyme treated libraries for the nuclear toll-like receptor 8 (TLR8) gene using RNA probes. In comparison to shotgun sequencing alone, the hybridization capture protocol resulted in over 100-fold enrichment of reads mapping to the TLR8 locus. USER treatment and the described hybridization capture protocol represent relatively simple and cost-effective methods to optimize aDNA sequencing libraries for generating data from endogenous molecules.

Introduction

For nearly two decades, the primary method of generating data from ancient DNA (aDNA) was Sanger sequencing of variable genetic loci following polymerase chain reaction (PCR) amplification [1]. However, PCR is a resource intensive methodology that quickly consumes an aDNA extract whilst producing relatively small amounts of

data, even with the amplification of several loci in the same reaction using multiplexing. The development of high throughput sequencing (HTS), which produces thousands of sequence reads in parallel, has drastically changed the paradigm of data generation from aDNA extracts. Most HTS platforms require aDNA extracts to be converted into sequencing libraries, by enzymatic ligating oligonucleotide adapters to the aDNA template. Adapters serve multiple functions in HTS including acting as binding sites for universal primers that can be used to amplify the entire library [2]. The ability to amplify the entire contents of an extract is extremely valuable for aDNA studies where starting materials can be quite small and the concentration of endogenous molecules low. The ligation of adapters converts an aDNA extract into an immortal resource that can be used in a nearly unlimited number of experiments [2], including those that require large amounts of DNA such as genome sequencing [3,4].

A complication of the HTS approach is that aDNA sequencing libraries are complex mixtures, containing not only endogenous molecules from the target organism but also exogenous DNA contamination and molecules with nucleotide damage. For most ancient extractions, endogenous molecules represent only a small proportion of the total aDNA whilst the vast majority will be exogenous contamination, likely to have originated from microorganisms that colonized the specimen *post-mortem* [5-7]. The ligation of adapters is non-specific and any sequencing library made from aDNA will contain both endogenous and exogenous molecules.

All aDNA contains some degree of nucleotide damage. After death, an organism's cellular repair machinery ceases to function and DNA molecules begin to accrue

damage through biological and chemical mechanisms. Biological factors that contribute to DNA damage include degradative enzymes liberated from the break down of cellular compartments *post-mortem* and organisms that feed on dead tissue. After the initial biological assault, surviving DNA then undergoes a slow chemical modification over time [7,8]. This *post-mortem* degradation manifests as abasic sites, inter- and intra-strand crosslinking, base modification, and fragmentation of aDNA into short molecules (generally less than 150 base pairs) [6,9]. Specific base modifications are of particular concern for aDNA studies because they act as miscoding lesions that cause some DNA polymerases to misincorporate nucleotides [10]. Deamination of cytosine to uracil is the major miscoding lesion present in aDNA and many common DNA polymerases, including Taq, read uracil as thymine and misincorporate an adenosine into the new DNA strand. Deamination of cytosine primarily occurs at the ends of aDNA because these regions tend to be single stranded and presumably more vulnerable to chemical attack [11]. Consequently, depending on the strand sequenced, misincorporation of adenosine creates the appearance of elevated 3' C → T or 5' G→A transitions in HTS data [9].

The complexity of sequencing libraries has important implications for aDNA research as sequencing exogenous molecules wastes resources and miscoding lesions can be mistaken for true genetic variations during downstream analysis [9]. Therefore, it is desirable to optimize library construction to minimize the proportion of exogenous DNA and limit the impact of damage induced miscoding. One of the aims of the present study was to compare the effects of four different enzymatic treatments (Figure 1) to find an optimal protocol for the construction of aDNA shotgun libraries. While prior aDNA studies have utilized these treatments [12-14], none have directly

compared the effect of these treatments in shotgun sequencing of aDNA libraries or investigate the treatments in combination.

Another aim of this study was to investigate the effects of the enzymatic treatments on a novel hybridization capture procedure, which uses RNA probes generated from long range PCR amplicons. Hybridization capture is a technique that utilizes RNA or DNA oligonucleotide probes to isolate loci of interest from a sequencing library [15-17]. Probes, which can be suspended in solution or attached to a solid support, are designed to hybridize and immobilize complimentary targets in a library, thus allowing unwanted sequences to be removed by washing [18]. Hybridization capture is a particularly valuable technique in aDNA research, as it allows the exclusion of exogenous molecules prior to sequencing.

All the enzymatically treated libraries in this study were prepared with aDNA extracted from a 26,000 year-old steppe bison (*Bison priscus*) bone. For hybridization capture, the toll-like receptor 8 (TLR8) locus was the target for enrichment in all cases.

Methods

A brief description of the enzymatic treatments is listed below and flow diagrams for all library construction protocols are given in Figure 1. A detailed description of all methods can be found in the Supplemental Methods section of this paper. Standard precautions were taken to ensure the authenticity of the aDNA data [19]. Extraction of aDNA and library construction were performed at the Australian Centre for Ancient

DNA (University of Adelaide, South Australia, Australia), in a dedicated aDNA laboratory that is cleaned routinely with bleach and is exposed to UV light after each procedure. The aDNA laboratory contained positive air pressure to minimize contamination from outside sources and access to the laboratory is limited to properly trained personnel. Control libraries were constructed from extraction blanks for each protocol and all PCRs included no template negative controls. Extraction blank libraries were tested for bison sequences with quantitative PCR (qPCR) and found to be negative. All no template controls for PCR were negative for product as assayed by gel electrophoresis.

Sample

The study aDNA was extracted from an astragalus bone of a steppe bison collected from a Late Pleistocene deposit in the Canadian Yukon Territory. The bone (ACAD Sample # A3133) was carbon dated at the Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom) with accelerator mass spectrometry. The sample produced an uncalibrated age of 26,360 ± 220 radiocarbon years before present using a ^{14}C half-life of 5,568 years. No permits were required for the described study, which complied with all relevant regulations.

aDNA Extraction

A Dremel tool with a carborundum cutting disk was used to remove a section of the astragalus bone, which was subsequently powdered in a Braun mikro-dismembrator. aDNA was extracted from 200 mg of bone powder using an established silica based extraction method and then stored at -20°C until needed [15,20].

Library Construction and Amplification

Library construction was performed according to a standard protocol that was altered for each of four enzymatic treatments (Figure 1):

• *Phusion*: Initial library amplification was performed with Phusion DNA polymerase, which does not efficiently amplify templates containing uracil and produces a library containing fewer uracil induced misincorporations [21].

•*USER+Phusion*: During library construction, aDNA was treated with the enzyme cocktail USER to remove uracils produced from deaminated cytosines [11]. USER is a combination of the enzymes uracil DNA glycosylate (UDG) and endonuclease VIII (EndoVIII). In aDNA treatment, UDG excises uracil bases forming abasic sites and then EndoVIII cleaves the sugar-phosphate backbone at the abasic sites generating templates ready for library construction.

•*Restriction Enzymes (RE) + Phusion*: After ligation of adapters, the library was treated with specific restriction enzymes to preferentially cut prokaryotic DNA [12], preventing the amplification of these contaminating sequences. The initial library amplification was then performed with Phusion to limit the amplification of templates containing uracil.

•*Combined*: All of the above treatments were used in library construction.

Each aDNA library was constructed using truncated versions of Illumina adapters [22] to improve enrichment efficiencies, as longer adapters can interfere with hybridization capture [23]. After construction, libraries were amplified by two rounds of low cycle PCR designed to produce the DNA concentrations needed for hybridization capture whilst minimizing the introduction of amplification biases [24]. The first and second amplifications were named whole extract amplification 1(WEA1) and whole extract amplification 2 (WEA2) respectively. In both WEA1 and WEA2, amplification was performed with universal primers that anneal to the library adapters. A fraction of the WEA2 from each library was further amplified using fusion primers (Table S1) to allow shotgun sequencing on an Ion Torrent PGM. *RE+Phusion* and *Combined* protocols utilized WEA2s made from pooling of two

separate libraries each treated with a different restriction enzyme cocktail (Figure 1). The pooling of libraries was designed to minimize the loss of steppe bison reads in the final sequencing data. The restriction enzymes should primarily cut prokaryotic contamination but some eukaryotic molecules will also be cleaved. Pooling of two libraries digested with different restriction enzymes should reduce the loss of endogenous molecules because different bison sequences will be lost in each digestion.

Generation of TLR8 Probe

The enzymatically treated steppe bison libraries were enriched for the TLR8 locus. TLR8 is a gene from the innate immune system and was chosen as a target for hybridization capture because its small size (3,102 base pairs) simplifies data processing and evaluation. Long range PCR was used to amplify a segment of the cattle (*Bos taurus*) X chromosome (AC_000187.1) spanning several hundred base pairs upstream and downstream of the TLR8 locus. The T7 RNA polymerase promoter was incorporated in the 5' end of the reverse primer (Table S1) allowing the amplicon to be used as a template for *in vitro* transcription. The TLR8 amplicon was transcribed in to RNA, which was subsequently converted into probe by fragmentation and biotinylation (Figure S1).

Hybridization Capture of TLR8

WEA2s from each library construction method were subjected to hybridization capture. Again, the *RE+Phusion* and *Combined* protocols utilized WEA2s made from pooling of two separate libraries each treated with a different restriction enzyme cocktail (Figure 1). DNA recovered from the first hybridization capture reaction was

amplified and then used in a second hybridization capture. Library recovered from the second enrichment was amplified once and a fraction of this material was then amplified a second time with fusion primers (Table S1) to attach full-length sequencing adapters to allow sequencing on an Ion Torrent PGM.

Ion Torrent Sequencing

Fusion primer-amplified shotgun and enriched libraries were pooled with libraries from other studies in equimolar amounts after quantification on an Agilent 2200 TapeStation and amplified with emulsion PCR using the Ion Torrent OneTouch DL system. The emulsion-amplified libraries were sequenced using an Ion Torrent PGM running Ion 316 chips. Each library was sequenced using a quarter of the capacity of a 316 chip.

Data Analysis

Analysis of sequencing data was performed with a pipeline that implemented several publically available software packages. Demultiplexing was carried out using FastX toolkit (version: 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit) and Cutadapt was used to remove adapter sequences and apply quality filters that retained reads 25 to 130 base pairs in length with a quality score ≥ 20 [25]. Filtered reads were mapped to the cattle genome (Btau 4.0) or TLR8 (chromosome X, AC_000187.1) references with BWA [26] and TMAP (https://github.com/nh13/TMAP) respectively. Filtered reads from the shotgun libraries were also mapped to a *Bison bison* mitochondrion reference (NC_012346.1) using TMAP. After mapping, clonal sequences were collapsed to produce unique mapped reads with Picard Tools (http://picard.sourceforge.net). Filtered reads were uploaded into the MG-RAST

metagenomics analysis server to evaluate the bacterial and steppe bison sequence content [27]. Since there is no bison reference genome for metagenomics analysis, reads that were classified as Bovidae (the family which includes bison) by MG-RAST were considered to be steppe bison sequences. Unique mapped reads were analyzed for patterns of deaminated cytosine-induced misincorporation using mapDamage [28,29].

Results

Library Characteristics

Ancient bison libraries from the four different enzymatic treatments and hybridization capture procedures were loaded in equimolar amounts for sequencing on an Ion Torrent PGM. Concordantly, the shotgun (unenriched) libraries produced roughly the same number of barcoded reads (raw sequences that contained the correct barcode) and filtered reads (sequences 25 to 130 base pairs in length with a minimum quality score ≥ 20). The number of barcoded reads ranged from 763,812 - 835,629 sequences whilst between 534,884 - 563,172 passed quality filters (Table 1A). Greater variation was observed among the TLR8 enriched libraries with barcoded reads ranging from 140,680 - 1,022,750 sequences whilst 101,282 - 713,391 reads passed quality filters (Table 1B).

Bacterial sequences tend to have higher GC content than eukaryotic genomes and consequently elevated GC content is an indicator of environmental contamination in ancient sequencing libraries [24]. In boxplot analyses the interquartile range (IQR;

which represents 50% of the data) of GC content for the filtered reads from the *Combined*-treated shotgun library was 37.0% – 52.1%, a range that covered the 42.4% GC content of the cattle genome. GC content IQRs for the filtered reads from the other shotgun libraries spanned 44.7% – 59.7%, values that were higher than the reference genome (Figure 2A). The GC content IQR for the filtered reads from the enriched *Combined* treated library ranged from 43.8% to 58.4%, the closest of any of the enriched libraries to the 42.1% GC content of the cattle TLR8 gene. Filtered reads from the remaining TLR8 enriched libraries yielded GC content IQRs spanning 48.9% – 62.8% (Figure 3A).

To further investigate the proportion of bacterial and endogenous reads, the shotgun and TLR8 enriched filtered data were uploaded to the MG-RAST metagenomics analysis server. MG-RAST determines the taxonomic distribution of a HTS dataset by identifying reads as belonging to taxonomic groups through comparison to references of annotated predicted proteins and ribosomal RNA genes. For each library, the percentage of reads identified as prokaryotic (bacterial contamination) or from the Bovidae family (endogenous bison) was determined (Table 1). It was expected that libraries producing a low fraction of prokaryotic reads and a high fraction of bovid reads would contain a high proportion of steppe bison reads in mapping analysis. For the shotgun libraries, the *Combined* treatment produced the lowest fraction of exogenous contamination (15.5% prokaryotic reads) and the highest percentage of endogenous DNA with 62.5% of the identified reads being from a bovid origin. For the other shotgun libraries, prokaryotic and bovid read fractions ranged 18.0% - 27.1% and 39.7% - 59.0% respectively (Table 1A). In the enriched libraries, the *Combined* treatment again produced filtered reads with the lowest percentage of prokaryotic reads (24.1%) and the highest percentage of bovid reads (63.2%). The

remaining treatments generated filtered reads where 26.7% - 30.3% of the identified sequences were called as prokaryotic and 51.3% - 57.6% bovid (Table 1B). Because library construction method is known to influence sequence length in HTS data, filtered read length distributions were examined for all libraries [24]. There was considerable overlap in read length between the treatments for both the shotgun and TLR8 enriched libraries. The length IQRs for the shotgun libraries spanned from 49.0 – 95.0 base pairs whilst the TLR8 enriched libraries spanned from 65.0 – 108.0 base pairs (Figures 2B and 3B). Generally, aDNA is <150 base pairs in length but a vast majority of unamplified molecules have been shown to be < 50 base pairs long [9]. The *Combined* shotgun and *Phusion* TLR8 enriched libraries produced mean read lengths of 67.9 and 81.7 base pairs respectively, the closest size to the majority of unamplified aDNA of all the libraries.

Shotgun Reads Mapped to Genome and Mitochondria References

The *USER+Phusion* treatment produced a shotgun library with the highest fraction (28.56%) and total number (156,608) of unique cattle genome-mapped reads (Phred score ≥ 30). The other treatments produced between 118,797 - 131,220 unique reads, which represented from 22.20% to 23.69% of the filtered reads for these libraries (Table 2). Additionally, the *USER+Phusion* shotgun library consistently had the most unique reads that mapped to the individual chromosomes of the reference. For the shotgun libraries, only a few (19 to 30) unique reads mapped to the Y chromosome (Figure 4).

In contrast to the nuclear genome results, the *Combined* treatment produced the highest number of unique reads that mapped to the *Bison bison* mitochondrial

reference with 49 sequences whilst the *USER+Phusion* treatment generated the second highest number of reads with 28 sequences. This relationship does not change when unique reads are examined as a fraction of the filtered reads (Table 2). However, these results might be a stochastic effect due to such a small data set. For all of the libraries, the number of reads mapping to the bison mitochondrial genome reference was so small that no interpretation of the effects of the enzymatic treatments on mitochondrial sequences could be made.

Shotgun and Hybridization Capture Reads Mapped to TLR8 Reference

For the shotgun libraries only one unique read (out of >750,000 barcoded reads, Table 1A) from both the *USER+Phusion* and *Combined* treatments mapped (Phred score ≥ 30) to the cattle TLR8 locus and no reads were reported for the other treatments (Table 3A). In contrast, hybridization capture-enriched libraries produced between 107 and 842 unique reads that mapped to the TLR8 reference. The *Combined* treatment produced the highest number of unique reads mapping after enrichment and *RE+Phusion* the least. The unique reads represented the highest fraction of the filtered reads in the *USER+Phusion* library at 0.57% whilst in the *Combined* library, unique reads were only 0.12% of the filtered reads. The *RE+Phusion* library contained the smallest percentage of unique TLR8 reads at 0.02%. Unique reads from the TLR8 enriched libraries covered 69.1% to 100% of the TLR8 gene with the *Combined* treatment producing the only library that covered the entire reference with a minimum depth of 1x (Table 3B, Figure S4).

The read length distributions for the *USER+Phusion* and *Combined* unique TLR8 reads both have a roughly Gaussian distribution weighted towards 75 and 65 base

pairs respectively (Figures 5B and 5D). This Gaussian distribution of read lengths has been reported in another hybridization capture study of aDNA [17]. In contrast, the results from the *Phusion* and *RE+Phusion* TLR8 unique reads are few and there appears to be no pattern to the read length distribution (Figures 5A and 5C).

Uracil-Induced Misincorporation

The mapDamage software was used to examine the shotgun and enriched libraries for patterns of deaminated cytosine-induced misincorporation [28,29]. Libraries that were not treated with the USER enzyme cocktail (and left uracils intact) exhibited elevated 3' C → T and 5' G→A transitions typical of deaminated cytosine induced miscoding (Figures 6 and 7). In contrast, the libraries treated with the USER enzyme cocktail exhibited no or minimal misincorporation.

Discussion

Library Characteristics

In this study enzymatic treatments and a hybridization capture method were investigated as possible techniques to optimize endogenous data generation from aDNA sequencing libraries. None of the enzymatic procedures appeared to have a serious impact on library quality as all the protocols produced a similar fraction of filtered reads (Table 1). However, differences were evident in the GC content of filtered reads and the relative concentration of endogenous and exogenous DNA sequences. The agreement of the GC content and the MG-RAST analysis suggests that the *Combined* treatment produced the libraries with the lowest level of bacterial contamination. However, a positive relationship between the GC content and

annotated prokaryotic reads in the MG-RAST analysis was not consistently observed in the other libraries. For example, in the *RE+Phusion* shotgun library there was an apparent negative relationship between these variables.

It is not clear why there is a large difference in GC content between the *RE+Phusion* and *Combined* libraries, as the restriction digestion should remove prokaryotic sequences similarly in both treatments. The only difference between these libraries was the removal of uracil by USER in the latter. If the endogenous aDNA was more heavily damaged than the prokaryotic contamination, then USER treatment may have rescued more steppe bison sequences for amplification by Phusion. Amplifying a greater fraction of bison aDNA may have shifted the GC content of the *Combined* filtered reads towards that of the cattle genome.

Another possibility is that USER treatment caused preferential cleavage of prokaryotic sequences in the *Combined* library. USER primarily acts by cleaving off the single strand ends of aDNA containing deaminated cytosine, however the enzyme cocktail will repair approximately 20% of the uracils in the double stranded portion of aDNA [11]. All of the restriction enzymes in this study recognize sequences that contain cytosine (Figure S2). In mammals, a fraction of the genomic DNA is protected from restriction enzyme cleavage by methylation of cytosines [30] and methylation does survive in aDNA [11,31]. Consequently, the steppe bison aDNA may have exhibited some protection from these enzymes whilst exogenous bacterial molecules remained susceptible to cleavage. Additionally, in the steppe bison aDNA a portion of the methylated cytosines would have under gone deamination to thymine [32], which is not repaired by USER [11], rendering the restriction site

unrecognizable to cognate restriction enzymes. Consequently, USER treatment may have repaired uracil damage in more restriction sites found in bacterial contamination than in endogenous aDNA, leading to a smaller fraction of prokaryotic sequences after amplification in the *Combined* treatment libraries.

Shotgun Reads Mapped to the Cattle Reference Genome

The *USER+Phusion* shotgun library produced the most unique reads that mapped to the cattle genome and consistently had the most unique reads that mapped to the individual chromosomes of the reference (Table 2, and Figure 4). The unique reads in the *USER+Phusion* shotgun library also represented the highest fraction of the filtered reads of any library. Removal of uracil by the enzymes in USER likely rescued damaged templates for amplification by Phusion and possibly enabled the analysis software to recognize and map additional bison sequences. The *USER+Phusion* treatment produced the optimal shotgun library despite the fact that the procedure did not produce the filtered reads with the lowest GC content (Figure 2A). Additionally, there appeared to be no relationship between the high number of mapped unique reads produced by the *USER+Phusion* shotgun library and the relative fractions of prokaryotic and bovid reads identified by MG-RAST analysis (Table 1A).

In shotgun sequencing, *Phusion* treatment was the worst performing with the lowest fraction of unique mapped reads (Table 2) and the least number of unique reads mapping to 20 out of the 31 cattle reference chromosomes (Figure 4).

In the genomic mapping data, there appeared to be little correlation between filtered GC content, the proportions of reads identified as prokaryotic versus bovid, and the

number of unique reads mapping to the cattle reference. Because bacterial DNA tends to have a higher GC content than eukaryotes, it would be expected that libraries with lower GC content would also have a lower fraction of reads identified as prokaryotic by MG-RAST analysis. However this was not always shown and there are several possible explanations for these inconsistencies. First, there may be a relationship between all these factors but a correlation may require replicate experiments with larger data sets to be reliably observed. Second, the reads mapped by BWA and identified as bovid by MG-RAST are different subpopulations of the filter reads in a library and these subpopulations may possess different characteristics.

In a previous study, restriction enzyme treatment increased mapped reads in a shotgun library from 3.1% to 13.1% [12] whilst in the current study, a similar endonuclease digestion increased reads in the *RE+Phusion* and *Combined* libraries by 1.1% and 1.49% in comparison to the *Phusion* treatment (Table 2). The variable effects of restriction enzyme treatment in these two studies could stem from several factors. The previous study used different restriction enzymes and these endonucleases may have been more effective in removing bacterial DNA. Additionally, the specimen from the previous study would have contained very different populations of aDNA fragments in terms of both endogenous and exogenous molecules. The previous specimen was a Neanderthal bone collected in central Europe and the aDNA from this sample may have been more receptive to restriction digestion than the aDNA extracted from the permafrost deposit steppe bison of the current study. Lastly, the two specimens had different starting fractions of endogenous aDNA: 3.1% for the Neanderthal and approximately 22% for the steppe bison, suggesting that restriction enzyme treatment may be more effective for lower quality samples.

Very few reads from the different shotgun libraries mapped to the reference Y chromosome. In contrast, the X chromosome which is approximately 3.5 times larger than the Y, consistently produced >100 fold more mapped reads regardless of treatment. Two explanations could account for the low number of reads mapping to the Y chromosome. The steppe bison used in this study could have been female and therefore there would have been no reads to map to the Y chromosome. Another possibility is that the bison was male, but the reads from the highly repetitive Y chromosome were not able to be mapped efficiently to the reference [33]. A recent study has published an analytical approach that predicts the sex of an ancient sample from aDNA shotgun data using the ratio of reads that mapped to the Y chromosome and the reads that mapped to both sex chromosomes [34]. Analysis of the steppe bison shotgun data using this approach strongly supported identifying this specimen as female.

Enriched Reads Mapped to the Cattle TLR8 Reference

In comparison to the shotgun libraries, the enriched libraries have a far higher number of unique reads that mapped to the cattle TLR8 reference (Table 3). For each treatment, enrichment increased the number of unique reads mapping to the TLR8 locus over 100 fold. This increase in reads clearly demonstrates that the hybridization capture protocol described in this study was successful in enriching for the TLR8 locus from steppe bison aDNA. Although successful, the hybridization capture system needs improvement as the sequences that mapped to the TLR8 reference remain only a small portion of the total number of the filtered reads (Tables 1B and 3B). Further

optimization of the hybridization conditions or the post-capture washes is likely to increase the fraction of TLR8 reads in an enriched library [35,36].

USER+Phusion treatment provided the best performance for TLR8 enrichment despite producing fewer unique reads and slightly lower coverage than the *Combined* library (Table 3B and Figure S4). This discrepancy was largely due to the relatively low number of barcoded reads obtained for the *USER+Phusion* TLR8 enriched library (Table 1B), possibly resulting from a pipetting or dilution error for this sample. The fraction of filtered reads uniquely mapping to the TLR8 reference was 0.57% for the *USER+Phusion* library compared to 0.12% for the next best performing library (*Combined)*. Given a similar number of barcoded reads as the other treatments, the enriched *USER+Phusion* library may have produced additional unique reads mapping to the reference and complete coverage of the TLR8 gene (Table 1B). The poor performance of the *Phusion* and *RE+Phusion* treatments in hybridization capture suggests that limiting the amplification of templates containing uracils was detrimental to enrichment.

Similar to the shotgun data, after hybridization capture there appeared to be no relation between filtered read GC content, MG-RAST metagenomics taxonomic assignment, and the number of unique mapped reads. It should be noted that the two extra rounds of PCR needed for hybridization capture could be responsible for the higher GC content of the *Combined* TLR8 enriched library in comparison to the *Combined* shotgun library (Figures 2A and 3A) as PCR is biased towards templates with higher GC content [24].

Several companies now commercially produce custom hybridization capture probes that can enrich for any chosen target, but these systems are expensive. The need to minimize costs during HTS is driving many laboratories to develop in-house hybridization capture procedures for aDNA. Recently, there have been several studies published using long-range PCR amplicons to produce probes to enrich for a diverse range of aDNA targets including human mitochondrial genomes and a virulence-associated plasmid from *Yersinia pestis* [15,37]. However, these studies differ from the enrichment protocol presented here in several ways. First, the previous studies fragmented and labeled the long-range product to produce DNA probes instead of using the amplicons as template for *in vitro* transcription of RNA probes. Additionally, an aqueous hybridization buffer was used in comparison to the formamide buffer used here. The hybridization capture system in the current study may offer better enrichment because the formamide buffer system can be manipulated to favor the formation of RNA-DNA hybrid molecules (i.e. probe-target) over the formation of DNA-DNA complexes (re-annealing of library) [38]. Furthermore, unlike the fragmented amplicon methods, the RNA used to make probes in this study was single stranded so there can be no probe-to-probe hybridization. A direct comparison will be needed to determine how the efficiency of the hybridization capture method described in the current paper compares to these alternative protocols.

The size of any target locus for hybridization capture is an important consideration for any enrichment strategy. Consequently, in-house protocols have been reported that enrich for a full chromosome [17] and target complete genomes [39]. Although the hybridization capture method described here was used to enrich a small nuclear locus, to be truly useful it must be scalable to targets much larger and complex than the

TLR8 gene. Furthermore, a recent study has described an ancient hybridization capture method that increases the recovery of small endogenous aDNA sequences. By optimizing a silica extraction protocol and utilizing a single strand library construction procedure, the authors of the study reported a greatly improved recovery of small (< 50 base pairs) aDNA fragments [36]. The majority of aDNA is composed of small fragments and these short sequences contain a large amount of information but these molecules are lost in most protocols for preparing sequencing libraries [9,36]. Further development of the hybridization capture method presented in the current study will involve modification to allow the enrichment of shorter aDNA fragments.

Uracil-Induced Misincorporation

Libraries constructed with USER treatment contained no or minimal evidence of uracil-induced incorporation whilst libraries that kept uracil intact show signs of miscoding lesions (Figures 6 and 7). Uracil does not completely inhibit replication by archaeal DNA polymerases such as Phusion, so it is not unexpected that the *Phusion* and *RE+Phusion* libraries contain deaminated cytosine induced misincorporation [21]. Miscoding from the TLR8 captured libraries of *Phusion* and *RE+Phusion* appear to favor the 5' and 3' end of reads respectively (Figure 7A and 7C), but this is likely a stochastic effect of the small number of sequences (Table 3B).

Conclusion

Of the enzymatic methods examined in the current study, *USER+Phusion* treatment proved to be the most effective in maximizing the number of steppe bison reads in both shotgun and TLR8 enriched libraries whilst simultaneously reducing uracil-

induced misincorporation. The USER enzyme cocktail is an inexpensive treatment and although it requires a relatively long incubation step, the process is not labor intensive [2]. The *USER+Phusion* protocol used Phusion DNA polymerase for amplification, but other DNA polymerases, such as Taq, would probably give similar results in combination with USER treatment. Because USER treatment increases endogenous data and eliminates damage-induced misincorporation, treating with the enzyme cocktail should be considered for any aDNA study, although it may be necessary to construct libraries in parallel without USER treatment so that uracil misincorporation patterns can be used to verify the authenticity of the aDNA [17].

RNA probes transcribed from a cattle DNA sequence were shown to be suitable for the enrichment of an entire nuclear gene from an extinct species of bison. This represents a useful method for enriching relatively short genomic loci, which is theoretically scalable to much larger target regions. Furthermore, it has been demonstrated that probes from cattle are capable of enriching for DNA from phylogenetically distant bovid taxa, as has recently been shown for other taxonomic groups [35].

Acknowledgments
We thank the miners of the Yukon Territory, especially the Johnson family, for assistance in collecting ancient vertebrate bones. We also acknowledge the Yukon Heritage Branch, especially Grant Zazula for field assistance.

Author Contributions
Conceived and designed experiment: SMR AC. Performed experiments: SMR. Analyzed the data: SMR BL JB. Wrote the paper: SMR. Edited the paper: AC BL JB. Provided the ancient sample: AC.

Supporting Information:
Figure S1. TLR8 RNA probe synthesis
Figure S2. Restriction enzyme recognition sites
Figure S3. Examples of qPCR amplification curves
Figure S4. TLR8 unique read coverage
Supplemental Method

References

1. Cooper A, Mourer-Chauviré C, Chambers GK, von Haeseler A, Wilson AC, et al. (1992) Independent origins of New Zealand moas and kiwis. Proceedings of the National Academy of Sciences 89: 8741-8744.
2. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.
3. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.
4. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, et al. (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499: 74-78.
5. Jackes M, Sherburne R, Lubell D, Barker C, Wayman M (2001) Destruction of microstructure in archaeological bone: A case study from Portugal. International Journal of Osteoarchaeology 11: 415-432.
6. Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 1939-1943.
7. Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, et al. (2004) Genetic analyses from ancient DNA. Annual Review of Genetics 38: 645-679.
8. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.
9. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, et al. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35: 5717-5728.
10. Hoss M, Jaruga P, Zastawny TH, Dizdaroglu M, Pääbo S (1996) DNA Damage and DNA Sequence Retrieval from Ancient Tissues. Nucleic Acids Research 24: 1304-1307.
11. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 38: 1-12.
12. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft Sequence of the Neandertal Genome. Science 328: 710-722.
13. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757-762.
14. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.
15. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
16. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, et al. (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. Nature 478: 506-510.

17. Fu QM, Meyer M, Gao X, Stenzel U, Burbano HA, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences of the United States of America 110: 2223-2227.
18. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111-118.
19. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
20. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. Biotechniques 42: 343-352.
21. Greagg MA, Fogg AM, Panayotou G, Evans SJ, Connolly BA, et al. (1999) A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. Proceedings of the National Academy of Sciences of the United States of America 96: 9045-9050.
22. Knapp M, Stiller M, Meyer M (2012) Generating barcoded libraries for multiplex high-throughput sequencing. Methods in Molecular Biology 840: 155-170.
23. Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research 22: 939-946.
24. Dabney J, Meyer M (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.
25. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200. EMBnetjournal.
26. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
27. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.
28. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153-2155.
29. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29: 1682-1684.
30. Chen Z, Riggs D (2011) DNA Methylation and Demethylation in Mammals. Journal of Biological Chemistry 286: 18347-18353.
31. Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, et al. (2012) High-Resolution Analysis of Cytosine Methylation in Ancient DNA. PLoS ONE 7: 1-6.
32. Shen JC, Rideout WM, Jones PA (1994) The Rate of Hydrolytic Deamination of 5-Methylcytosine in Double-Stranded DNA. Nucleic Acids Research 22: 972-976.
33. Carvalho AB, Clark AG (2013) Efficient identification of Y chromosome sequences in the human and Drosophila genomes. Genome Research 23: 1894-1907.

34. Skoglund P, Storå J, Götherström A, Jakobsson M (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. Journal of Archaeological Science 40: 4477-4482.
35. Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ (2013) Capturing protein-coding genes across highly divergent species. Biotechniques 54: 321-326.
36. Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, et al. (2013) Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proceedings of the National Academy of Sciences 110: 15758-15763.
37. Schuenemann VJ, Bos K, DeWitte S, Schmedes S, Jamieson J, et al. (2011) Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. Proceedings of the National Academy of Sciences 108: 746–752.
38. Casey J, Davidson N (1977) Rates of formation and thermal stabilities of RNA:DNA and DNA:DNA duplexes at high concentrations of formamide. Nucleic Acids Research 4: 1539-1552.
39. Enk JM, Devault AM, Kuch M, Murgha YE, Rouillard J-M, et al. (2014) Ancient Whole Genome Enrichment Using Baits Built from Modern DNA. Molecular Biology and Evolution 31: 1292-1294.

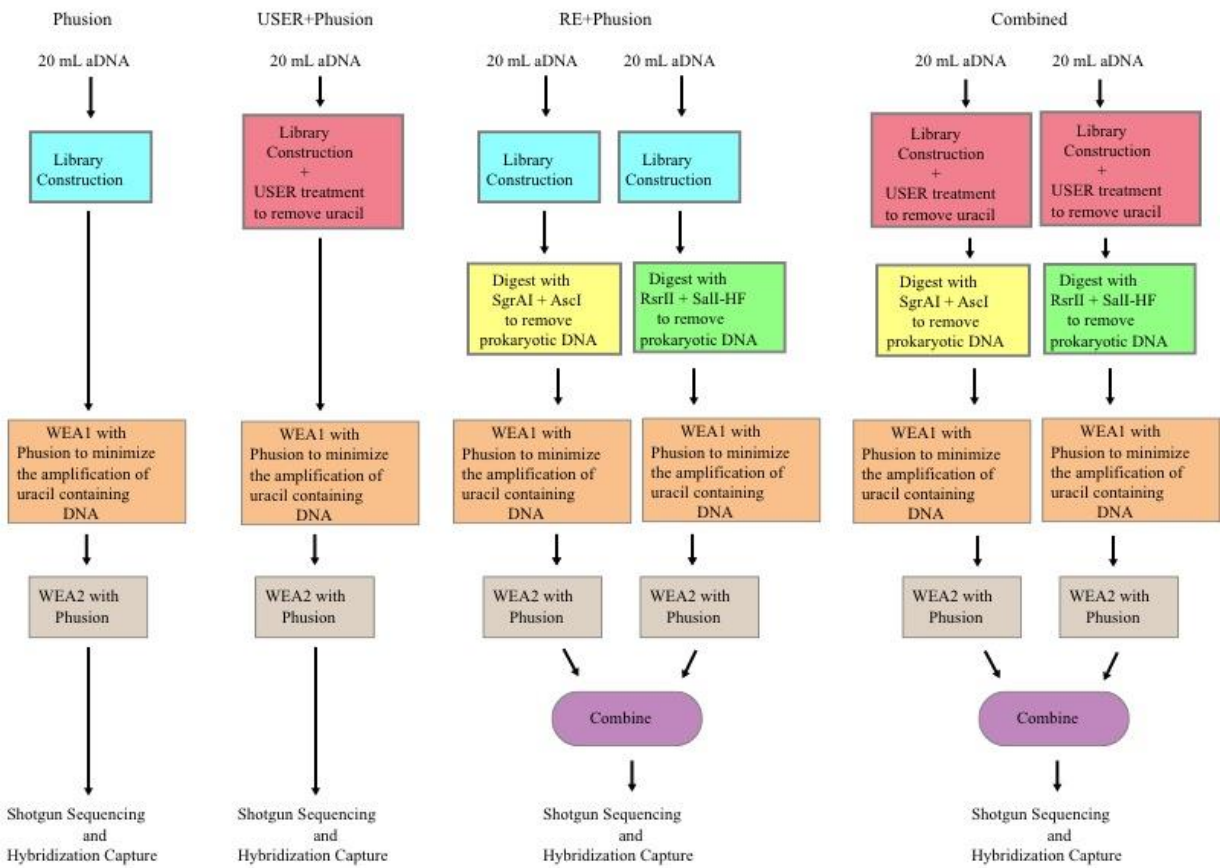


Figure 1. Library construction and amplification

Flow diagrams outlining the steps involved in the construction of the enzymatically treated steppe bison sequencing libraries. USER is an enzyme cocktail containing uracil DNA glycosylase and endonuclease VIII, which will remove uracil produced by deamination of cytosine from aDNA. SgrA1, AscI, RsrII, and SalI-HF are restriction enzymes. Phusion is a DNA polymerase that does not amplify templates containing uracil efficiently.
WEA1 =Whole Extract Amplification 1, WEA2 = Whole Extract Amplification 2

Figure 2A.

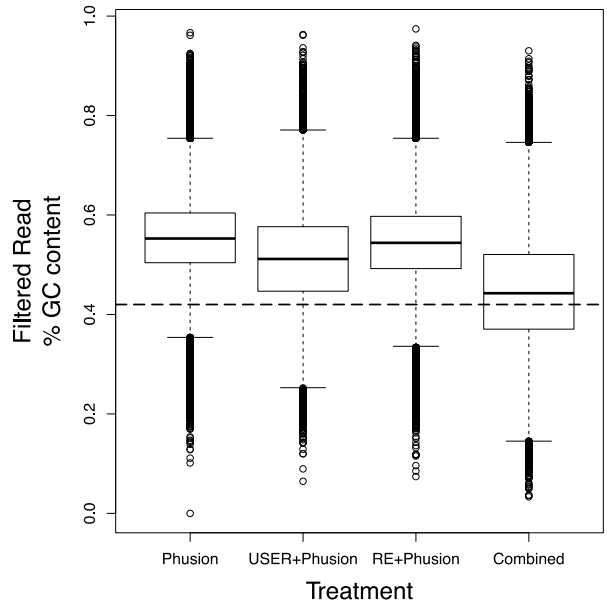


Figure 2B.

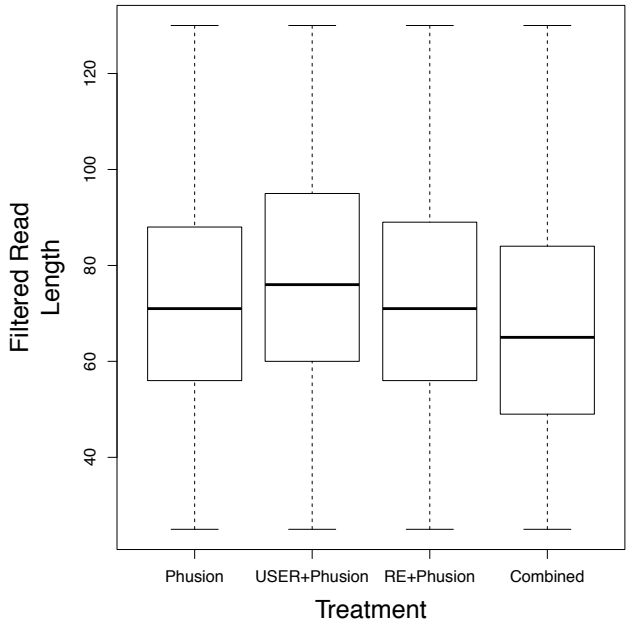


Figure 2. Boxplots for shotgun filtered read GC content and length

The R statistical package (http://CRAN.R-project.org/package=tweedie) was used to generate boxplots for the filtered read CG content and filtered read length for the enzymatically treated steppe bison shotgun libraries. The horizontal dashed line in 2A represents the 42.4% GC content of the cattle reference genome (Btau 4.0). The *Combined* treatment produced the shotgun library with the GC content closest to the reference (2A) and all the treatments produced libraries with approximately the same range of read lengths (2B).

Figure 3A.

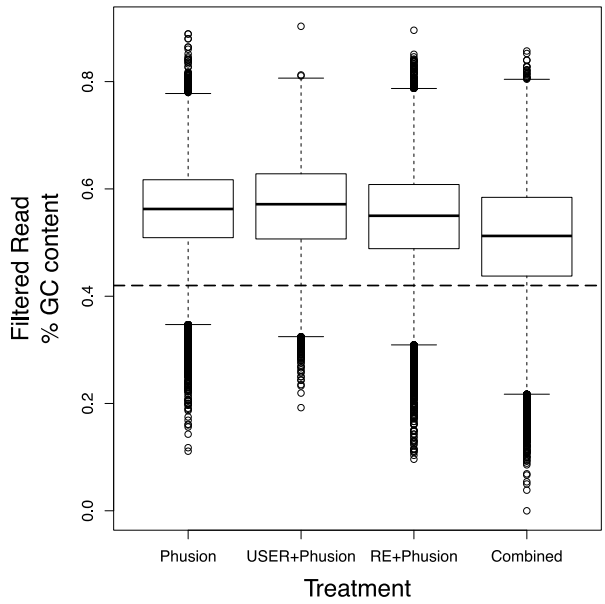


Figure 3B.

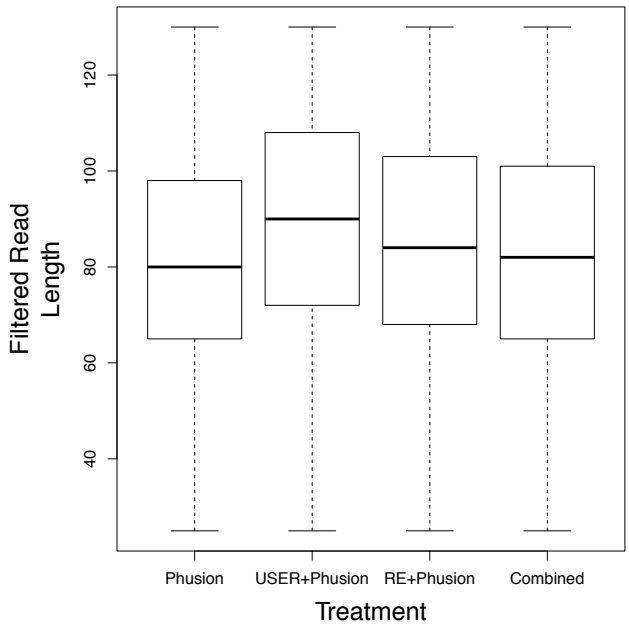


Figure 3. Boxplot for TLR8 enriched filtered read GC content and length

Filtered read GC content and filtered read length boxplots for the enzymatically treated steppe bison TLR8 enriched libraries were generated with the R statistical package (http://CRAN.R-project.org/package=tweedie). The horizontal dashed line in 3A represents the 42.1% GC content of the cattle TLR8 gene. All treatments produced enriched libraries with GC contents higher than the TLR8 reference (3A). TLR8 enriched libraries (3B) exhibited a shift towards longer reads in comparison to the shotgun libraries (Figure 2B) and this change is possibly due to larger reads having a higher enrichment efficiency in hybridization capture.

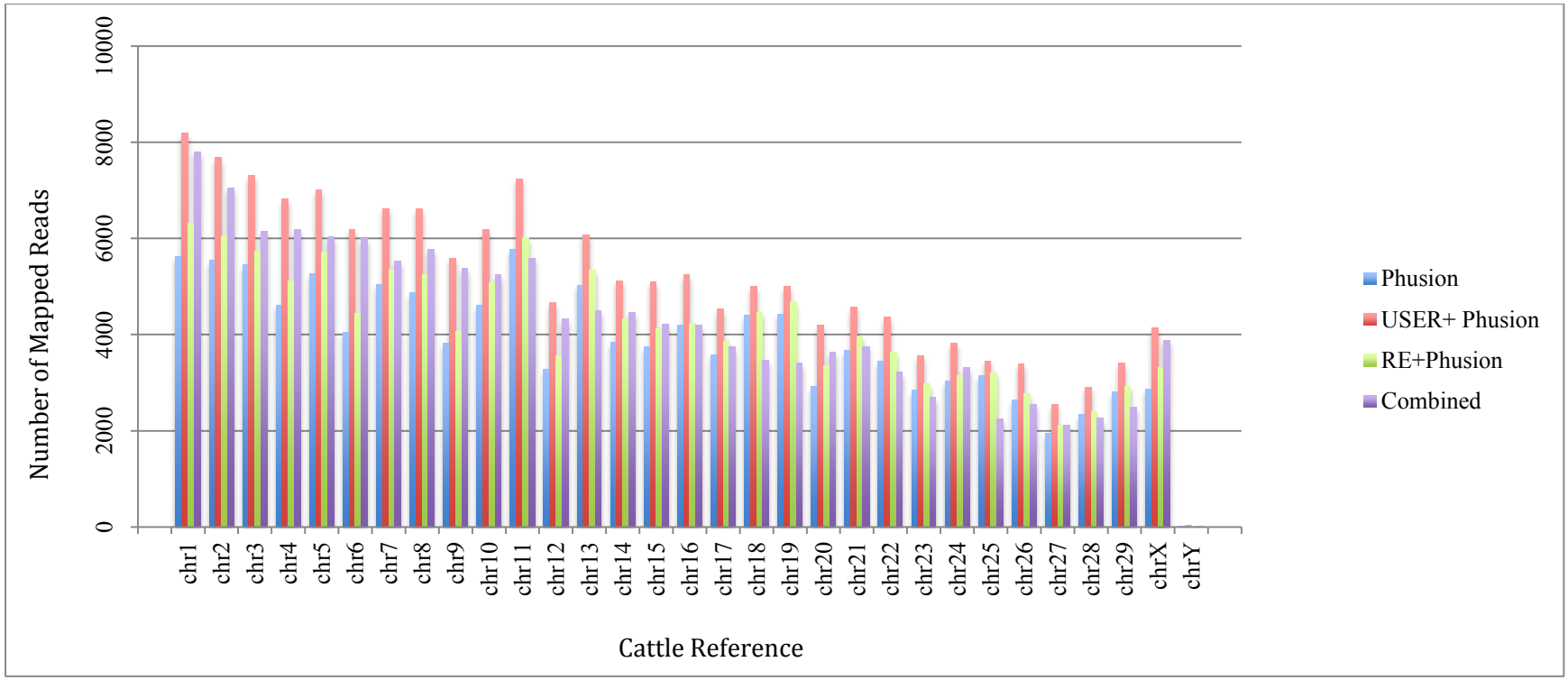
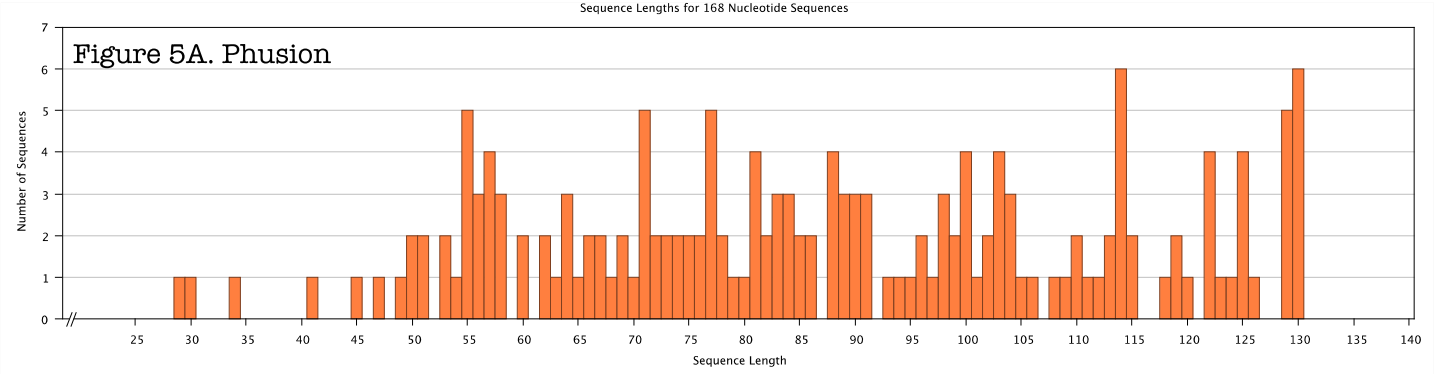
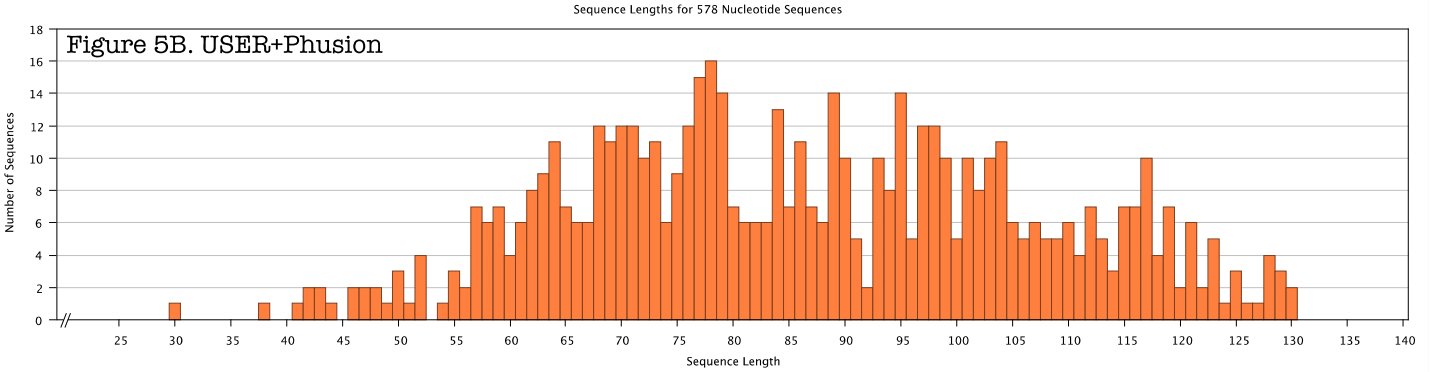


Figure 4. Chromosomal distribution of shotgun unique reads

Filtered reads from the enzymatically treated steppe bison shotgun libraries were mapped to the cattle reference (Btau 4.0) with the BWA program [26] and clonal sequences collapsed to generate unique reads. The unique reads were sorted to the individual chromosomes of the cattle reference and the number of unique reads that mapped to chromosome Y was small, ranging from 19 to 30 sequences. The *USER+Phusion* library consistently produced the highest number of unique reads that mapped to the individual chromosomes of the cattle genome.


Sequence Lengths for 168 Nucleotide Sequences
Figure 5A. Phusion
Number of Sequences
Sequence Length


Sequence Lengths for 578 Nucleotide Sequences
Figure 5B. USER+Phusion
Number of Sequences
Sequence Length

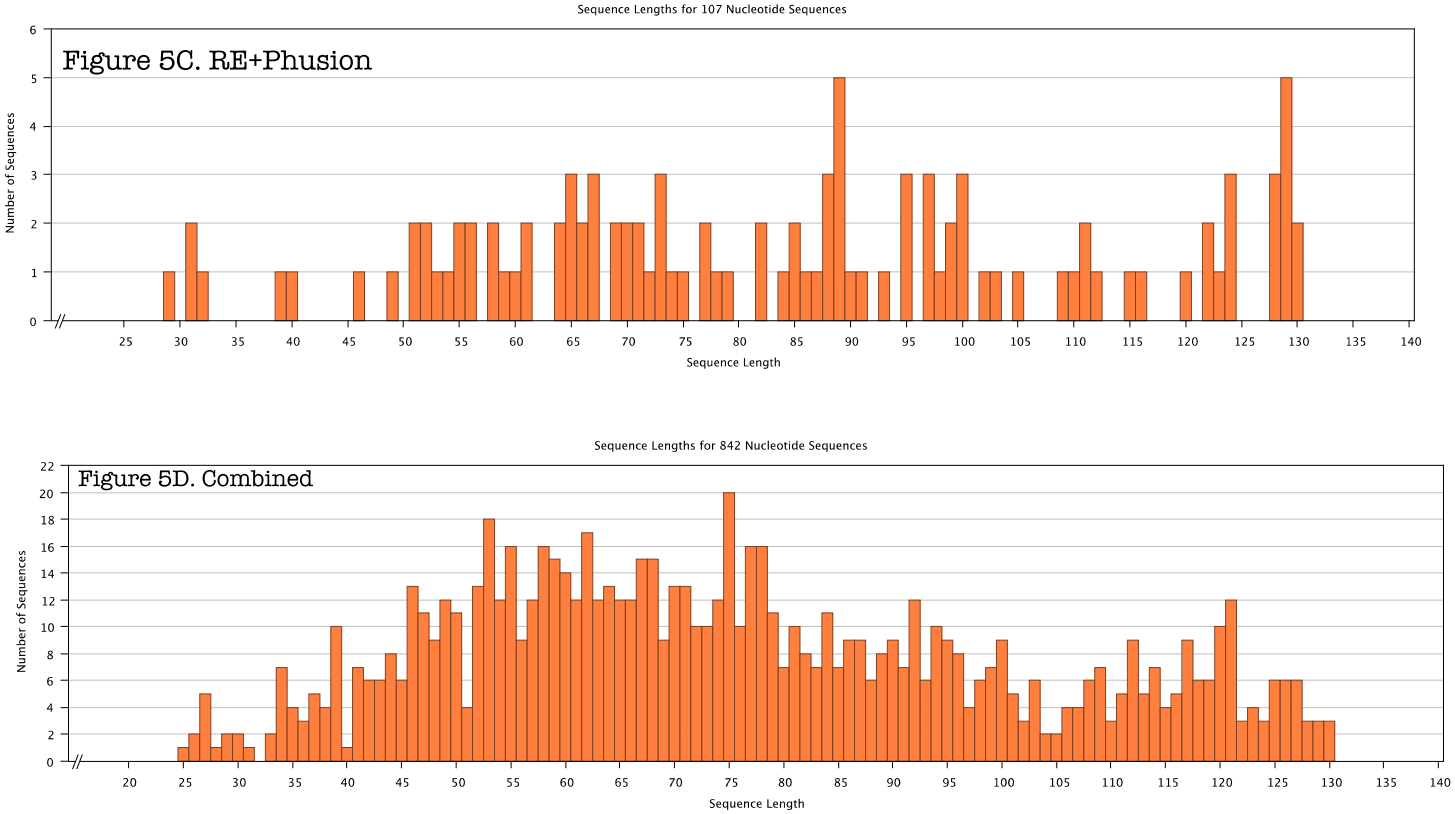


Figure 5. Length distribution for the unique TLR8 enriched reads

Length distributions for the unique reads that were mapped to the cattle TLR8 reference using the TMAP program (https://github.com/nh13/TMAP). Plots of the read length distribution were generated in the Geneious R6.1 analysis package using the TMAP data. The *USER+Phusion* (Figure 5B) and *Combined* TLR8 mapped read exhibit a roughly Gaussian distribution that has been reported previously in hybridization capture of aDNA [17].

Figure 6A.

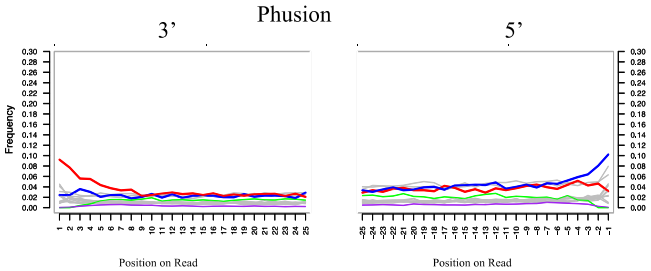


Figure 6B.

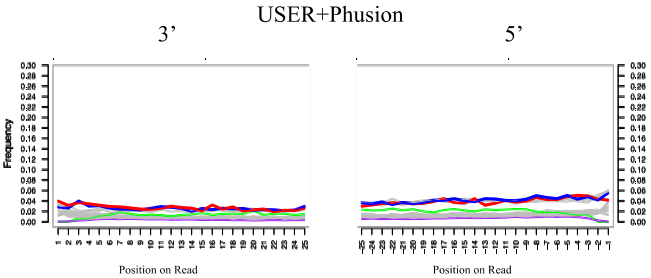


Figure 6C.

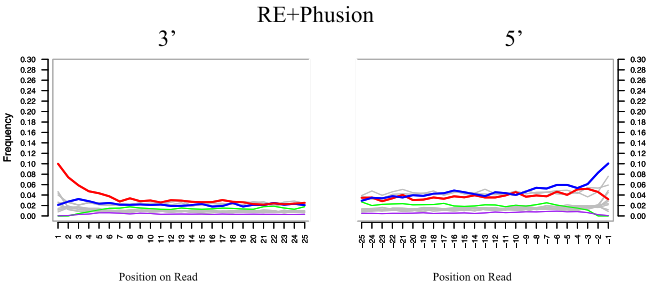


Figure 6D.

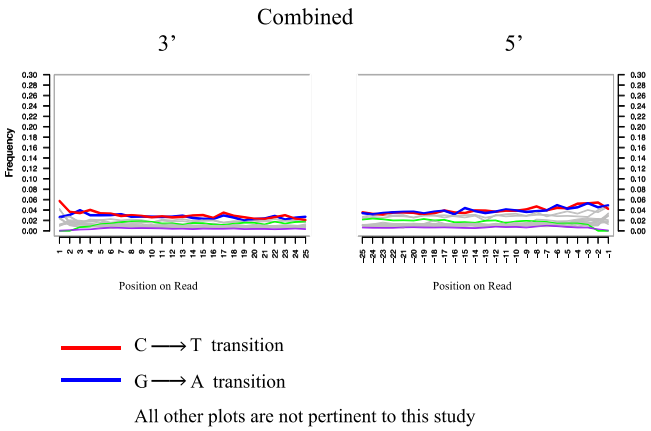


Figure 6. mapDamage plots for shotgun unique reads.
Shotgun unique reads from the enzymatically treated libraries (Table 1A) were compared to the cattle reference genome (Btau 4.0) and the frequency of 3’ C → T and 5’ G→A uracil induced transitions were plotted with mapDamage [28,29]. Deamination of cytosine to uracil primarily occurs at the ends of aDNA because these regions are single stranded and presumably more prone to chemical decomposition [11]. The *Phusion* (6A) and *RE+Phusion* (6C) libraries exhibited the typical pattern of deaminated cytosine misincorporation.

Figure 7A.

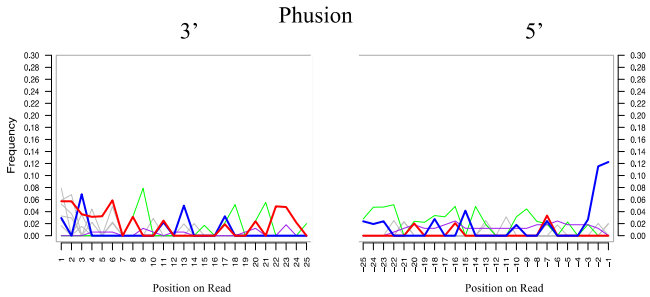


Figure 7B.

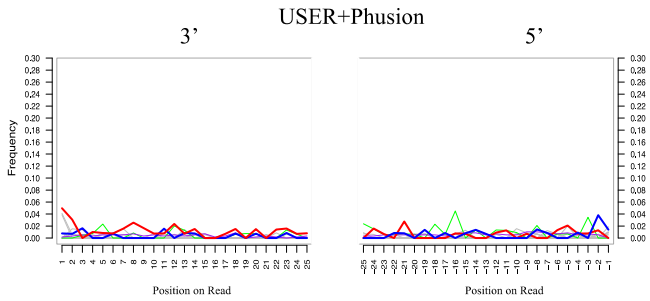


Figure 7C.

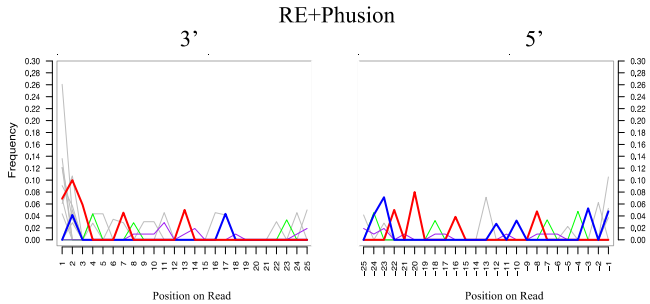


Figure 7D.

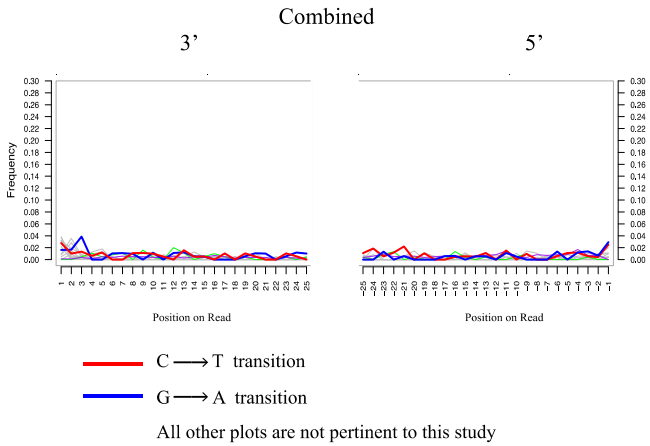


Figure 7. mapDamage plots for captured TLR8 unique reads

Unique reads from the secondary hybridization capture of TLR8 (Table 1B) were compared to the cattle reference TLR8 gene and the occurrence of 3’ C → T and 5’ G→A uracil induced misincorporations were plotted with mapDamage [28,29]. In aDNA, the majority of uracils produced from cytosine deamination are found at the ends of sequences because these regions tend to be single stranded and presumably more susceptible to chemical degradation [11]. The predominance of misincorporations in the 5’ end of the *Phusion* treatment and the 3’ end for the *RE+Phusion* treatment are possibly from stochastic effects of a small number of reads (Table 3B).

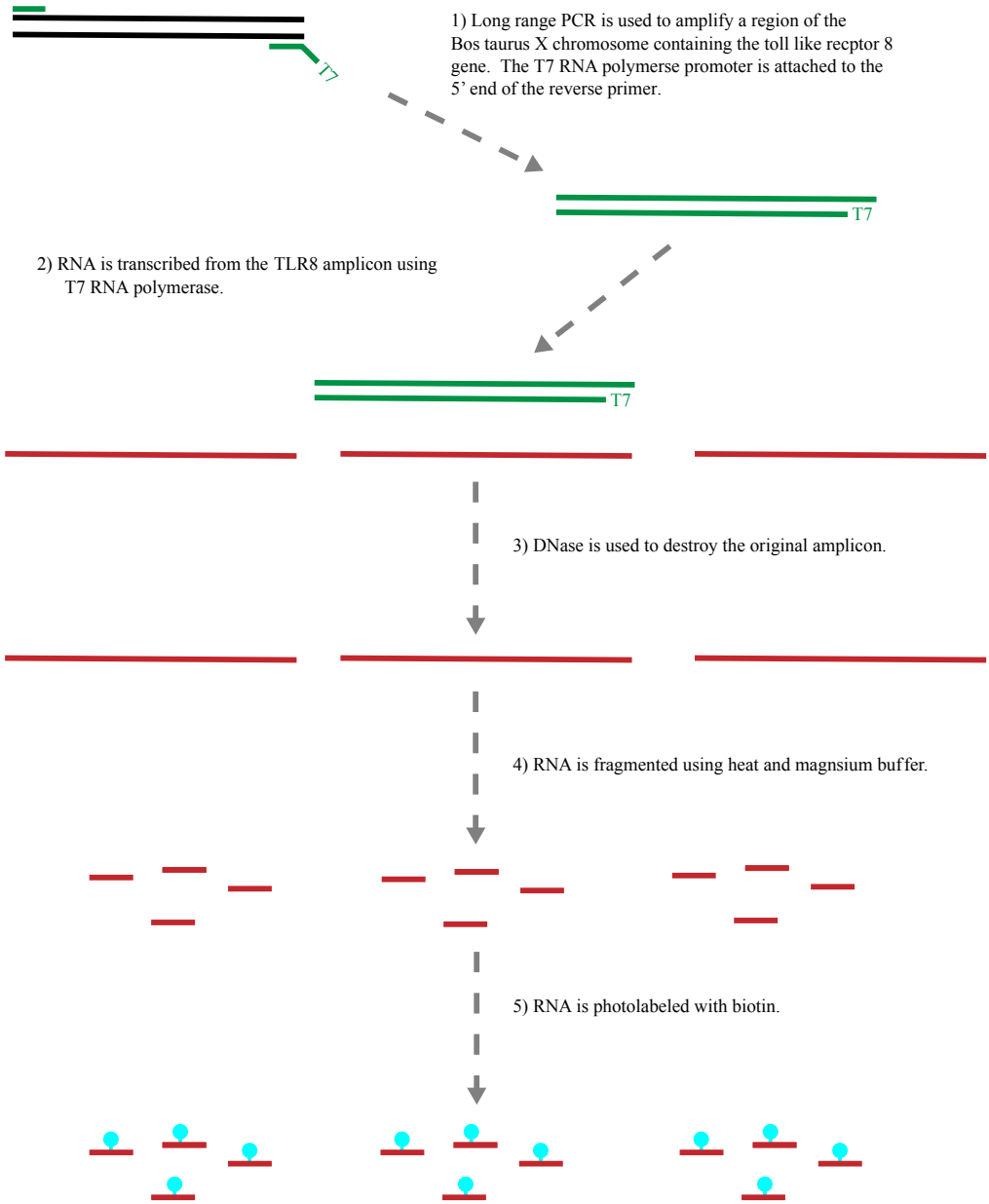


Figure S1. TLR8 RNA probe synthesis
A schematic illustrating the steps for producing the RNA probes used in hybridization capture of the TLR8 gene.

Figure S2. Restriction enzyme recognition sites
The recognition sequences for the restriction enzymes used to prepare steppe bison sequencing libraries in the *RE+Phusion* and *Combined* treatments. The arrows indicate where the restriction enzymes cut the sugar phosphate backbone of DNA.

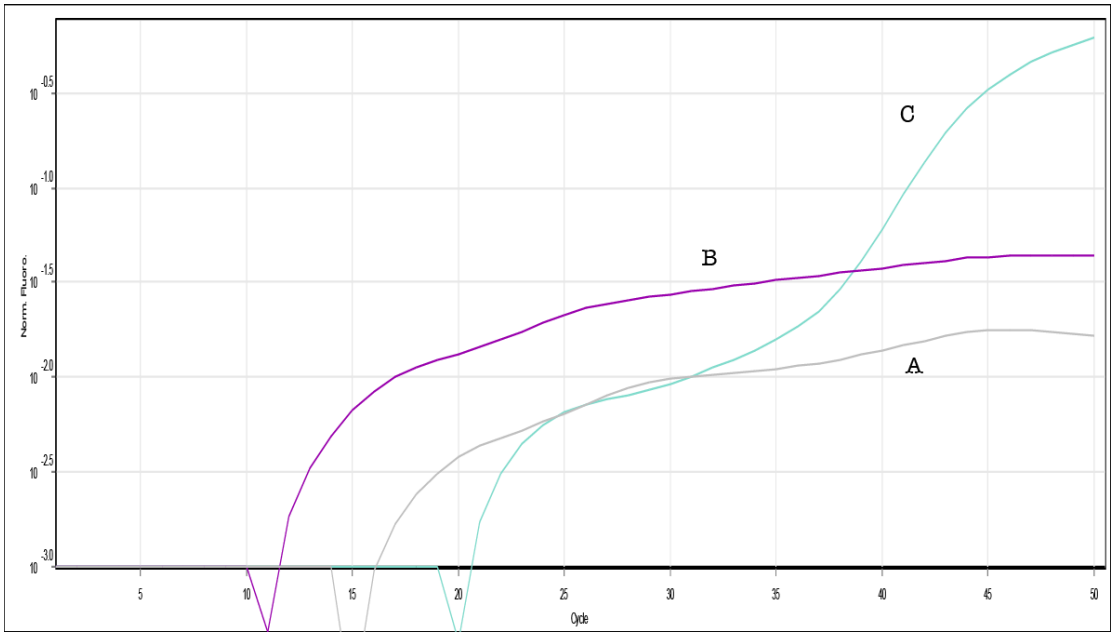


Figure S3. Examples of qPCR amplification curves

Examples of qPCR amplification curves for the TLR8 gene in steppe bison and extraction blank sequencing libraries. Bison aDNA and extraction blanks were converted into sequencing libraries using four enzymatic treatments and then taken through an initial round of PCR called Whole Extract Amplification 1 (WEA1). All library WEA1s were tested for the presence of the TLR8 gene. Extraction blank libraries produced amplification curves that did not cross a background threshold and were considered free of bison aDNA.

A – No Template Control
B – Extraction Blank *RE+Phusion* WEA1
C – Steppe bison *RE+Phusion* WEA1

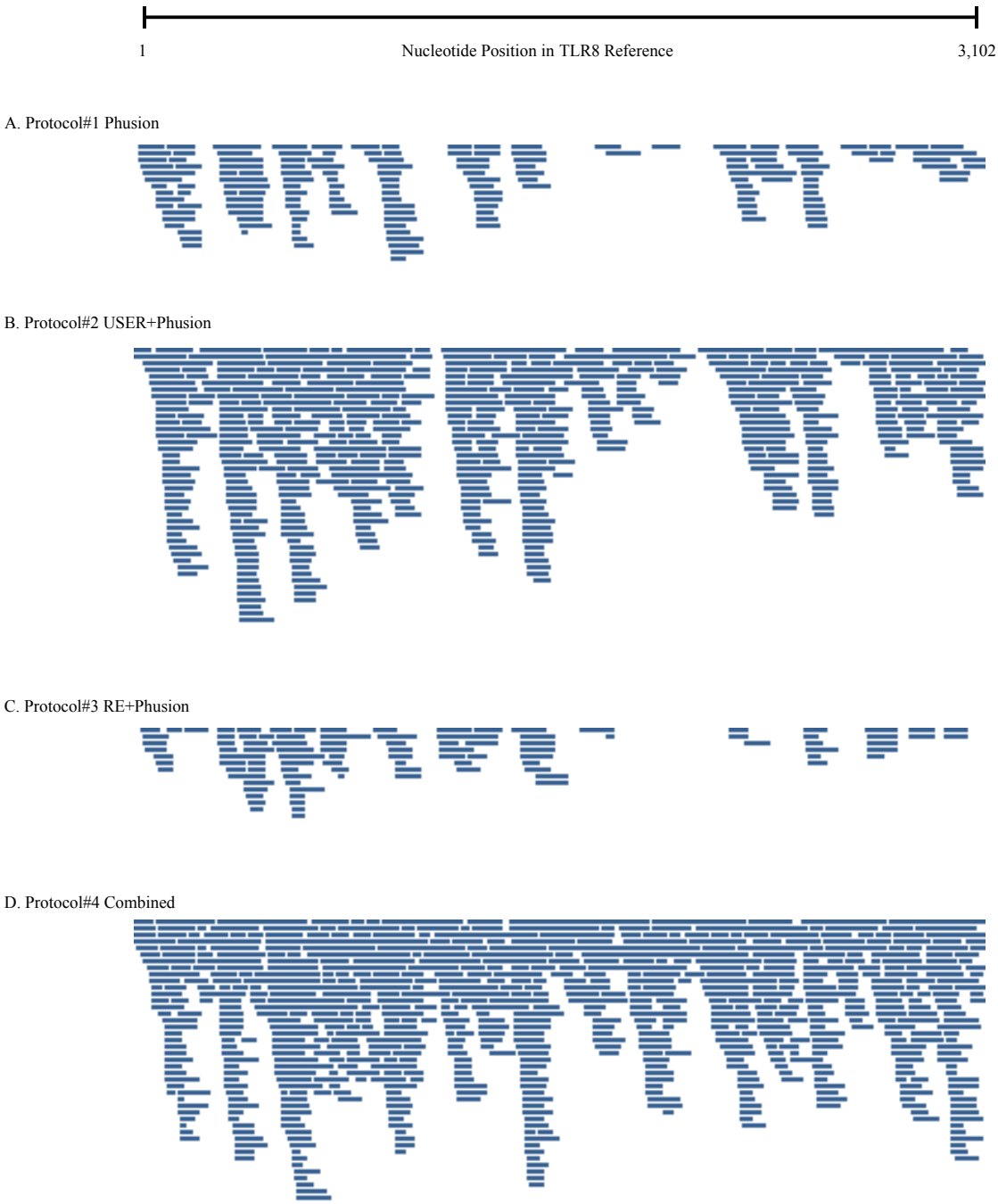


Figure S4. TLR8 unique read coverage

Visualizations of the unique read coverage from the enzymatically treated TLR8 enriched libraries generated using the graphical viewer Tablet (http://ics.hutton.ac.uk/tablet/). The detrimental effects of limiting the amplification of uracil containing templates by using Phusion DNA polymerase alone can be seen in the sparse number of reads and the gaps in coverage of the TLR8 gene exhibited by the *Phusion* (A) and *RE+Phusion* (C) libraries.

Table 1. Library characteristics

1A. ShotgunLibraries				MG-RAST Analysis	
Enzymatic Treatment	Total # Barcoded Reads	Total # Filtered Reads	% Filtered Reads	% of Annotated Reads Prokaryotic	% of Annotated Reads Bovidae
Phusion	809,828	534,884	66.0	27.1	39.7
USER+Phusion	763,812	548,257	71.8	23.7	54.2
RE+Phusion	829,115	538,170	64.9	18	59
Combined	835,629	563,172	67.4	15.5	62.5

1B. Captured Toll Like Receptor 8 Libraries (TLR8)				MG-RAST Analysis	
Enzymatic Treatment	Total # Barcoded Reads	Total # Filtered Reads	% Filtered Reads	% of Annotated Reads Prokaryotic	% of Annotated Reads Bovidae
Phusion	702,762	514,360	73.2	26.7	57.6
USER+Phusion	140,680	101,282	72.0	30.3	51.3
RE+Phusion	863,334	597,505	69.2	27.1	56.2
Combined	1,022,750	713,391	69.8	24.1	63.2

Four different enzymatic treatments (Figure 1) were examined in the construction of sequencing libraries made with steppe bison aDNA. A portion of each library was shotgun sequenced (1A) and another portion was taken through two sequential rounds of TLR8 hybridization capture (1B). All libraries were barcoded and sequenced on an Ion Torrent PGM sequencer. Reads from the PGM were binned according to barcodes and then passed through quality filters (25 to 130 base pairs in length and quality score ≥ 20). The filtered data were uploaded into the MG-RAST metagenomics analysis server, which determined the taxonomic distribution through similarity of reads to annotated predicted proteins and ribosomal RNA genes [27]. None of the enzymatic treatments appeared to have a substantial impact on library quality as all methods produced approximately the same fraction of reads passing quality filters. In the MG-RAST analysis, annotated reads identified as prokaryotic were considered to be environmental contamination and reads identified as belonging to the family Bovidae (which includes bison) were deemed to be endogenous steppe bison sequences. The enzymatic treatments did not have a clear impact on the MG-RAST analysis.

Table 2. Unique shotgun reads mapping to cattle and *Bison bison* references

Enzymatic Treatment	Cattle Genome (Btau 4.0)		*Bison bison* Mitochondrial Genome (NC_012346)	
	†Raw	‡Relative	†Raw	‡Relative
Phusion	118,797	22.20%	11	2.06×10^{-3} %
USER+Phusion	156,608	28.56%	28	5.11×10^{-3} %
RE+Phusion	127,500	23.69%	9	1.67×10^{-3} %
Combined	131,220	23.30%	49	8.70×10^{-3} %

Filtered reads from the four enzymatically treated steppe bison shotgun libraries (Table 1A) were mapped (Phred score ≥30) to the cattle (*Bos taurus*) reference genome (Btau 4.0) with the BWA program [26] and to the *Bison bison* mitochondrial genome (NC_012346.1) with TMAP software (https://github.com/nh13/TMAP). Clonal sequences were collapsed to produce the number of unique reads that mapped to each reference. Of all the enzymatic treatments, *USER+Phusion* prove to be the optimal library construction method for generating data from nuclear endogenous aDNA. The data generated for the mitochondrial genome was too small to make any evaluation of the effect of the enzyme treatments.

†Raw - The number of unique mapped reads.

‡Relative – The number of unique mapped reads as a percentage of the filtered reads for the library.

Table 3. Mapping of unique reads to the cattle TLR8 gene

3A. Filtered Reads from Shotgun Libraries						
Enzymatic Treatment	Unique Mapped Reads		Coverage of TLR8 Gene	Minimum Reads in Coverage	Maximum Reads in Coverage	Mean # of Reads in Coverage
	†Raw	‡Relative				
Phusion	0	0	0.00%	0	0	-
USER+Phusion	1	1.8×10^{-4} %	3.10%	1	1	-
RE+Phusion	0	0	0.00%	0	0	-
Combined	1	1.7×10^{-4} %	1.90%	1	1	-

3B. Filtered Reads from TLR8 Libraries						
Enzymatic Treatment	Unique Mapped Reads		Coverage of TLR8 Gene	Minimum Reads in Coverage	Maximum Reads in Coverage	Mean # of Reads in Coverage
	†Raw	‡Relative				
Phusion	168	0.03%	82.80%	0	18	4.7 ± 4.5
USER+Phusion	578	0.57%	98.90%	0	42	16.1 ± 9.7
RE+Phusion	107	0.02%	69.10%	0	14	2.9 ± 3.2
Combined	842	0.12%	100%	1	43	20.5 ± 7.9

Steppe bison filtered reads from the enzymatically treated shotgun libraries (3A) and TLR8 enriched libraries (3B) were mapped to the cattle TLR8 reference (chromosome X, AC_000187.1) using the TMAP program (https://github.com/nh13/TMAP) and clonal sequences were collapsed to generate unique reads. Unique reads were imported into the Geneious R6.1 analysis package to generate data on coverage of the TLR8 gene. In comparison to the shotgun data, the high number of TLR8 reads in the enriched libraries indicates that the hybridization capture procedure described in the current study was successful in recovering target sequences from the steppe bison libraries. The enzymatic treatments did have an impact on enrichment as the *USER+Phusion* and *Combined* treatments clearly out performed the other methods in recovery of the TLR8 gene from the steppe bison libraries.

†Raw - The number of TLR8 unique mapped reads.

‡Relative – The number of TLR8 unique mapped reads as a percentage of the filtered reads for the library.

Table S1 Primers and oligonucleotides

IS1_adapter.P5	A*C*A*C*TC TTTCCCTACACGACGCTCTTCCG*A*T*C*T
IS2_adapter.P7	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
IS3_adapter.P5+P7	A*G*A*T*CGGAA*G*A*G*C
IS7_short_amp.P5	ACACTCTTTCCCTACACGAC
IS8_short_amp.P7	GTGACTGGAGTTCAGACGTGT
Bovid_LR_TLR8_F	GGCAGAGCAGGCCAACTGTCA
Bovid_LR_TLR8_R(T7)	**AATTGTAATACGACTCACTATAGGG**TGATGGACTCGTCTCACCTCTGC
ITF_FOR_BC1	CCATCTCATCCCTGCGTGTCTCCGACTCAGtgacgtgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC2	CCATCTCATCCCTGCGTGTCTCCGACTCAGacagctgACACTCTTTCCTACACGACGCTCTTCCGATCTC
ITF_FOR_BC3	CCATCTCATCCCTGCGTGTCTCCGACTCAGagcactgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC5	CCATCTCATCCCTGCGTGTCTCCGACTCAGtctgatgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC8	CCATCTCATCCCTGCGTGTCTCCGACTCAGtacgtcgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC9	CCATCTCATCCCTGCGTGTCTCCGACTCAGtgtctcgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_REV	CCTCTCTATGGGCAGTCGGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
T7-A18B	**GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAG**(A)18[B]
B_bison_TLR8_40_F	AGCTAAGGTCAAAGGCTACAGG
B_bison_TLR8_128_R	TGACAGAAGCGTCTTTGGTG
P5_short_RNAblock	ACACUCUUUCCCUACACGAC
P7_short_RNAblock	GUGACUGGAGUUCAGACGUGU

* - Phosphorothioate bond
Bold Nucleotides – T7 RNA polymerase promoter
Lower Case Nucleotide – Barcodes
B = C or G or T

Optimizing Ancient DNA Sequencing Libraries:
Supplemental Methods

Standard precautions were taken to ensure the authenticity of the aDNA data [1]. Extraction of aDNA and library construction were performed at the Australian Centre for Ancient DNA (University of Adelaide, South Australia, Australia), in a dedicated aDNA laboratory that is cleaned routinely with bleach and is exposed to UV light after each procedure. The aDNA laboratory contained positive air pressure to minimize contamination from outside sources and access to the laboratory is limited to properly trained personnel. Control libraries were constructed from extraction blanks for each protocol and all PCRs included no template negative controls. Extraction blank libraries were tested for bison sequences with quantitative PCR (qPCR) and found to be negative. All of the no template controls for PCR were negative for product as assayed by gel electrophoresis.

0.00 Sample

aDNA was extracted from an astragalus bone of a steppe bison recovered from Irish Gulch in the Yukon Territory, Canada. The bone (ACAD sample # A3133) was carbon dated at the Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom) with accelerator mass spectrometry. The sample produced an uncalibrated age of 26,360 ± 220 radiocarbon years before present using a ^{14}C half-life of 5,568 years. No permits were required for the described study, which complied with all relevant regulations.

1.00 aDNA Extraction

A Dremel tool with a carborundum cutting disk was used to remove a section of the astragalus bone, which was subsequently powdered in a Braun mikro-dismembrator. aDNA was extracted from the bone powder using a silica based method [2,3]. To extract, 200 mg of bone powder was incubated with 4.4 mL lysis buffer (0.5M EDTA, pH 8.0; 0.5% N-lauroylsarcosine and 0.25 mg/mL proteinase K) overnight at 37 °C on a rotor. The tube was then centrifuged at 4,600 rpm for 5 minutes and the supernate transferred to a fresh 15 mL tube containing 125 μL of medium size silica particles[1] and 16 mL binding buffer (13.5 mL QG Buffer + pH indicator, 1.3% Triton x100, 20 mM NaCl, 250 mM sodium acetate, and H_2O). The extract and silica were then mixed overnight at room temperature on a rotor, after which the tube was centrifuged at 4,600 rpm for 5 minutes and the supernate discarded. The pelleted silica was suspended in 1 mL 80% ethanol, transferred to a 1.5 mL tube, and centrifuged at 15,000 g for 2 minutes. The supernate was discarded and the silica washed an additional two times with 1 mL 80% ethanol. Residual ethanol was removed by drying the silica at 37°C for 30 minutes. The aDNA was then eluted by suspending the silica in 200 μL 1x TE buffer and incubating at 50 °C for 10 minutes. The silica was then centrifuged at 15,000 g for 2 minutes and the supernate, containing the aDNA, transferred to an O-ring tube for storage at -20°C. An extraction blank using 200 μL H_2O instead of bone powder was also performed.

[1] Medium sized silica particles for aDNA extraction were prepared as follows. Six grams of silicon dioxide were suspended in 50 mL H_2O and left to settle for 1 hour. Approximately 40 mL of the supernate was then poured into a fresh 50 mL centrifuge tube and allowed to settle overnight on a bench top. The next morning, 30 mL of the supernate was discarded and the remaining 10 mL kept as the medium sized silica particles for aDNA extraction. The silica particles were kept at room temperature in the dark until needed.

2.00 Preparation of Truncated Adapter Stocks

Working stocks of short P5 and P7 adapters [4] were prepared using four oligonucleotides (Table S1) and 10x Oligo hybridization buffer (500 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA). The stocks were prepared as follows:

25 µM P5 Working Stock:	25 µM P7 Working Stock:
10 µL of 250 µM IS1_adapter.P5	10 µL of 250 µM IS2_adapter.P7
10 µL of 250 µM IS3_adapter.P5+P7	10 µL of 250 µM IS3_adapter.P5+P7
10 µL 10× Oligo hybridization buffer	10 µL 10× Oligo hybridization buffer
70 µL Molecular Biology Grade H_2O	70 µL Molecular Biology Grade H_2O

To anneal the adaptor oligonucleotides, the working stocks were heated to 95°C for 10 seconds and then cooled to 14°C at - 0.1°C/second. After annealing, stocks were aliquoted and stored at -20 °C.

3.00 Library Preparation and Enzymatic Treatment

3.01 Phusion: The aDNA was polished in a reaction that contained 20 µL bison A3133 extract, 1x NEB Buffer 2, 20U T4 polynucleotide kinase, 4.5U T4 DNA polymerase, 1mM ATP, 0.1 mM dNTPs, 8 µg rabbit serum albumin (RSA), molecular biology grade H_2O to 40 µL. The polishing reaction was incubated at 25° C for 15 minutes and then purified using a MinElute spin column following the provided PCR cleanup protocol and eluting with 20 µL EB + 0.05% Tween-20. Next, adapter ligation was performed by incubating 20 µL polished aDNA, 1 µL 25 µM P5 short adapter working stock, 1 µL 25 µM P7 short adapter working stock, 1x T4 Ligase buffer, 5% (w/v) polyethylene glycol 4000, 6U T4 DNA ligase and H_2O to 40

μL at 22°C for 1 hour. The ligation reaction was purified with a MinElute column as before. To complete the adapters a strand displacement reaction was performed in a tube containing 20 μL of ligated aDNA, 1x Thermopol buffer, 19.2U *Bst* DNA polymerase large fragment, 250 μM dNTPs and H_2O to 40 μL. The strand displacement reaction was incubated at 37°C for 10 minutes followed by heating to 80°C for 20 minutes to inactivate the *Bst* [4,5].

3.02 USER+Phusion: Removal of uracil and polishing of the aDNA was performed in a reaction that contained 20 μL bison extract, 1x NEB Buffer 2, 3U USER enzyme cocktail, 20U T4 polynucleotide kinase, 1mM ATP, 0.1 mM dNTPs, 8 μg RSA, and H_2O to 38.5 μL. The reaction was incubated at 37°C for 3 hours then 4.5U of T4 DNA polymerase was added and the reaction incubated at 25°C for a further 30 minutes. The polished aDNA was MinElute purified as before and then processed as in step *3.01* from adapter ligation onwards [6].

3.03 RE+Phusion[2]: Two library reactions were carried out as in step *3.01*, after which the libraries were digested with different mixtures of restriction enzymes. One library was digested in a reaction that contained 20 μL inactivated *Bst* reaction, 1x NEB Buffer 4, 10U SgrAI, 10U AscI, and H_2O to 50 μL whilst the other library was digested in a separate reaction containing 20 μL inactivated *Bst* reaction, 1x NEB Buffer 4, 5U RsrII, 20U SalI-HF, and H_2O to 50 μL. The digestion reactions were incubated at 37°C for 60 minutes followed by heating to 65°C for 20 minutes to

[2] To reduce the loss of bison aDNA, two digestion reactions were performed using different restriction enzyme cocktails and then combined after amplification (Figure 1). Each cocktail will cleave some eukaryotic DNA but the bison aDNA lost in one digestion reaction will be different from what is lost in the other. By combining the digestion reactions after amplification, the loss of endogenous bison sequences should be minimized.

inactivate the restriction enzymes. After inactivation, the digestion reactions were purified separately using MinElute columns as before.

3.04 Combined[2]*:* Two libraries were constructed using the uracil repair protocol outlined in step *3.02*. The repaired libraries were then taken through restriction enzyme digestion as described in step *3.03*.

4.00 WEA1

4.01 Phusion: The library was initially amplified in five PCRs each containing 5 µL inactivated *Bst* reaction, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM of each primer IS7_short_amp.P5 and IS8_short_amp.P7 (Table S1), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Amplification was performed in a heated lid thermal cycler programed as follows: 1 cycle: 98°C for 30 seconds; 14 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. After amplification, 2 µL of each PCR were gel electrophoresed and produced product smears of approximately 150 to 300 base pairs in length. PCRs were pooled and combined with 1.8 volumes of Ampure XP in a 1.5 ml low-bind tube for purification. Ampure and the pooled PCR products were mixed well and allowed to stand for 5 minutes after which the tube was placed in a magnetic rack for 3 minutes to pellet the beads. The supernate was discarded and the beads were washed three times with 800 µL 70% ethanol. After discarding the last wash, the beads were dried by leaving the uncapped tube in a magnetic rack for 10 minutes. Library was eluted by resuspending the beads in 30 µL 10 mM Tris pH 8.0 + 0.05% Tween-20 and incubating at room temperature for 5 minutes. The beads were then pelleted by placing the tube in a magnetic rack and after 3 minutes the supernate

(containing the library) was transferred to a fresh low bind tube. The library was quantified with a NanoDrop 2000 spectrophotometer and then stored at -20°C.

4.02 USER+Phusion: Library was amplified as in step *4.01*.

4.03 RE+Phusion: Both of the digested libraries were amplified as in *4.01* except DNA from the restriction enzyme treatments were used as template instead of the inactivated *Bst* reaction.

4.04 Combined: Both of the digested libraries were amplified as in *4.01* except DNA from the restriction enzyme treatments were used as template instead of the inactivated *Bst* reaction.

5.00 WEA2

5.01Phusion: Amplification was performed as in step *4.01* using 3 ng of purified WEA1 library as template instead of the inactivated *Bst* reaction.

5.02 USER+Phusion: Amplification was performed as in step *4.01* using 3 ng of purified WEA1 library as template instead of the inactivated *Bst* reaction.

5.03 RE+Phusion: For each of the digestions, amplification was performed as in step *4.01* using 3 ng of purified WEA1 library as template instead of the inactivated *Bst* reaction.

5.04 Combined: For each of the digestions, Amplification was performed as in step *4.01* using 3 ng of purified WEA1 library as template instead of the inactivated *Bst* reaction.

6.00 Extraction Blank

Extraction blanks were taken through similar library preparation protocols and WEA1s as those used for the steppe bison extract. Libraries made with bison aDNA and extraction blanks were assayed for TLR8 using quantitative PCR (qPCR) as a test for contamination. qPCRs were performed in triplicate reactions containing 3 ng WEA1, 1x Brilliant Green II SYBR Green Master Mix, 500 nM of each primer B_bison_TLR8_40_F and B_bison_TLR8_128_R (Table S1), 0.5 µg RSA, and H_2O to 10 µL. Amplifications were performed on a Roto-Gene 6000 thermal cycler running Roto-Gene v1.7 [Build 87] software and scanning on the SYBR green channel. The thermal cycling program was 1 cycle: 95°C for 5 minutes; 50 cycles: 95°C for 10 seconds, 58°C for 20 seconds, 72°C for 15 seconds followed by a melt curve that ramped from 55°C to 95°C over 10 minutes. Extraction blanks generated fluorescence curves similar to no template controls and were considered free of bison sequences. aDNA libraries produced amplification curves that rose above background after approximately 38 cycles (Figure S3).

7.00 Shotgun Libraries

7.01 Phusion: Purified library from WEA2 was used as the shotgun library.

7.02 *USER+Phusion*: Purified library from WEA2 was used as the shotgun library.

7.03 RE+Phusion: For each of the digestions, the molarity of the WEA2 was calculated using the NanoDrop measured concentrations and an average sequence size from the gel smears produced by the libraries. Library from the two WEA2s were combined in equimolar amount to form the *RE+Phusion* shotgun library. The pooled DNA from this step was also used as the WEA2 for hybridization capture.

7.04 Combined: For each of the digestions, the molarity of the WEA2 was calculated using the NanoDrop measured concentrations and an average sequence size from the gel smears produced by the libraries. Library from the two WEA2s were combined in equimolar amount to form the *Combined* shotgun library. The pooled DNA from this step was also used as the WEA2 for hybridization capture.

8.00 Synthesis of TLR8 RNA Probe

8.01 Primer Design for TLR8 In Vitro Transcription (IVT) Template: Primer-Blast software at NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) was used to design primers to amplify a 4,511 base pair segment of the cattle (*Bos taurus)* reference AC_000187.1 (X chromosome, cattle genome UMD 3.1). The amplicon spanned the entire TLR8 gene (3,102 base pairs) including several hundred nucleotides upstream and downstream of the gene. Primers were designed for two-step PCR and the T7 RNA polymerase promoter was included on the 5' end of the reverse primer allowing the amplicon to act as template for *IVT* [7-9].

8.02 Amplification of TLR8 IVT Template: Amplification of TLR8 *IVT* template was performed in PCRs containing: 1x Phire Buffer, 25 ng calf thymus DNA, 200 µM dNTPs, 500 nM of each primer: Bovid_LR_TLR8_F and Bovid_LR_TLR8_R(T7)

(Table S1), 0.5 µL Phire Hot Start II DNA polymerase, and H_2O to 25 µl. Amplifications were performed in a heated lid thermal cycler programed as follows: 1 cycle: 98°C for 30 seconds; 26 cycles: 98°C for 10 seconds and 72°C for 50 seconds; and 1 cycle: 72°C for 180 seconds. Product from 2 µL of each PCR was visualized by gel electrophoresis and UV illumination to verify amplicon size. Eighteen PCRs were pooled, purified with a QIAquick PCR Purification kit, and then quantified on a NanoDrop 2000 spectrophotometer.

8.03 Transcription of TLR8 IVT Template: TLR8 RNA was transcribed using a T7 High Yield RNA Synthesis kit in reactions containing: 150-200 ng TLR8 *IVT* template, 1x Reaction buffer, 10 mM rNTPs, 2 µL T7 enzyme mix, and H_2O to 20 µL. Reactions were incubated at 37°C for 16 hours after which the DNA template was destroyed by adding 2U of Turbo DNase and heating at 37°C for 15 minutes. Four IVT reactions were pooled and purified with a MEGAclear spin column using a modified protocol. In brief, the following were combined: 80 µL pooled *IVT* reactions, 20 µL Elution Solution, 350 µL Binding Solution, and 250 µL 100% ethanol and then centrifuged through a MEGAclear column at 15,000g for 30 seconds. The flow through was discarded and the column washed two times with 500 µL of Wash Solution with centrifugation as before. The column insert was transferred to a fresh tube and centrifuged at 15,000g for 60 seconds to remove residual liquid. The insert was again transferred to a fresh tube and 50 µL of nuclease free H_2O applied to the column. The column was heated at 65°C for 5 minutes to elute the RNA and then centrifuged at 15,000g for 60 seconds. The flow through was transferred to a fresh tube and a second elution was performed as before. The elutions were quantified on a NanoDrop 2000 spectrophotometer and stored separately at -80°C. To inspect the

integrity of the *IVT* product, 100 ng of the TLR8 RNA were visualized on an acrylamide gel. Water was used to recover the *IVT* product because the Elution buffer provided with the MEGAclear columns was found to inhibit fragmentation of RNA in the step below.

8.04 Fragmentation of TLR8 RNA: Fragmentation was performed using NEBNext Mg RNA Fragmentation Module in reactions containing: 45 µg TLR8 RNA, 1x RNA Fragmentation Buffer and H_2O to 20 µL. Reactions were heated at 94°C for 10 minutes and fragmentation was then stopped by adding 2 µL Stop Solution. Each fragmentation reaction was purified with a RNeasy MinElute spin column using the provided cleanup protocol.

8.05 Biotinylation of TLR8 RNA Fragments: Fragmented RNA was biotinylated using a Photoprobe (Long Arm) Biotin kit in 200 µL PCR tubes containing 20 µg of fragmented RNA, 40 µL of Photoprobe reagent and H_2O to 80 µL. Uncapped reactions were placed in a gel cooling rack and incubated under the UV bulb of a sterilization cabinet for 30 minutes. After the UV incubation, an organic extraction was performed on each labeling reaction by adding 64 µL H_2O, 16 µL 1 M Tris buffer, and 160 µL sec-butanol to each tube followed by vigorous shaking for 30 seconds. The phases in each tube were separated by centrifugation at 1,000 g for 1 minute after which the upper organic layers were discarded. A second round of organic extraction was performed on each labeling reaction by adding a fresh aliquot of 160 µL sec-butanol and processing as before. The second organic phases were discarded and the RNA in the remaining aqueous phases was purified with RNeasy MinElute spin columns using the provided reaction cleanup protocol. Elutions from the RNeasy

columns were pooled and quantified on a NanoDrop 2000 spectrophotometer. One hundred nanograms of the biotinylated RNA was visualized on an acrylamide gel and produced a smear in the range of 80-300 bases. The RNA was diluted to 100 ng/μL and stored as 5 μL aliquots at -80°C. Hereafter, biotinylated RNA will be called TLR8 probe.

9.00 Blocking RNA Synthesis

Repetitive sequence blocking RNA was transcribed from Bovine Hybloc DNA (Cot-1 DNA) using a previously published linear amplification protocol [10]. Briefly, Hybloc DNA was first polished with T4 polynucleotide kinase and T4 DNA polymerase and then purified with MinElute spin columns using the provided PCR cleanup protocol. Polished DNA was tailed using terminal transferase with a tailing solution containing 92 μM dTTP and 8 μM ddCTP and then MinElute purified as before. The tailed DNA was heat denatured and the T7-A18B primer (Table S1) annealed to the poly-T tail. After annealing of the primer, a second strand of the primed DNA was synthesized using DNA polymerase I Klenow fragment. The resulting double stranded product was MinElute purified as before and then transcribed using a T7 High Yield RNA Synthesis kit in reactions that contained 75 ng DNA from the second strand reaction, 1x Reaction buffer, 10 mM rNTPs, 2 μL T7 enzyme mix, and H_2O to 20 μL. Transcription reactions were incubated overnight at 37°C after which the original DNA template was destroyed by adding 2U Turbo DNase and heating at 37°C for 15 minutes. The RNA was purified with RNeasy MinElute spin columns following the provided reaction cleanup protocol and then quantified on a NanoDrop 2000 spectrophotometer. One hundred nanograms of the RNA were visualized on an acrylamide gel and produced a smear approximately 80 to

500 bases in length. Lastly the RNA was diluted to 1µg/µL in H_2O and stored at -80°C. The product transcribed in this step will be called Hybloc RNA in all further procedures.

10.00 Hybridization Capture Buffers

Hybridization Buffer:	75% formamide, 75 mM HEPES, pH 7.3, 3 mM EDTA, 0.3% SDS, and 1.2 M NaCl [11]
Wash Buffer 1:	2.0x SSC and 0.05% Tween-20
Wash Buffer 2:	0.75x SSC and 0.05% Tween-20
Wash Buffer 3:	0.20x SSC and 0.05% Tween-20
Release Buffer:	0.1 M NaOH

Neutralization Buffer: 1M Tris-HCl pH 7.5

11.00 Primary TLR8 Hybridization Capture

For the primary TLR8 hybridization capture of each library, three 200 µL Reagent Tubes were set up as follows:

Reagent Tube 1-	3.5 µL WEA2 at 35 ng/µL
Reagent Tube 2 -	5 µL TLR8 probe (Step 8), 1 µL Hybloc RNA (Step 9), and 0.5 µL of stock containing 50 µM P5_short_RNAblock and P7_short_RNAblock block (Table S1)
Reagent Tube 3-	30 µL Hybridization Buffer

The P5/P7 short_RNAblock are RNA oligonucleotides that are complementary to the library adapters and were included in the reaction to prevent reannealing of the library adapters. Hybridization capture was carried out in a heated lid thermal cycler with the following program:

Step 1- 94°C for 2 minutes

Step 2- 65°C for 3 minutes

Step 3- 42°C for 2 minutes

Hold 4- 42°C hold

To pre-warm solutions, Reagent Tubes were placed in the thermal cycler at the start of each program Step in the following order:

Step 1- Reagent Tube 1

Step 2- Reagent Tube 2

Step 3- Reagent Tube 3

Once the Hold cycle began, 20 µL of the hybridization buffer from Reagent Tube 3 was mixed with the RNA in Reagent Tube 2. The entire content of Tube 2 was then mixed with the DNA in Reagent Tube 1 to begin the hybridization capture reaction. The hybridization reaction was incubated at 42°C for 48 hours.

Just prior to the end of the hybridization reaction, magnetic streptavidin beads were washed and blocked. To wash, 50 µL beads were added to 0.5 mL Wash Buffer 1 and vortexed briefly. The beads were centrifuged, pelleted in a magnetic rack, and the supernate discarded. Beads were washed a second time as above. For blocking, beads were suspended in 0.5 ml Wash Buffer 1 + 100 µg yeast t-RNA and incubated for 30 minutes at room temperature on a rotor. Beads were subsequently pelleted in a magnetic rack, washed once with 0.5 mL Wash Buffer 1, and suspended in 0.5 mL Wash Buffer 1.

At the end of the 48 hour incubation, each hybridization reaction was combined with blocked beads and mixed on a rotor for 30 minutes at room temperature. The beads were then pelleted with a magnetic rack and the supernate discarded. The beads were then taken through a series of washes with increasing stringency as outlined below:

Wash 1 - 0.5 mL Wash Buffer 1 at room temperature for 10 minutes

Wash 2 - 0.5 mL Wash Buffer 2 at 50°C for 10 minutes

Wash 3 - 0.5 mL Wash Buffer 2 at 50°C for 10 minutes

Wash 4 - 0.5 mL Wash Buffer 3 at 50°C for 10 minutes

After discarding the last wash, captured library was liberated from probe by incubating the beads in 50 µL Release buffer for 10 minutes at room temperature and then adding 70 µL Neutralization buffer. The beads were then pelleted with a magnetic rack and the supernate, containing the captured library, was transferred to a fresh tube for purification with a modified MinElute protocol. In brief, captured library was combined with 650 µL PB buffer and 10 µL 3 M sodium acetate to adjust the pH for efficient DNA binding to the MinElute column. Purification was performed with the standard PCR cleanup protocol except the library was eluted from the spin column with 35 µL EB + 0.05% Tween-20.

12.00 Amplification of the Primary TLR8 Hybridization Capture

Libraries recovered from the primary hybridization capture were amplified as in step *4.01*, except that in each PCR 5 µL of recovered library from the primary hybridization capture was used as template instead of the inactivated *Bst* reaction. The DNA produced from this amplification was called Primary TLR8 Library.

13.00 Secondary TLR8 Hybridization Capture

The Primary TLR8 Library from each enzymatic treatment was taken through a second round of hybridization capture as done in step *11.00*, except Reagent Tube 1 contained:

Reagent Tube 1- 3.5 µL Primary TLR8 Library at 50 ng/µL

14.00 Amplification of the Secondary TLR8 Hybridization Capture

Libraries recovered from the secondary hybridization capture were amplified as in step *4.01*, except that in each PCR 5 µL of recovered library from the secondary hybridization capture was used as template instead of the inactivated *Bst* reaction. The DNA produced from this amplification was called Secondary TLR8 Library.

15.00 Conversion to Ion Torrent Libraries

Shotgun libraries (from step *7.00*) and Secondary TLR8 Libraries (from step *14.00)* were converted into Ion Torrent sequencing libraries by amplification with fusion primers. Each library was amplified with a different barcoded primer allowing for pooling with samples from other studies. For each library, the fusion primer amplification was performed in four PCRs containing 3.0 ng library, 1x Phusion HF buffer, 200 µM each of primers ITF_FOR_BC*x* and ITF_REV (Table S1), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Fusion primer amplifications were performed in a heated lid thermal cycler programed as follows: 1 cycle: 98°C for 30 seconds; 7 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. PCRs were pooled and processed as in step *4.01*.

16.00 Ion Torrent Sequencing

Molarity and size of the fusion primer libraries from this study and others were determined with an Agilent 2200 TapeStation running a High Sensitivity D1K ScreenTape. Using the TapeStation molarity and the Template Dilution Factor (Ion OneTouch System User Guide, page 97, 4472430 Rev. E), several library pools were made with each pool containing 4 libraries in equimolar amounts. Each library pool was separately emulsion amplified with an OneTouch DL System using reagents from an Ion OneTouch 200 Template Kits v2 DL and following the protocol provided in Ion OneTouch Quick Guide (publication# Man006959 Rev 5.0). The qualities of the emulsion amplified library pools were checked with a Qubit 2.0 fluorometer using an Ion Sphere Quality Control Kit. Each library pool was sequenced on an Ion 316 chip running in an Ion Torrent PGM. Sequencing was performed with reagents from an Ion PGM Sequencing 200 Kits v2 using the provided protocol (publication# Man0007273 Rev 1.0).

17.00 Data Analysis

Sequence data was obtained from the PGM in BAM format and analyzed using a bioinformatics pipeline that implemented several publicly available software packages. Demultiplexing was carried out using FastX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit), with a strict zero mismatches threshold (--bol --mismatches 0) and Cutadapt v1.2 [12] was used to trim adapter sequences using a maximum error rate of 0.33 (-e 0.3333), and to remove short (-m 25 bp), long (-M 130 bp) and low quality sequences (-q 20), with a total of five passes (-n 5). The characteristics of the filtered reads were checked with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Shotgun reads were mapped

against the *cattle* genome (Btau 4.0) using BWA [13] and to the *Bison bison* (*B. bison*) mitochondrial genome (NC_012346) using TMAP v3.2.1 (https://github.com/nh13/TMAP) with the following options (-g 3 -M 3 -n 7 -v stage1 --stage-keep-all map1 --seed-length 12 --seed-max-diff 4 stage2 map2 --z-best 5 map3 --max-seed-hits 10). TMAP does not work well with large references so BWA was needed for mapping the shotgun libraries to the cattle genome. Shotgun and enriched TLR8 reads were also mapped to the cattle TLR8 reference (AC_000187.1) using TMAP. Mapped reads with mapping quality below Phred 30 were discarded with SAMtools [14] and duplicate sequences collapsed into unique reads with the MarkDuplicates tool of Picard Tools v1.79 (http://picard.sourceforge.net). GC content of mapped reads was analyzed using the CollectGcBiasMetrics tool of Picard Tools v1.79 and misincorporation patterns were assessed using MapDamage v0.3.6 [15,16]. The resulting TLR8 unique read coverage was visualized using Biomatters Geneious Pro v R6.1 software (http://www.geneious.com/) and the cattle TLR8 reference. Mapped unique reads from the enriched libraries and the TLR8 reference were also imported into the assembly graphical viewer Tablet [17] to generate the schematics of the unique read coverage. To determine the fractions of bacterial and endogenous sequences in the libraries, the shotgun and TLR8 enriched reads were up-loaded into the MG-RAST metagenomics server and analyzed using the default settings except 'none' was selected for the sequence screening option [18]. Metagenomics analysis is performed by MG-RAST by examining the similarity of reads to known predicted proteins and ribosomal RNA genes. For this study, annotated reads categorized as prokaryotic were considered to be bacterial contamination. As there is no bison reference genome for comparison, reads categorized as belonging to the family Bovidae, which includes bison, were considered to be endogenous steppe bison

sequences. Boxplots for read GC content and length of the filtered reads were generated using the R statistical package (http://CRAN.R-project.org/package=tweedie).

Vendors for Reagents and Equipment

Hot Start Phire II DNA Polymerase	Thermo Fisher Scientific, Vic, AU
Calf Thymus DNA	Affymetrix, CA, USA
Oligonucleotides	Intergraded DNA Techn., IA,USA
dNTPs mix	New England Biolabs, MA, USA
QIAquick PCR Purification kit	Qiagen, Vic, AU
NanoDrop 2000 spectrophotometer	Thermo Fisher Scientific, Vic, AU
T7 High Yield RNA Synthesis kit	New England Biolabs, MA, USA
Turbo DNase	Life Technologies, Vic, AU
MEGAclear spin columns	Life Technologies, Vic, AU
NEBNext Mg RNA Fragm. Module	New England Biolabs, MA, USA
RNeasy MinElute spin columns	Qiagen, Vic, AU
Photoprobe (Long Arm) Biotin	Vector Lab. LTD, CA, USA
Hybloc DNA	Applied Genetics Lab., FL,USA
T4 polynucleotide kinase	New England Biolabs, MA, USA
T4 DNA polymerase	New England Biolabs, MA, USA
MinElute Spin columns	Qiagen, Vic, AU
Terminal transferase	New England Biolabs, MA, USA
dTTP	New England Biolabs, MA, USA
ddCTP	Affymetrix, CA, USA
DNA poly. I Klenow fragment	New England Biolabs, MA, USA
Dremel tool	Dremel, CA, USA
Braun micro-dismembrator	Braun, Hesse, DE
Silica	Sigma-Aldrich, NSW, AU
QG Buffer	Qiagen, Vic, AU
MinElute Spin columns	Qiagen, Vic, AU
Triton x100	Sigma-Aldrich, NSW, AU
NaCl	Sigma-Aldrich, NSW, AU
Sodium acetate	Sigma-Aldrich, NSW, AU
Ethanol	Sigma-Aldrich, NSW, AU
1M Tris-HCl pH 8.0	Life Technologies, Vic, AU
1M Tris-HCl pH 7.5	Sigma-Aldrich, NSW, AU
O.5 M EDTA pH 8.0	Life Technologies, Vic, AU
NEB Buffer 2	New England Biolabs, MA, USA
10 mM ATP	New England Biolabs, MA, USA
Tween-20	Thermo Fisher Scientific, Vic, AU
T4 DNA ligase	Thermo Fisher Scientific, Vic, AU
5% (w/v) polyethylene glycol 4000	Thermo Fisher Scientific, Vic, AU
Bst DNA polymerase large fragment	Thermo Fisher Scientific, Vic, AU
1x Thermopol buffer	Thermo Fisher Scientific, Vic, AU
Phusion Hot Start II DNA poly.	New England Biolabs, MA, USA
Ampure XP beads	Beckman Coulter, NSW, AU
Formamide	Sigma-Aldrich, NSW, AU
1M HEPES, pH 7.3	Affymetrix, CA, USA
20x SSC buffer	Sigma-Aldrich, NSW, AU
Magnetic streptavidin beads	New England Biolabs, MA, USA
Yeast t-RNA	Life Technologies, Vic, AU
USER enzyme cocktail	New England Biolabs, MA, USA

Rat serum albumin	Life Technologies, Vic, AU
QIAquick Gel Extraction kit	Qiagen, Vic, AU
2x Brilliant Green II SYBR Green	Life Technologies, Vic, AU
Roto-Gene 6000 thermal cycler	Qiagen, Vic, AU
2200 TapeStation	Agilent, Vic, AU
High Sensitivity D1K ScreenTape	Agilent, Vic, AU
OneTouch DL System	Life Technologies, Vic, AU
Ion OneTouch 200 Templ Kits v2 DL	Life Technologies, Vic, AU
Qubit 2.0 fluorometer	Life Technologies, Vic, AU
Ion Sphere Quality Control Kit	Life Technologies, Vic, AU
Ion Torrent PGM	Life Technologies, Vic, AU
Ion PGM Sequencing 200 Kits v2	Life Technologies, Vic, AU
Ion 316 chip	Life Technologies, Vic, AU
H_2O (Molecular Biology Grade)	Life Technologies, Vic, AU
SgrAI	New England Biolabs, MA, US
AscI	New England Biolabs, MA, US
RsrII	New England Biolabs, MA, US
SalI-HF	New England Biolabs, MA, US
NEB Buffer 4	New England Biolabs, MA, USA

Supplementary References

1. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
2. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. Biotechniques 42: 343-352.
3. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
4. Knapp M, Stiller M, Meyer M (2012) Generating barcoded libraries for multiplex high-throughput sequencing. Methods in Molecular Biology 840: 155-170.
5. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.
6. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 38: 1-12.
7. Ichijo T, Yamaguchi N, Tani K, Nasu M (2008) 16S rRNA sequence-based rapid and sensitive detection of aquatic bacteria by on-chip hybridization following multiplex PCR. Journal of Health Science 54: 123-128.
8. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182-189.
9. Cone RW, Schlaepfer E (1997) Improved in situ hybridization to HIV with RNA probes derived from PCR products. Journal of Histochemistry & Cytochemistry 45: 721-727.
10. Liu C, Bernstein B, Schreiber S (2005) DNA linear amplification; Hughes S, Lasken R, editors. Bloxham, Oxfordshire, United Kingdom: Scion Publishing Ltd. 77-98 p.
11. Konietzko U, Kuhl D (1998) A subtractive hybridisation method for the enrichment of moderately induced sequences. Nucleic Acids Research 26: 1359-1361.
12. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200.EMBnetjournal.
13. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
14. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
15. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153-2155.
16. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29: 1682-1684.
17. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, et al. (2013) Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14: 193-202.

18. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

Statement of Authorship

Title of Paper	Isothermal Amplification in Hybridization Capture of Ancient DNA
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards		
Contribution to the Paper	Helped conceive study design, performed all experiments, helped analyze data, wrote paper		
Signature		Date	June 18,2015

Name of Co-Author	Julien Soubrier		
Contribution to the Paper	Helped analyze data, helped edit paper		
Signature		Date	22.06.15

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Helped conceive study design, helped edit paper		
Signature		Date	24/06/2015

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Isothermal Amplification in Hybridization Capture of Ancient DNA

Stephen M. Richards*, Julien Soubrier, and Alan Cooper

Australian Centre for Ancient DNA, University of Adelaide, South Australia, Australia

*Corresponding author
E-mail: steve.richards@adelaide.edu.au

Abstract

Polymerase chain reaction (PCR) is commonly used to amplify ancient DNA (aDNA) sequencing libraries during hybridization capture procedures. However, PCR does not amplify with complete fidelity and will alter the sequence composition of a library by introducing GC content and length biases. Isothermal amplification methods are possible alternatives to PCR that may introduce less bias. In this study, two different isothermal protocols: rolling circle amplification (RCA) and recombinase polymerase amplification (RPA), were compared to PCR in the hybridization capture of a single nuclear gene from aDNA sequencing libraries. Both isothermal protocols produced similar results to PCR in terms of coverage of the target gene and identification of single nucleotide polymorphisms (SNPs), however the RPA protocol produced an enriched library with less GC bias. The other protocols produced enriched libraries with higher GC biases suggesting preferential amplification of contaminating bacterial molecules. Additionally, RPA produced an enriched library with a mean read length closest to that of the raw aDNA. These results show that isothermal amplification can be used effectively in place of PCR in hybridization capture aDNA studies.

Introduction

Hybridization capture for high-throughput sequencing (HTS) is becoming a routine procedure in ancient DNA (aDNA) research, as it allows the initially low relative concentration of target DNA in an extract to be increased substantially. This increases efficiency and cost effectiveness of sequencing. In hybridization capture, complementary oligonucleotide probes are annealed to target molecules in a sequencing library. Probes can be in solution or attached to a solid support and will immobilize target molecules upon annealing, allowing unwanted DNA to be eliminated through washing. Target molecules are then released from the probe and

processed for sequencing [1]. However, many factors can influence the success of hybridization capture, and optimization is necessary to ensure the captured library represents an unbiased sampling of the target.

Libraries made from aDNA will typically require amplification at several steps in hybridization capture procedures due to low concentration. After library construction, the concentration of aDNA is low and must be amplified in order to produce enough material for hybridization capture. As relatively small quantities of DNA are recovered following hybridization capture, ancient libraries will require amplification after enrichment. Ancient samples usually contain high levels of exogenous DNA that can make two sequential rounds of hybridization capture necessary [2-4]. Inevitably, aDNA libraries that undergo hybridization capture are taken through three or more rounds of amplification prior to sequencing, and PCR is the most common method used for these amplification steps. However, PCR does not amplify with complete fidelity and will alter the composition of a library by introducing sequence biases.

PCR biases result in certain molecules being preferentially amplified. For example, different DNA polymerases are known to influence the composition of a library in terms of template size and GC content [5]. Low template concentration can also introduce stochastic bias; if by random chance a template is not amplified in the first few cycles of PCR, then this sequence can be lost from the final amplicon population [6]. Additionally, ramp rates of thermal cyclers can alter the template composition of a library. Fast ramp rates may not provide enough time to efficiently denature high melting temperature molecules and consequently bias a library against templates with high GC content [7]. Optimizing amplification conditions can minimize biases, but

maintaining the fidelity of a complex mixture of templates, such as a sequencing library, is particularly challenging with PCR [5,8].

Isothermal amplification comprises a group of methodologies that use enzymatic activity to denature DNA instead of heat, and amplify at a constant relatively low temperature. Isothermal methods lack many of the characteristics of PCR that are known to introduce biases, and may therefore improve the fidelity of an aDNA library taken through hybridization capture. Several isothermal methodologies are currently utilized for DNA amplification [9]. For example, multiple displacement amplification (MDA) is used in many whole genome amplification protocols because of the high fidelity and large yield the method produces [10]. While MDA is known to introduce sequence biases, these biases have been shown to be less severe than for PCR [11]. However, MDA has not been applied to aDNA research because the amplification efficiency of this methodology is correlated with template length [12], with the latter being characteristically low in aDNA extracts [13].

Rolling circle amplification (RCA), a variant of the isothermal MDA, requires a circular DNA template and is dependent on the properties of phi29 DNA polymerase. In RCA, templates can be innately circular, such as plasmids [14], or be constructed *in vitro* using enzymes such as CircLigase to circularize single stranded DNA [15,16]. To initiate RCA, a single stranded circular DNA template is primed with random hexamer oligonucleotides and phi29 binds to the primer/template complexes. Phi29 then initiates DNA synthesis by extending the primers along the template and using a strong displacement activity, dislodges any downstream complementary strands encountered on the template. Dislodged strands then become sites for further DNA

synthesis, which cause a large replication structure to develop that produces exponential amplification (Figure S1). The product of RCA is a high molecular concatemer of the original circular template [14].

Another isothermal methodology, recombinase polymerase amplification (RPA), utilizes proteins involved in genetic recombination to amplify target DNA. First, recombinase enzymes form a complex with a primer that scans the template DNA for complimentary sequences. Once found, the primer is annealed to the complimentary sequence and the non-complimentary template strand is displaced. The recombinase enzymes then disassociate from the primer and a DNA polymerase with a strand displacement activity, such as *Bst* DNA polymerase, binds to the double stranded DNA formed by the primer and template. DNA single strand binding proteins attach to the displaced strand stabilizing the formation of a replication fork. The DNA polymerase then extends the primer to produce a copy of the original template. Repeated extension of two opposing primers produces exponential amplification of the target DNA (Figure S2) [17]. RPA produces a product similar to that of PCR, an amplicon constrained in size to the binding sites of the primers.

Whatever the library amplification technique used, it is important to understand the downstream effects the method will produce when applied to aDNA. All aDNA suffers from a variety of *post-mortem* damage caused through biological and chemical mechanisms [18]. Deamination of cytosine to uracil is one of the most commonly observed forms of damage in aDNA [13], which primarily occurs at the ends of ancient molecules because these regions tend to be single stranded and therefore more susceptible to chemical modification [19,20]. Deaminated cytosines are problematic

for aDNA research because many DNA polymerases will read uracil as thymine and misincorporate an adenosine on the complimentary strand, thus complicating downstream analysis.

Previous studies have successfully used isothermal amplification on degraded DNA from forensic samples and formalin fixed paraffin embedded tissues [16,21], as well as modern sequencing libraries [22-25]. However, there appears to be no study that has used isothermal methods to amplify sequencing libraries made from degraded DNA or in the hybridization capture of degraded DNA. The current study serves as an initial evaluation of isothermal amplification in the hybridization capture of aDNA through direct comparison to PCR. We compare four amplification protocols (Figure 1): two of the protocols exclusively used PCR, whilst the remaining protocols used isothermal methods in at least one of the amplification steps.

Methods

A detailed description of the methods for this study can be found in the Supplemental Methods section accompanying this paper and flow diagrams for each of the protocols are shown in Figure 1. All pre-amplification procedures were performed in a dedicated low DNA laboratory at the Australian Centre for Ancient DNA (University of Adelaide, South Australia, Australia) and strict guidelines were followed to ensure the authenticity of the aDNA results [26]. The low DNA laboratory was located in a separate building from where all DNA amplifications were performed. All workspaces were cleaned regularly with bleach and exposed to UV light after every procedure. Access to the low DNA laboratory was limited to properly trained staff.

Extraction blank controls were performed with every aDNA library construction and all amplifications included no template controls. After construction, all libraries were taken to a post-amplification laboratory for all further procedures. Extraction blank libraries were found to be negative for bison sequences by quantitative PCR (qPCR) and all no template controls were negative for bands in gel electrophoresis.

Amplification Protocols

Four protocols were compared in the hybridization capture of the toll-like receptor 8 (TLR8) gene from sequencing libraries made with the same steppe bison (*Bison priscus*) specimen (Figure 1). The first PCR protocol performed all amplifications with Phusion, a DNA polymerase that does not efficiently amplify templates containing uracil and consequently, has been used in previous aDNA studies to reduce the amount of deaminated cytosine induced misincorporation in sequencing data [27,28]. The second PCR protocol used Phusion DNA polymerase but treated the aDNA with the enzyme cocktail USER to remove uracil prior to amplification. USER is a mixture of uracil DNA glycosylase and DNA endonuclease VIII, which will act on aDNA to remove any uracils and effectively eliminate misincorporation caused by deaminated cytosine [29]. In the remaining two protocols, RCA and RPA were used to reduce or eliminate the number of PCR steps in the hybridization capture procedure.

The isothermal methods used in this study make use of different proteins, including DNA polymerases to amplify DNA. Phi29 the DNA polymerase used for RCA is known to read through uracil and misincorporate incorrect bases [30]. A TwistDx kit was used for RPA and at the time the current study was started, the identity of the provided DNA polymerase was proprietary and the ability of the enzyme to read

through uracil was unknown (Personal Communication: Matthew Forrest, TwistDx Limited).

Sample

The sample used in this study was a steppe bison astragalus bone collected from a Late Pleistocene permafrost deposit in the Canadian Yukon Territory. The bone (ACAD reference number A3133) was carbon dated using accelerator mass spectrometry at the Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom). Using a ^{14}C half-life of 5,568 years, the astragalus bone produced an uncalibrated date of 26,360 ± 220 radiocarbon years before present. No permits were required for the described study, which complied with all relevant regulations.

Extraction of aDNA

A section of the astragalus bone was removed using a Dremel tool with a carborundum cutting disk and then powdered in a Braun mikro-dismembrator. aDNA was extracted from 200 mg of bone powder using a previously established silica slurry method that resulted in 200 μL of aDNA in 1x TE buffer [31,32]. Extracted aDNA was stored at -20°C until needed.

Library Construction

Initially, aDNA for protocols 1, 2, and 4 was converted into an Illumina sequencing library using truncated adapters [33], because complete adapters can reduce enrichment efficiency in hybridization capture [34]. To produce a library stock with a concentration sufficient for enrichment, the libraries from **Protocol#1 Phusion** and

Protocol#2 USER+Phusion were taken through two sequential rounds of amplifications called whole extract amplification 1 (WEA1) and whole extract amplification 2 (WEA2). For **Protocol#4 RPA**, the first amplification, WEA1, produced a library stock with an adequate concentration for hybridization capture and no further amplification was needed. These whole extract amplifications were performed with universal primers that annealed to the library adapters.

The first step in **Protocol#3 RCA** was circularization of the aDNA with the enzyme CircLigase II. The circular aDNA was amplified with RCA and this isothermal amplification was called WEA1. The high molecular weight RCA product was fragmented and sequences 100 to 200 base pairs in length were size-selected for conversion into a truncated Illumina library (as above). Only a small amount of DNA was recovered from these procedures and the library had to be amplified using PCR with universal adapter primers to generate the concentration needed for hybridization capture; this PCR amplification was called WEA2. In protocol 3, RCA was only used for the initial amplification out of concern that repeated circularization and fragmentation would render the steppe bison sequences unrecognizable to capture probe and mapping software.

Generation of TLR8 Probe

TLR8 was chosen for this evaluation of isothermal technologies primarily because it is a small nuclear gene that could be easily captured and sequenced. TLR8 is a member of the toll-like receptor (TLR) family of transmembrane proteins, which are part of the innate immune system involved in recognizing generic pathogen molecules.

TLR8 in particular binds viral and bacterial RNA [35]. In domesticated cattle (*Bos taurus*), TLR8 is 3,102 base pairs in length and is located on the X chromosome [36].

Probes for enrichment of the TLR8 gene were produced using an in house protocol. The TLR8 gene was amplified from calf DNA using long-range PCR and a primer pair that contained the T7 RNA polymerase promoter attached to the 5' end of the reverse primer. T7 RNA polymerase was used to transcribe RNA from the long-range amplicon and the RNA was subsequently fragmented and biotin labeled to produce TLR8 probes.

Hybridization Capture of TLR8

Libraries from protocols 1, 2, and 3 were taken through two rounds of sequential TLR8 hybridization capture. After the first enrichment, the recovered libraries were amplified with universal adapter primers and PCR to produce material for the second TLR8 enrichment. Library from **Protocol# 4 RPA** was also taken through two rounds of TLR8 hybridization capture. However, since RPA can be easily substituted for PCR, the isothermal method was used for all amplifications. For all protocols, library recovered from the second enrichment was amplified with fusion primers to convert the molecules from a truncated Illumina format to full-length Ion Torrent libraries.

Ion Torrent Sequencing

Ion Torrent sequencing libraries were quantified on an Agilent 2200 TapeStation and pooled in equimolar amounts with samples from other studies. Library pools were emulsion amplified with an Ion OneTouch DL system using an Ion OneTouch 200 Template Kit v2 DL. Emulsion amplified libraries were sequenced on an Ion Torrent

PGM with Ion PGM 200 Sequencing kits and Ion 316 chips. Each library was sequenced with a quarter of the capacity of a 316 chip.

Data Analysis

A pipeline that implemented several publically available software packages was used to analyze the sequence data generated in this study. FastX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit) was used to demultiplex the raw data from the Ion Torrent PGM according to library specific barcodes. Cutadapt was then used to trim adapter sequences and apply quality filters to retain sequences that contained a minimum quality score of ≥ 20 and were 25 to 130 base pairs in length [37]. The fractions of steppe bison and prokaryotic sequences in the filtered reads from each library were determined via metagenomics analysis using the MG-RAST server [38]. Since there is no bison reference genome for metagenomics analysis, reads that were classified as Bovidae (the family which includes bison) by MG-RAST were considered to be steppe bison sequences. Filtered reads were mapped to the cattle TLR8 reference using TMAP (https://github.com/nh13/TMAP) and clonal sequences were collapsed to unique reads with Picard Tools (http://picard.sourceforge.net). Unique read coverage from each protocol was compared to the cattle TLR8 gene using Geneious 6.1.2 to determine the single nucleotide polymorphism (SNP) profile of each enriched library. Unique mapped reads were also analyzed for patterns of deaminated cytosine induce misincorporation using the mapDamage software [39,40].

Results

Whole Extract Amplification 1 (WEA1)

DNA yields for WEA1 (the initial library amplification) were examined, as this was the only step where the different amplification methods could be directly compared. This is not an ideal comparison because the protocols used varying amounts of aDNA as template, which were at different stages of library processing (Table 1, Supplemental Methods). Regardless, all protocols used small amounts of aDNA for this initial amplification. In WEA1, **Protocol#1 Phusion** and **Protocol#2 USER+Phusion** produced the lowest yields generating 0.137 and 0.434 μg of product. In contrast, the isothermal protocols generated multiple micrograms of product. **Protocol#3 RCA** had the highest amplicon yield with 26.3 μg of product and **Protocol#4 RPA**, had the second highest yield with 2.08 μg of library. However, the RCA product is a high molecular weight concatemer that cannot be used directly in hybridization capture (Figure S3).

Characteristics of Filtered Reads from TLR8 Hybridization Capture

Approximately the same relative number of reads passed quality filters after processing with Cutadapt (25 to130 base pairs in length and quality score ≥ 20) for all amplification protocols. Filtered reads ranged from 66% to 72% of the raw sequences (Table 2A). The similarity of these values suggests that the different protocols did not have a profound effect on library quality.

The cattle TLR8 reference gene has a 42.1% GC content, which provides a standard to which the enriched steppe bison libraries can be compared. In boxplot distributions

(Figure 2A), **Protocol #4 RPA** produced a GC content inter-quartile range (IQR: a value that represents 50% of the library data) of 37.9% to 50.7%. All of the other protocols produced GC content IQR values that were higher than the cattle reference ranging from 49% to 64%.

A previous study demonstrated that unamplified aDNA is generally <150 base pairs long with the majority of the molecules being < 50 base pairs in length [13]. **Protocol #4 RPA** produced filtered reads with the smallest length IQR at 42 to 62 base pairs a range that was the closest in size to what had been observed for the majority of unamplified aDNA molecules. The other protocols produced read length IQR ranging from 60 to 111 base pairs (Figure 2B).

To investigate the level of bacterial sequences in each library, the filtered reads from each protocol were analyzed with the MG-RAST metagenomics server. The filtered reads were uploaded into MG-RAST and analyzed using the default parameters with the exception that sequence screening was disabled. MG-RAST performs metagenomics analysis by using comparisons to annotated predicted proteins and ribosomal RNA genes to identify sequences [38]. For the metagenomics analysis, reads identified as prokaryotes were considered to be bacterial contamination and because there is no reference genome for bison, reads identified as belonging to the family Bovidae, which includes bison, were considered to be endogenous steppe bison aDNA. In the MG-RAST analysis, the **Protocol #4 RPA** library contained identified reads that were 12.6% prokaryotic and 61.3% bovid, which was the lowest bacterial and highest endogenous sequence content. The other libraries contained

identified reads that were between 27.7% to 28.7% prokaryotic and 42.7% to 53.5% bovid (Table 2A).

Mapping of Filtered TLR8 Enriched Reads

After sequencing, the number of filtered reads assigned to each captured library that mapped to the TLR8 locus with high confidence (mapping score of Phred ≥ 30) varied from 2,801 to 43,180. When clonal sequences were collapsed the range was from 203 to 1,608 unique mapped reads (Table 2B). In the enriched libraries, the **Protocol#4 RPA** library contained the highest percentage of filtered reads that were TLR8 unique mapped reads at 0.27%. The other treatments produced unique reads that were 0.04% to 0.19% of the filtered reads. The lowest coverage was produced by **Protocol #1 Phusion,** where unique reads covered only 86% of the TLR8 reference with at least 1x coverage. Unique reads from the other protocols completely covered the TLR8 reference, with the least coverage ranging from 1x to 3x and the highest between 63x and 173x (Table 2B and Figure 3).

Length distributions for the unique reads that mapped to the TLR8 reference were generated using Geneious. **Protocol #1 Phusion** produced a sparse distribution with no discernable pattern (Figure 4A). **Protocols #2 USER+Phusion** and **Protocol #3 RCA** produced similar read distributions that were weighted towards larger sequences and with modest peaks at approximately 75 base pairs (Figure 4B and 4C). In contrast, **Protocol #4 RPA** produced a roughly Gaussian distribution with read length centered at approximately 55 base pairs (Figure 4D).

Using the cattle TLR8 gene as a reference, the SNP profiles of the mapped read pileups were called using Geneious (Table 3) with a minimum 5x coverage used to call a SNP [41]. **Protocols #2 USER+Phusion, #3 RCA, and #4 RPA** all led to the identification of the same 14 SNPs in comparison to the TLR8 reference, whereas **Protocol #1 Phusion** could only call four of these variations due to low coverage.

Nucleotide Misincorporation

The mapDamage software was used to compare the unique read data from each library to the cattle TLR8 reference and plot the frequency of 5' C→T and 3' G→A misincorporations (Figure 5). The **Protocol#1 Phusion** library exhibited both 5' C→T and 3' G→A increases typical of uracil-induced nucleotide misincorporation (Figure 5A). As expected, **Protocol #2 USER+Phusion** produced an enriched library that did not exhibit a significant uracil-induced misincorporation pattern (Figure 5B). Despite the ability of phi29 to read through uracil and misincorporate, **Protocol #3 RCA** produced a library that does not have evidence of unique reads with incorrect bases (Figure 5C). In contrast, **Protocol #4 RPA** produced a captured library with a strong pattern of uracil-induced misincorporation (Figure 5D).

Discussion

Library Yields for WEA1

Overall, protocols that solely used PCR for amplification produced the least amount of product in WEA1, with the impact of uracil residues blocking amplification by the Phusion DNA polymerase clearly evident in the low yield of **Protocol#1 Phusion**. Removal of uracil with USER treatment prior to amplification produced a yield three

times higher (**Protocol #2 USER+Phusion**), presumably because more templates could be amplified (Table 1). Isothermal methods consistently produced higher product yields for WEA1 in this study. **Protocol #3 RCA** had the highest yield of all protocols, but this yield is not readily comparable to what the other methods produced, as the RCA amplicon is a high molecular weight concatemer. The RCA amplicon had to be converted through fragmentation and size-selection to a usable form for hybridization capture, which resulted in a large loss of DNA. Approximately 100 ng of DNA were recovered from 10 μg of RCA product taken through fragmentation and size-selection. This is in contrast with the other protocols, which produced a version of an Illumina library that could be used directly in hybridization capture. **Protocol #4 RPA** produced the second highest yield of library for WEA1 and this yield was sufficient such that no further amplification was necessary for the first hybridization capture of TLR8. All the other protocols required an additional amplification (WEA2) to produce a library stock with sufficient concentration for hybridization capture. In the comparisons of yields for WEA1, **Protocol #4 RPA** used less inactivated *Bst* reaction as template than the two PCR based protocols yet produced a considerably higher yield (Table 1). The reduced number of amplification steps needed in the RPA protocol may have contributed to the lower GC and size biases observed in this library. The high yields for RCA and RPA amplification of aDNA were not unexpected as isothermal amplification has previously been reported to generate micrograms of product using as little as 10 ng of degraded DNA isolated from formalin-fixed, paraffin-embedded tissues [42].

Characteristics of Filtered Reads from TLR8 Hybridization Capture

In this study, all of the protocols that employed PCR for any amplification step produced enriched libraries with mean GC contents that were at least 10% higher than the cattle TLR8 gene (Figure 2). The higher GC content for the filtered reads of **Protocols #1 Phusion**, **#2 USER+Phusion,** and **Protocol #3 RCA** (which had one fewer PCR amplification), suggests that these libraries have been biased towards bacterial sequences [5]. **Protocol #4 RPA** produced filtered reads with a mean GC content of 44.2%, close to the 42.1% of the reference TLR8 gene implying that this library has minimal bias towards bacterial sequences (Figure 2A). A previous study reported that the TwistDx kit amplified a modern sequencing library with a high level of GC fidelity [23], which corroborates the low GC content observed for **Protocol #4 RPA** filtered reads.

In MG-RAST analysis, **Protocol#4 RPA** produced the lowest fraction of filtered reads identified as prokaryotic, which is in accordance with the low GC content found in this library. The filtered reads of the **Protocol#4 RPA** library also contained the highest fraction of reads identified as bovid and the largest relative fraction of unique TLR8 mapped reads (Figure 2A and Table 2). Taken together, these results suggest that RPA amplified with minimal bias towards bacterial sequences. For the other protocols the data from the MG-RAST analysis were not as conclusive. Protocols #1, #2, and #3 all produced filtered reads with a comparable GC content (Figure 2A) and contained a similar fraction of reads identified as prokaryotic (Table 2), implying these libraries were biased to a similar degree towards bacterial sequences. Conversely, in these libraries there was no correlation between the numbers of TLR8 mapped reads and the fraction of reads identified as bovid. It is not clear why the

metagenomics analysis produced results that did not correlate well with the mapped data. Results from programs such as MG-RAST and TMAP represent small subpopulations of the total number of reads in a library and a larger dataset may be required to observe any relationship.

In hybridization capture of aDNA in particular, there are two opposing factors affecting length bias. First, there is PCR, which preferentially amplifies smaller amplicons [6] and second, there is hybridization capture, which is more efficient in enriching longer sequences [31]. A previous study that used a primer extension technology and an extract from the same steppe bison bone used in the current study determined that the vast majority of unamplified endogenous aDNA in this sample was < 50 base pairs in length [13]. The RPA protocol in the current study produced a mean length of 53.2 base pairs for the filtered reads after two rounds of hybridization capture suggesting that little size bias was introduced into this library. In contrast, the other protocols produced filtered reads with mean lengths that were > 24 base pairs larger than what was found in the RPA library (Figure 2B). The minimal size bias of **Protocol #4 RPA** is also apparent in the smaller size distributions of the TLR8 unique mapped reads (Figure 4). In the case of **Protocol #3 RCA** the size bias of filtered reads is likely not relevant because sequence length will be heavily influenced by the size-selection step in this protocol (Figure 1). However, given that Protocols #1, #2, and #4 used the same hybridization capture method it is not clear why only the filtered reads and mapped reads from **Protocol #1 Phusion** and **#2 USER+Phusion** are biased upwards in size (Figures 2B and 4). It is possible that RPA may produce an even stronger size bias for shorter sequences than PCR and is therefore constraining the read length of **Protocol #4 RPA** even after hybridization capture.

Mapped TLR8 Unique Reads

For both isothermal protocols the higher yields from the library amplifications allowed more library to be added to the hybridization capture steps (Supplemental Methods). However, this did not appear to affect the coverage of the TLR8 gene in comparison to **Protocol#2 USER+Phusion** (Table 2B and Figure 3).

Protocol#4 RPA produced the highest fraction of filtered reads that were unique mapped reads, which represents a 42% increase in comparison to **Protocol #2 USER+Phusion**, the best performing PCR protocol (Table 2B). **Protocol #3 RCA** produced a similar relative fraction of unique mapped reads as **Protocol #2 USER+Phusion**. **Protocol #1 Phusion** produced the smallest percentage of unique reads and lowest coverage of the TLR8 reference (Table 2B and Figure 3), suggesting that limiting the amplification of damage aDNA templates by using Phusion DNA polymerase alone was detrimental to the hybridization capture of TLR8. Removal of uracil prior to amplification (**Protocol#2 USER+Phusion**) allowed the hybridization capture of the entire TLR8 gene. While additional sequencing of the enriched library produced with **Protocol #1 Phusion** might have produced enough unique reads to cover the TLR8 gene, USER treatment is much more cost effective, which in this study, resulted in a 4.75 fold increase in the relative number of TLR8 reads (Table 2B).

In comparison to the cattle TLR8 gene, the isothermal methods and **Protocol#2 USER+Phusion** identified the same 14 SNPs in the steppe bison aDNA. This is an important observation because it demonstrates that at least with a small nuclear target, isothermal methods have a similar accuracy to PCR for identifying informative

variations in aDNA. This gives strong support that isothermal methods are viable alternatives to PCR in hybridization capture of aDNA.

Nucleotide Misincorporation

Deaminated cytosine induced misincorporation was examined to determine how the isothermal methods dealt with uracil. In most sequencing library protocols, there is a repair step that removes 3' overhangs and fills in 5' overhangs. Consequently, damage in single strand 3' ends of aDNA is lost during library construction whilst miscoding lesions will be retained in the 5' end and will cause misincorporation during subsequent PCR. Depending on what strand of the amplicon is sequenced, a deaminated cytosine in the original aDNA will produce either 5' C $\rightarrow$ T or 3' G $\rightarrow$ A transitions in HTS data [13].

Despite being constructed with Phusion, which stalls on uracil, the mapped reads produced with **Protocol#1 Phusion** exhibited 5' C $\rightarrow$ T and 3' G $\rightarrow$ A transitions typical of deaminated cytosine induced miscoding (Figure 5A). This is consistent with the observation that uracil reduces but does not eliminate amplification by thermophilic archeal DNA polymerases such as Phusion [39,43]. While **Protocol#3 RCA** used phi29, which reads through uracil, for the initial amplification step, the captured library lacks the typical misincorporation patterns of cytosine deamination (Figure 5C). However, this absence is likely caused by the circularization of the aDNA. Cytosine deamination primarily occurs at the ends of aDNA fragments and therefore will occur next to circularization junctions after treatment with CircLigase. After fragmentation of the RCA product, sequences containing circularization junctions will not be captured efficiently by probe and any reads that are captured will

not map well to the reference sequence. **Protocol#4 RPA** exhibited the typical pattern for uracil-induced misincorporation, demonstrating the DNA polymerase provided in the TwistDx kit can read through and misincorporate when encountering a uracil during DNA synthesis (Figure 5D).

Future Directions

RCA and RPA are known to introduce biases into libraries made with modern DNA, but a more extensive study will be needed to determine how these biases will effect hybridization capture of aDNA. Amplification of low concentrations of genomic DNA with MDA (which uses phi29 and random hexamer primers similar to RCA) produced libraries with allelic dropout and locus copy number errors [44-46]. However, it is not clear how these biases will transfer to RCA of aDNA. It has been hypothesized that at least some of the biases produced by MDA are caused by secondary structures in genomic DNA blocking replication by phi29 [47,48] and small circular aDNA templates may not provide enough sequence for these blocking structures to form. RPA is a very new methodology and consequently little information is available on the biases this technology may produce. One study has found that RPA introduced a relatively high number of clonal and chimeric sequences during amplification of a modern sequencing library [23].

With modern DNA, RCA is currently used to amplify circular templates such as viral and chloroplast genomes [49,50]. Additionally, CircLigase and RCA have shown encouraging results in the amplification of short tandem repeat profiles from degraded forensic samples [21]. The high fidelity and large amplification yields produced with RCA makes the methodology enticing for aDNA research. Many large-scale

hybridization capture technologies require up to 20 µg of DNA, a mass difficult to produce with aDNA and PCR without introducing biases [5]. For aDNA studies, RCA may be easily adapted for genotyping where the intent is to call SNPs. However, if the aim is to generate data such as exome sequences from aDNA then specific analysis pipelines will need to be developed to cope with the concatemer data produced by RCA. Future experiments will also have to enrich for larger fragments than those selected in this study (100-200 base pairs) and use HTS methods that produce longer reads, so that data is generated from a circular template multiple times. Consensus sequences could then be generated from the multiple reads in a concatemer and the RCA library analyzed for damage and clonal sequences before mapping to a reference.

Conclusion

This study was performed as an initial investigation of isothermal amplification in hybridization capture of aDNA, with the intent of avoiding biases introduced by PCR. Both RCA and RPA produced results that were comparable to PCR in coverage of the TLR8 gene and SNP identification. Importantly, RPA reduced the number of amplification steps and improved library fidelity by amplifying with less bias towards GC rich sequences and producing read lengths more typical of unamplified aDNA. In contrast, the only clear advantage of RCA over PCR was in the yield of the initial amplification, by producing nearly two orders of magnitude more DNA after WEA1. However, the additional processing the RCA product required prior to hybridization capture mitigated the advantage of this large yield. In the current study, we have demonstrated that isothermal amplification can be used effectively in place of PCR in hybridization capture of aDNA. However, further investigations of library fidelity with complex genomic targets are required to fully characterize the advantages that

these technologies may provide in comparison to PCR. Evaluation of hybridization capture of multiple heterozygous loci will be needed so that statistical tests on biases such as allelic dropout can be performed.

Acknowledgements:
The authors would like to thank Grant Zazula and the Yukon Heritage Branch for their assistance in our fieldwork. The authors would also like to acknowledge the miners of the Yukon Territory, including the Johnson family, for their help in the collection of ancient vertebrate bones.

Author Contributions:
Conceived and designed experiments: SMR AC. Performed experiments: SMR. Analyzed data: SMR JS. Wrote paper SMR. Edited paper: JS AC. Provided ancient sample: AC.

Supporting Material:

References

1. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111-118.
2. Fu QM, Meyer M, Gao X, Stenzel U, Burbano HA, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences of the United States of America 110: 2223-2227.
3. Handt O, Höss M, Krings M, Pääbo S (1994) Ancient DNA: Methodological challenges. Cellular and Molecular Life Sciences 50: 524-529.
4. Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ (2013) Capturing protein-coding genes across highly divergent species. Biotechniques 54: 321-326.
5. Dabney J, Meyer M (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.
6. Walsh PS, Erlich HA, Higuchi R (1992) Preferential PCR amplification of alleles: mechanisms and solutions. Genome Research 1: 241-250.
7. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. (2012) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12: 1-14.
8. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nature Methods 6: 291-295.
9. Gill P, Ghaemi A (2008) Nucleic acid isothermal amplification technologies - A review. Nucleosides Nucleotides & Nucleic Acids 27: 224-243.
10. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 99: 5261-5266.
11. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7: 1-21.
12. Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, et al. (2003) Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. Genome Research 13: 294-307.
13. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, et al. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35: 5717-5728.
14. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Research 11: 1095-1099.
15. Gadkar VJ, Filion M (2011) A novel method to perform genomic walks using a combination of single strand DNA circularization and rolling circle amplification. Journal of Microbiological Methods 87: 38-43.

16. Wang G, Maher E, Brennan C, Chin L, Leo C, et al. (2004) DNA amplification method tolerant to sample degradation. Genome Research 14: 2357-2366.
17. Piepenburg O, Williams CH, Stemple DL, Armes NA (2006) DNA Detection Using Recombination Proteins. PLoS Biology 4: e204.
18. Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 1939-1943.
19. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 38: 1-12.
20. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.
21. Tate CM, Nuñez AN, Goldstein CA, Gomes I, Robertson JM, et al. (2011) Evaluation of circular DNA substrates for whole genome amplification prior to forensic analysis. Forensic Science International: Genetics 6: 185-190.
22. Jasmine F, Ahsan H, Andrulis IL, John EM, Chang-Claude J, et al. (2008) Whole-genome amplification enables accurate genotyping for microarray-based high-density single nucleotide polymorphism array. Cancer Epidemiology, Biomarkers & Prevention 17: 3499-3508.
23. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, et al. Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13: 12.
24. Ma Z, Lee RW, Li B, Kenney P, Wang Y, et al. (2013) Isothermal amplification method for next-generation sequencing. Proceedings of the National Academy of Sciences 110: 14320-14323.
25. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, et al. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proceedings of the National Academy of Sciences 110: 19872-19877.
26. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
27. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757-762.
28. Edwards CJ, Magee DA, Park SDE, McGettigan PA, Lohan AJ, et al. (2010) A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (*Bos primigenius*). PLoS One 5: 1-13.
29. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.
30. Serrano-Heras G, Bravo A, Salas M (2008) Phage phi 29 protein p56 prevents viral DNA replication impairment caused by uracil excision activity of uracil-DNA glycosylase. Proceedings of the National Academy of Sciences of the United States of America 105: 19044-19049.
31. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
32. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. Biotechniques 42: 343-352.
33. Knapp M, Stiller M, Meyer M (2012) Generating barcoded libraries for multiplex high-throughput sequencing. Methods in Molecular Biology 840: 155-170.

34. Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research 22: 939-946.
35. Cervantes JL, Weinerman B, Basole C, Salazar JC (2012) TLR8: the forgotten relative revindicated. Cellular & Molecular Immunology 9: 434-438.
36. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, et al. (2009) A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biology 10: R42.
37. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200. EMBnetjournal.
38. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.
39. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153-2155.
40. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics 29: 1682-1684.
41. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443-451.
42. Aviel-Ronen S, Qi Zhu C, Coe B, Liu N, Watson S, et al. (2006) Large fragment Bst DNA polymerase for whole genome amplification of DNA from formalin-fixed paraffin-embedded tissues. BMC Genomics 7: 312.
43. Greagg MA, Fogg AM, Panayotou G, Evans SJ, Connolly BA, et al. (1999) A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. Proceedings of the National Academy of Sciences of the United States of America 96: 9045-9050.
44. Ballantyne KN, van Oorschot RA, Muharam I, van Daal A, John Mitchell R (2007) Decreasing amplification bias associated with multiple displacement amplification and short tandem repeat genotyping. Analytical Biochemistry 368: 222-229.
45. Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvänen AC (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. Nucleic Acids Research 31: 1-9.
46. Arriola E, Lambros MB, Jones C, Dexter T, Mackay A, et al. (2007) Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridisation. Laboratory Investigation 87: 75-83.
47. Murthy V, Meijer WJJ, Blanco L, Salas M (1998) DNA polymerase template switching at specific sites on the phi 29 genome causes the in vivo accumulation of subgenomic phi 29 DNA molecules. Molecular Microbiology 29: 787-798.
48. Panelli S, Damiani G, Espen L, Sgaramella V (2005) Ligation overcomes terminal underrepresentation in multiple displacement amplification of linear DNA. Biotechniques 39: 174-180.
49. Johne R, Muller H, Rector A, van Ranst M, Stevens H (2009) Rolling-circle amplification of viral DNA genomes using phi29 polymerase. Trends in Microbiology 17: 205-211.

50. Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, et al. (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 6: 22.

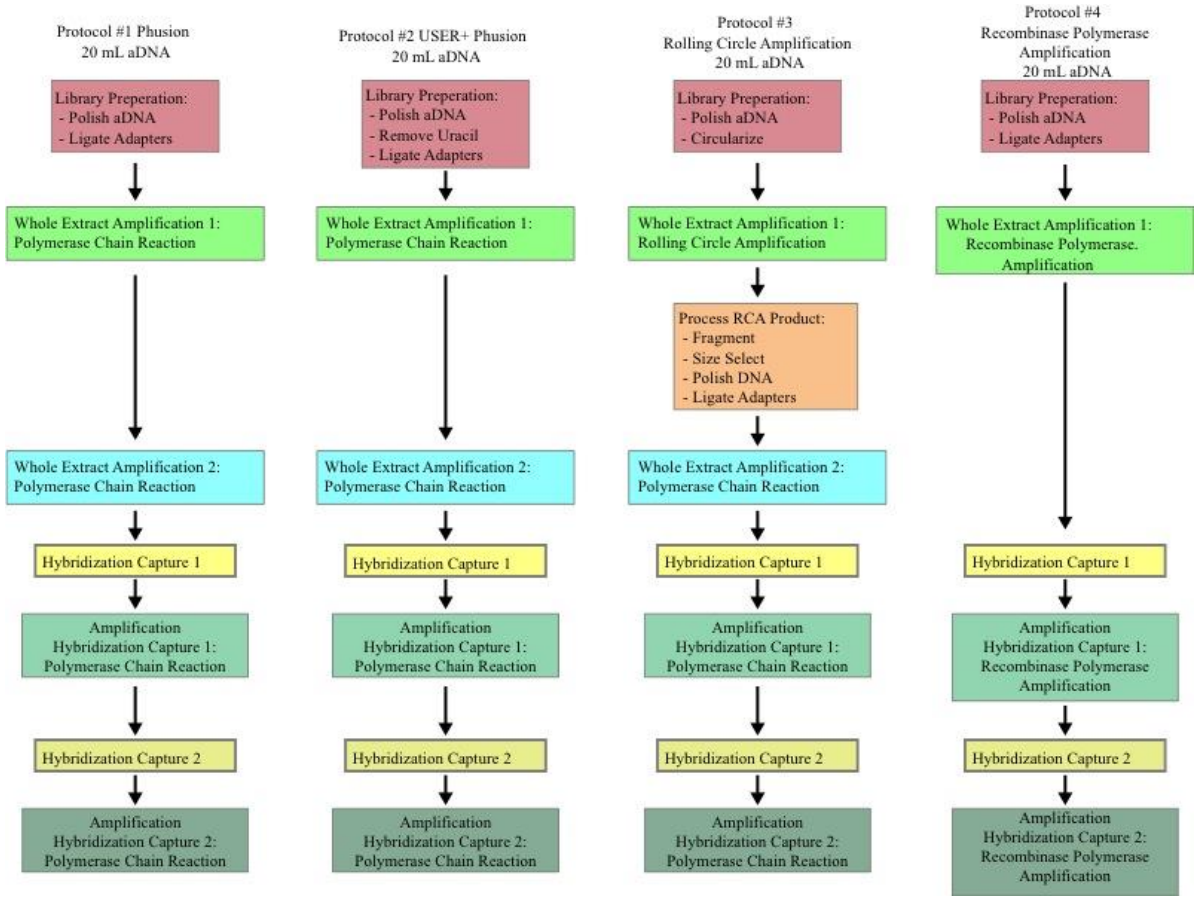
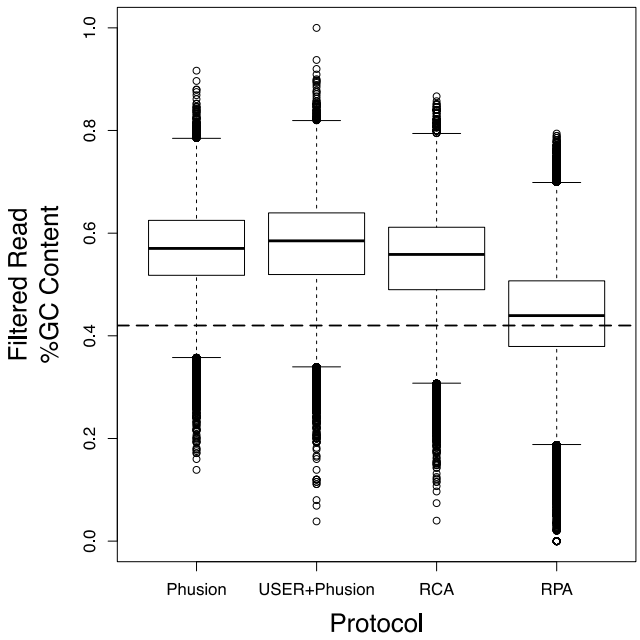


Figure 1. Flow diagrams for the hybridization capture protocols
In this study, isothermal amplification was examined as an alternative to PCR in the hybridization capture of aDNA. PCR is known to introduce a variety of biases into sequencing libraries including changes in GC content and sequence length [5] and isothermal methods may produce libraries with fewer biases. Steppe bison aDNA was converted into sequencing libraries and taken through two sequential rounds of hybridization capture for the toll-like receptor 8 gene using PCR and isothermal protocols. Illustrated are the steps in the two PCR (#1 and #2) and two isothermal (#3 and #4) protocols investigated.

2A.



2B.

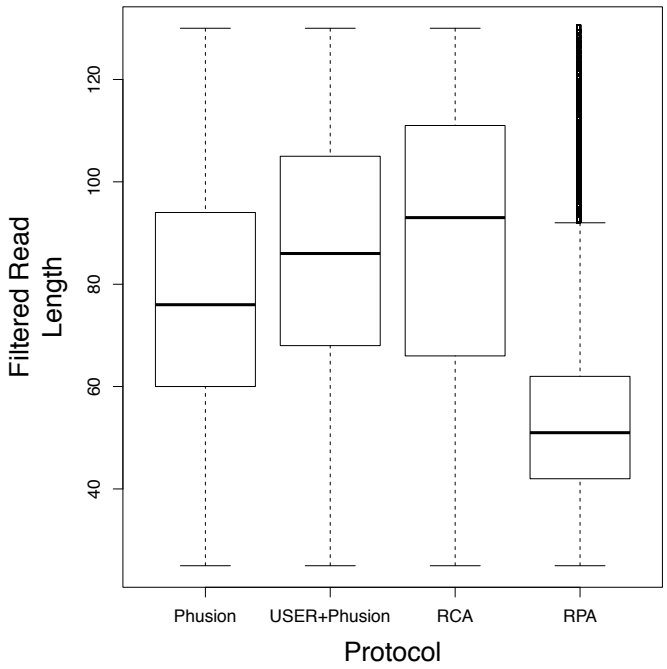


Figure 2. Boxplots of filtered read GC content and read length distribution
Boxplots for filtered read GC content (2A) and read length (2B) were generated using the R statistical package (http://CRAN.R- project.org/package=tweedie). The dotted line in Figure 2A represents the 42.1% GC content of the cattle TLR8 reference. **Protocol#4 RPA** produced filtered reads with a mean GC content that was the closest to the reference TLR8 gene and read lengths nearest to what has been observed in unamplified aDNA [13].

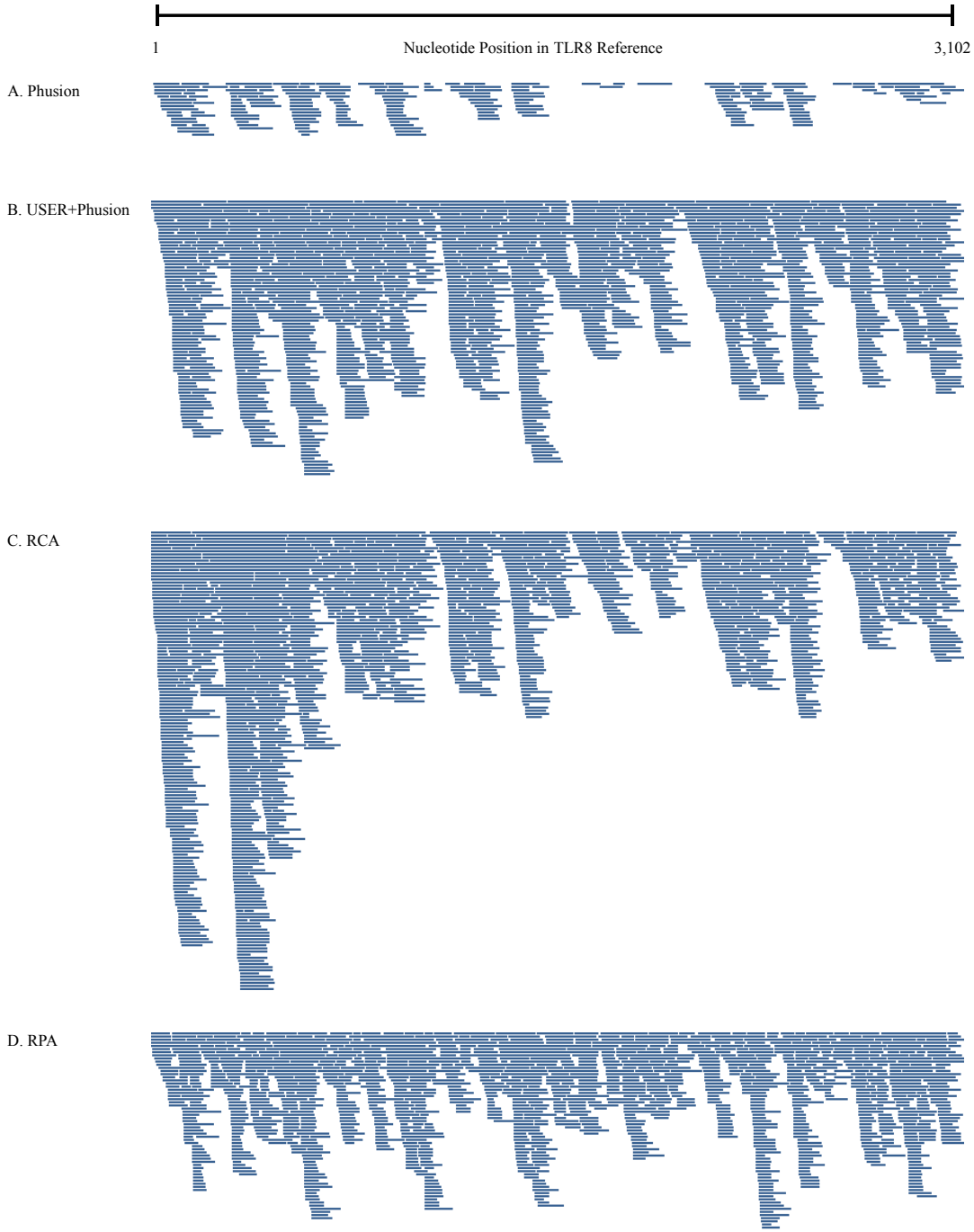
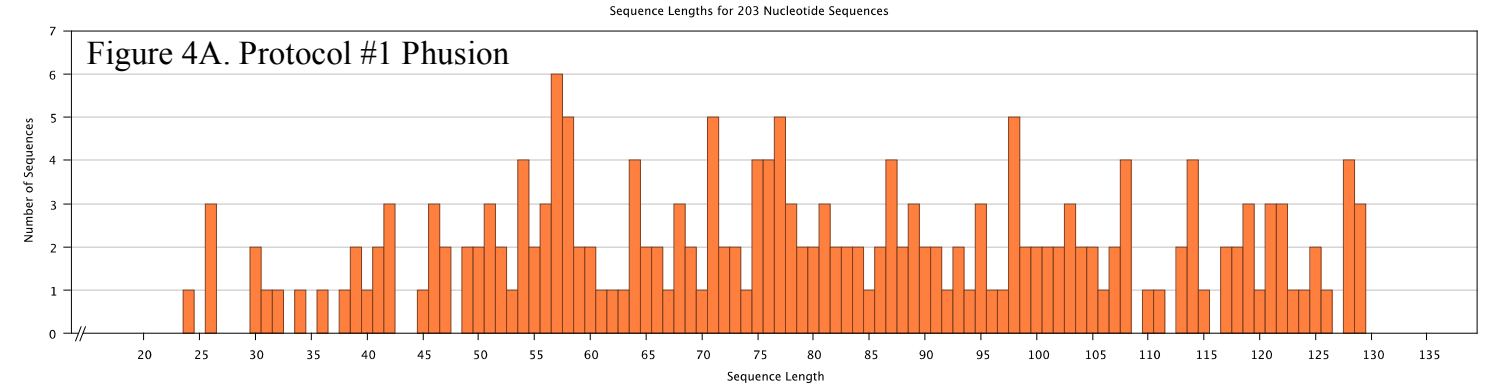
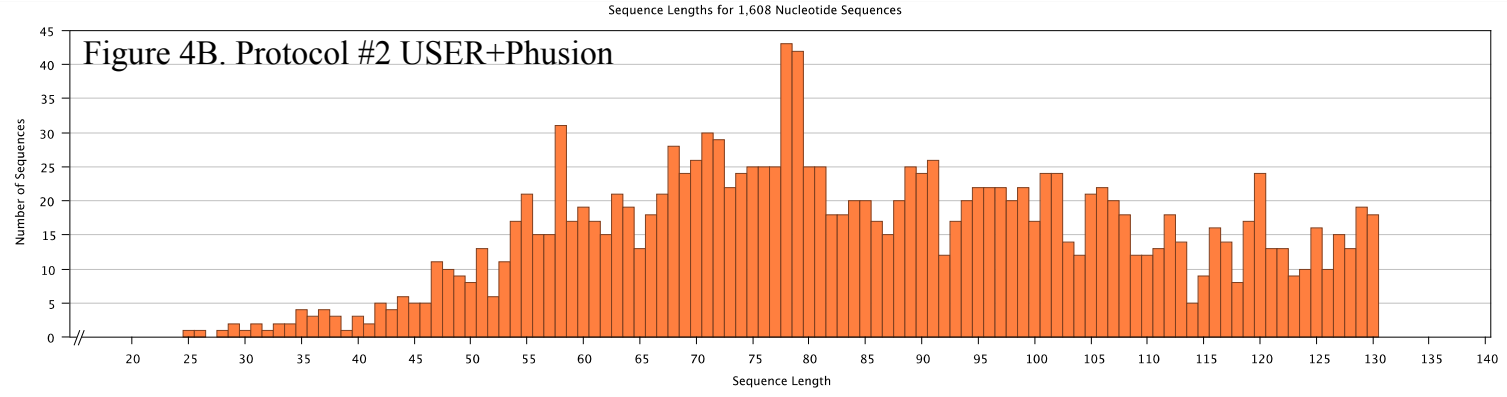


Figure 3. TLR8 unique read coverage

Schematics of the TLR8 unique read coverage were generated in the Tablet graphical viewer (http://ics.hutton.ac.uk/tablet/). **Protocol #2 USER+Phusion, Protocol#3 RCA** and **Protocol#4 RPA** all produced comparable pileups that completely covered the cattle TLR8 reference. **Protocol#1 Phusion** generated a pileup with minimal unique reads and reduced coverage of the reference TLR8 gene. Limiting the amplification of templates containing uracil by using Phusion DNA polymerase alone likely caused the poor performance of protocol #1.


Sequence Lengths for 203 Nucleotide Sequences
Figure 4A. Protocol #1 Phusion
Number of Sequences
Sequence Length


Sequence Lengths for 1,608 Nucleotide Sequences
Figure 4B. Protocol #2 USER+Phusion
Number of Sequences
Sequence Length

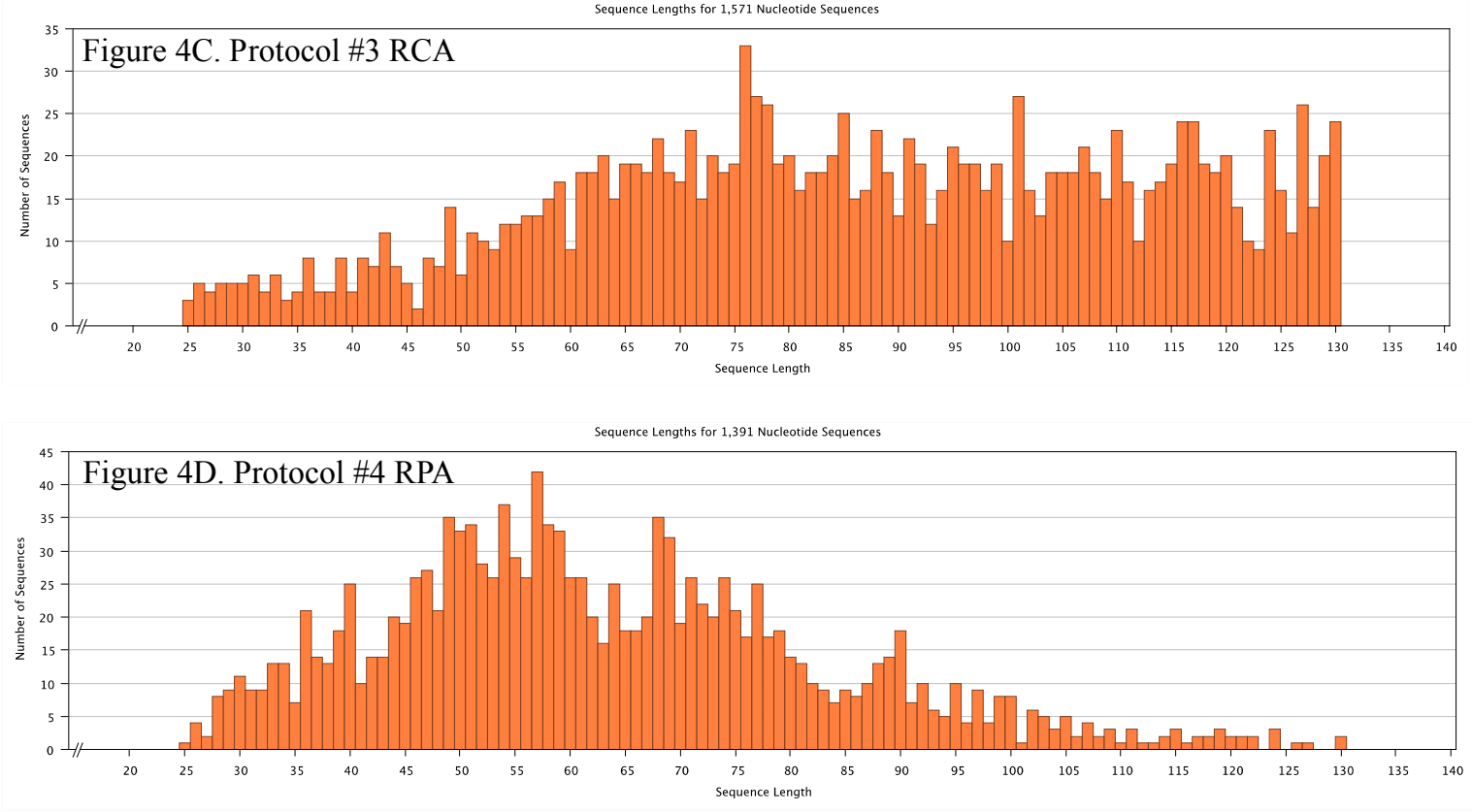


Figure 4. TLR8 unique read length distribution

The length distributions for the TLR8 unique reads were generated using Geneious 6.1.2. Unique reads from protocol #4 exhibited a distribution towards smaller sequences and a size that is more representative of unamplified aDNA.

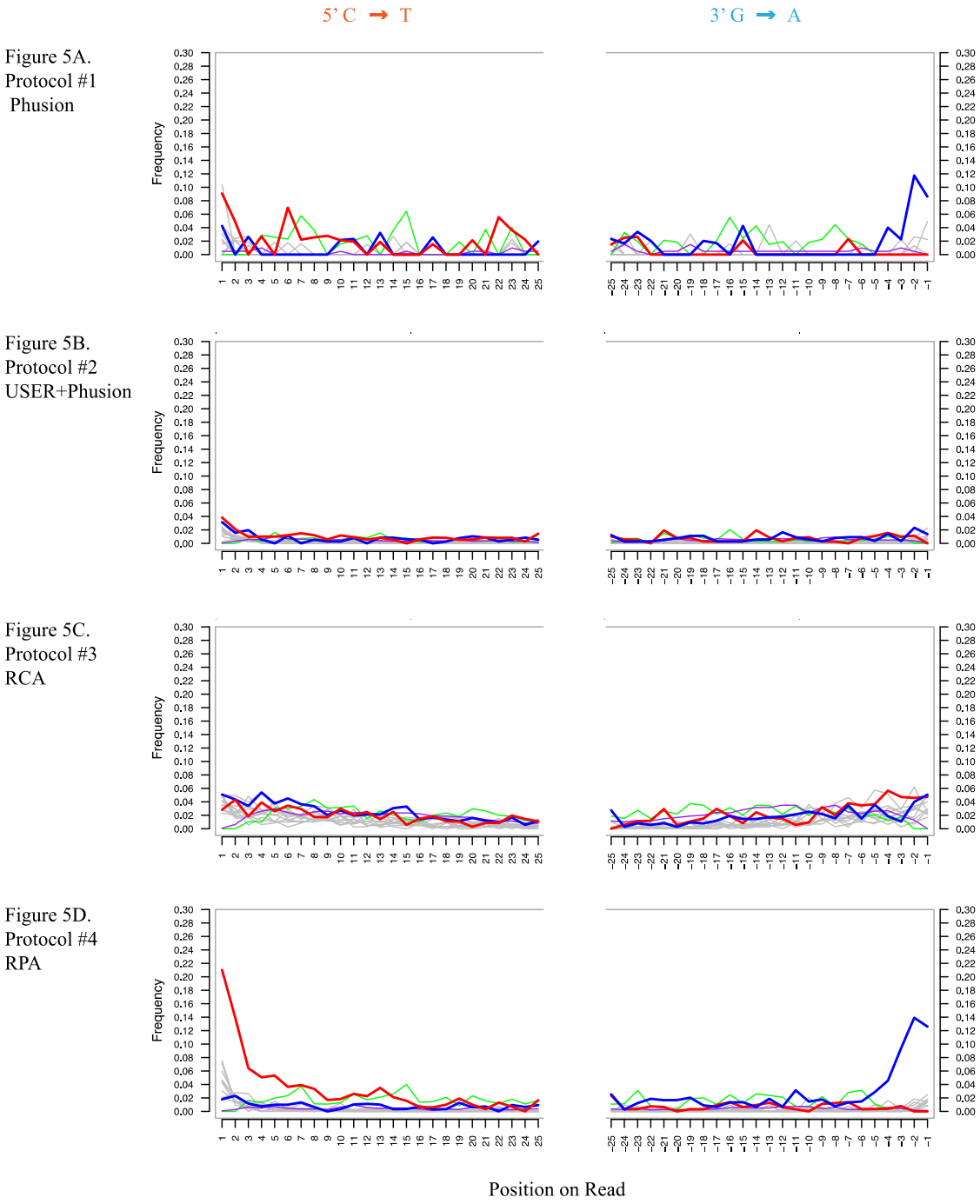


Figure 5. mapDamage misincorporation plots

Frequencies of deaminated cytosine induced misincorporations for the unique reads from each protocol were plotted using the mapDamage script [39]. Red lines represent frequencies of C→T substitutions whilst the blue lines represent G→A substitutions in comparison to the cattle reference. The other plot lines in the graphs are not relevant to this study. Protocols #1 and #4 exhibit the typical increased frequency of 5' C→T and 3' G→A misincorporations generated from deaminated cytosine.

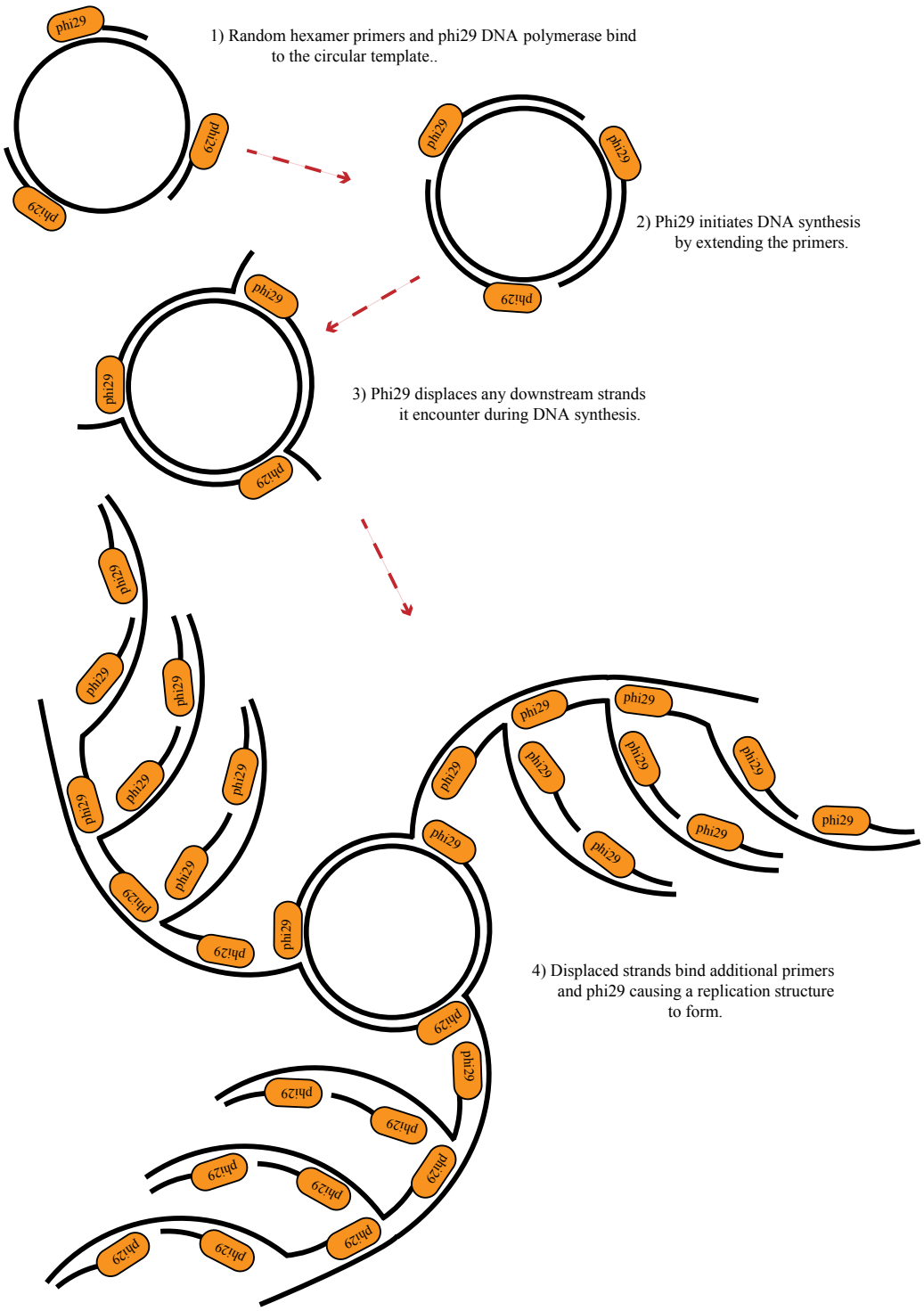


Figure S1. Rolling circle amplification
A schematic illustrating the steps involved in rolling circle amplification (RCA) of aDNA. RCA is a isothermal amplification method that makes use of a circular single stranded DNA template and is a variant of multiple displacement amplification [14].

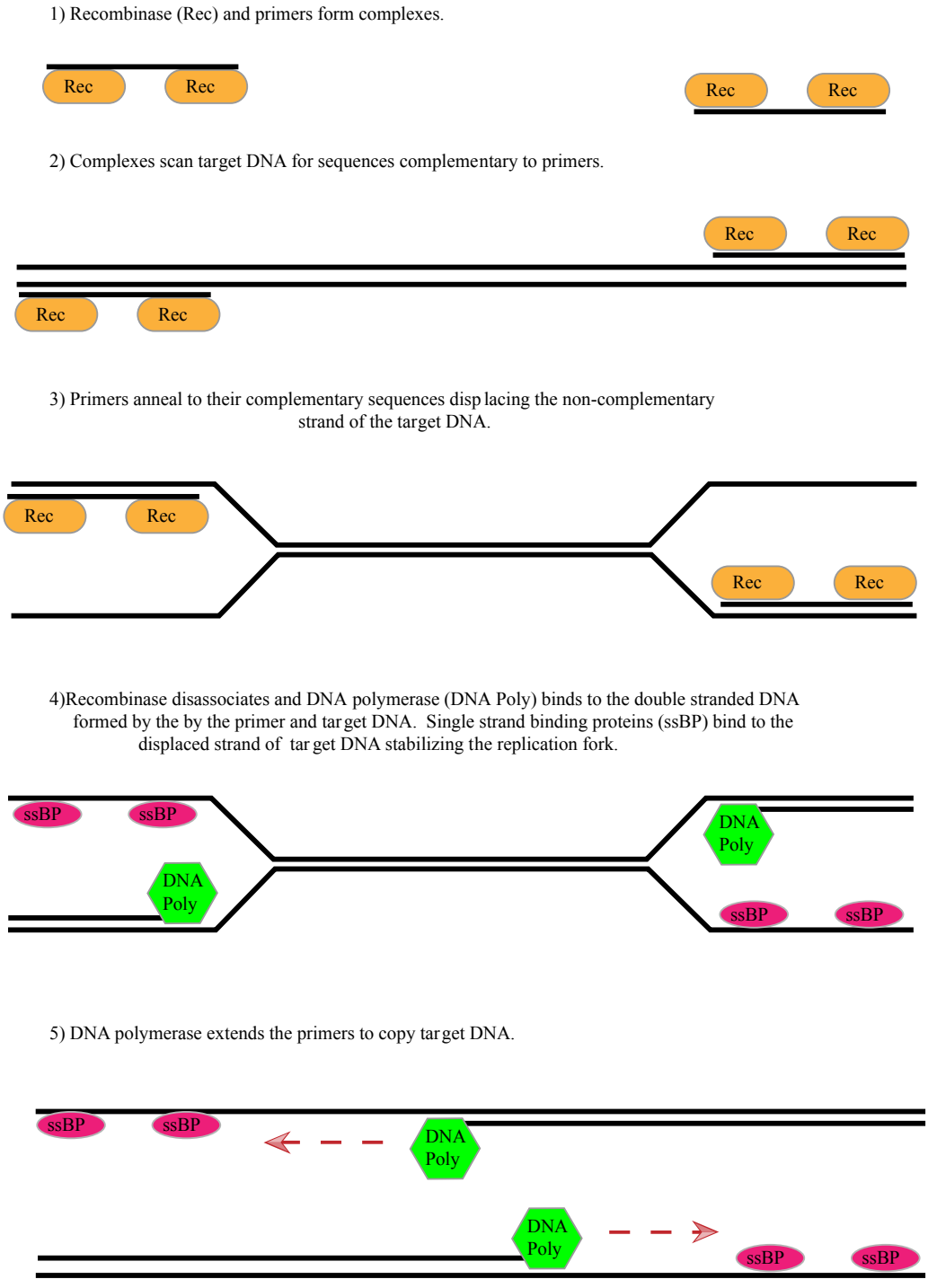


Figure S2. Recombinase polymerase amplification
A schematic illustrating the steps in recombinase polymerase amplification (RPA). RPA is an isothermal method that uses proteins involved with genetic recombination to amplify target DNA [17].

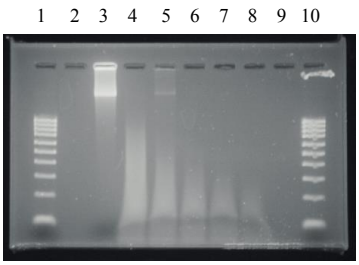


Figure S3a.

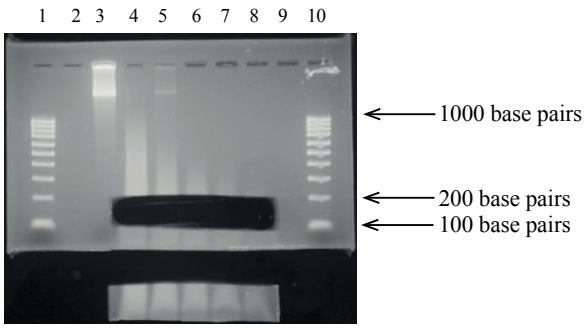


Figure S3b

Figure S3. Agarose gel of RCA product from steppe bison aDNA

An agarose gel showing the time course fragmentation and size-selection of steppe bison aDNA amplified by RCA. At each time point 2 μg of steppe bison aDNA was fragmented for the indicated length of time with the enzyme cocktail Fragmentase. For the time points, fragments 100-200 base pairs in size were excised together, purified by silica column, and then converted into a truncated Illumina sequencing library for hybridization capture of the TLR8 gene. Size-selection was performed on fragments produced from a time course of treatments in an attempt to minimize any biases that Fragmentase may introduce.

S3a - Fragmented RCA Product from steppe bison aDNA
S3b - Excised 100-200 base pair fragments of steppe bison RCA product

Lane: 1- Ladder
2- Untreated RCA Product
3- 30 minutes Fragmentase treated RCA product
4- 45 minutes Fragmentase treated RCA product
5- 60 minutes Fragmentase treated RCA product
6- 75 minutes Fragmentase treated RCA product
7- 90 minutes Fragmentase treated RCA product
8- Ladder

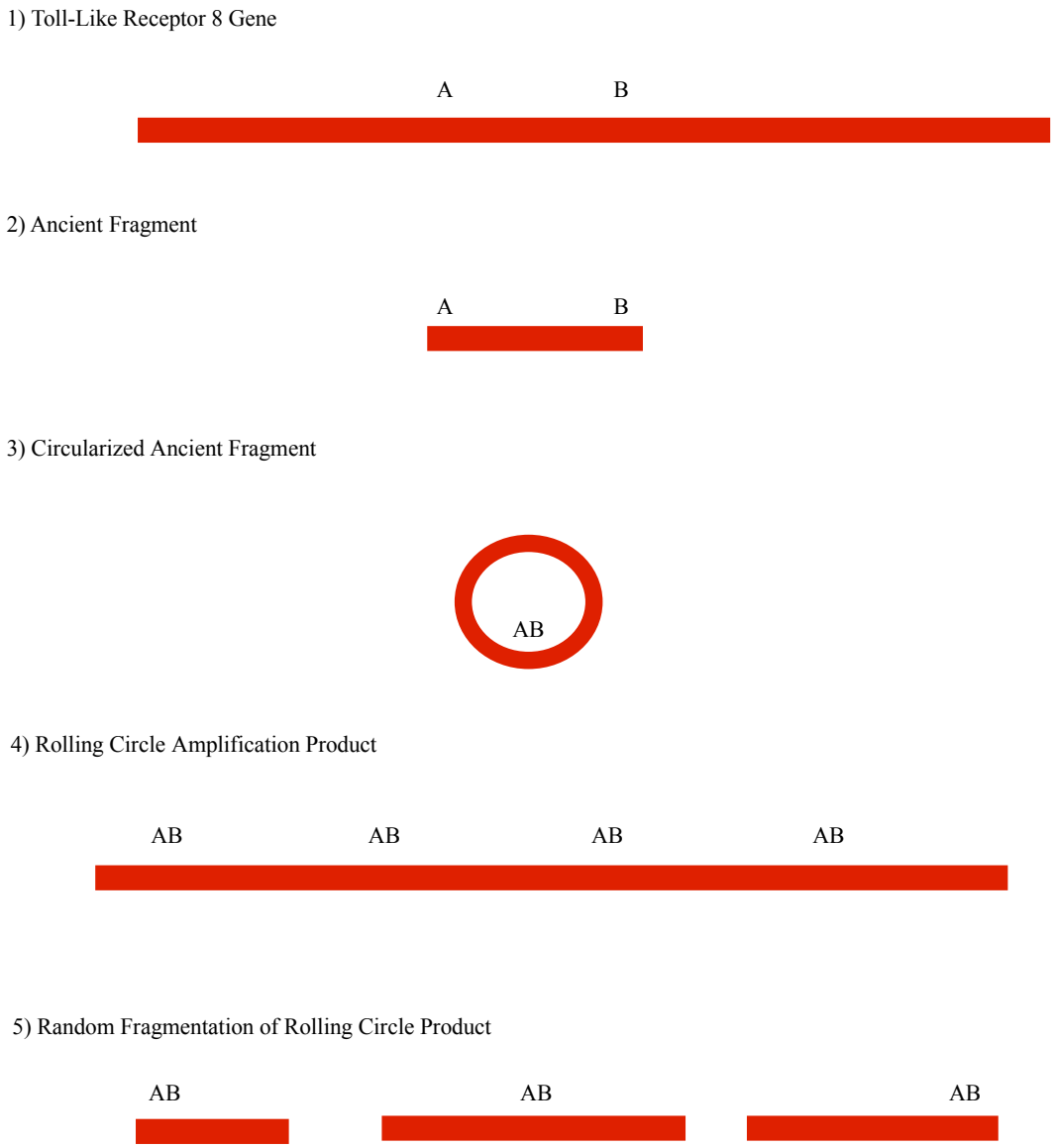


Figure S4. Illustration of the fragmentation of RCA product from aDNA
A schematic illustrating the steps involved in producing fragmented sequences from the high molecular weight RCA concatemer. Many of the fragments will contain a scrambled upstream/downstream orientation, which will complicate hybridization capture and mapping.

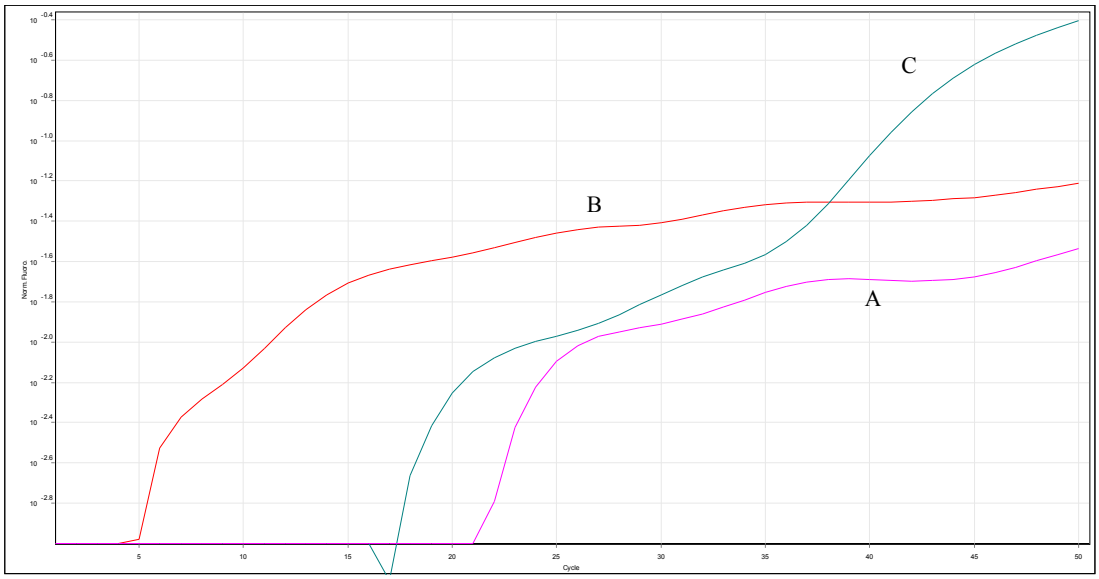


Figure S5. Example of TLR8 qPCR amplification curves
WEA1 of libraries made with steppe bison extracts and extraction blank were assayed for the presence of bison TLR8 DNA by qPCR. In the assay, all extraction blanks produce amplification curves that did not rise above a background threshold and were considered to be free of bison DNA. The WEA1 from steppe bison libraries produced amplification curves that passed the background threshold.

A – No template control
B – Fragmented extraction blank RCA product (WEA1)
C – Fragmented steppe bison (A3133) RCA product (WEA1)

Table 1. Whole extract amplification 1 (WEA1) yields

	Whole Extract Amplification 1		
	aDNA Input	Format	Product Yield (DNA)
Protocol#1 Phusion	25 μL inactivated *Bst* reaction	5 x 25 μL	0.137 μg
Protocol#2 USER+Phusion	25 μL inactivated *Bst* reaction	5 x 25 μL	0.434 μg
Protocol#3 RCA Rolling Circle Amplification	18 μL polished aDNA	50 μL	26.3 μg
Protocol#4 RPA Recombinase Polymerase Amplification	13.2 μL inactivated *Bst* reaction	50 μL	2.08 μg

Whole extract amplification 1 (WEA1) DNA yields for the two PCR and two isothermal protocols investigated in this study. Yields were examined for WEA1 because this was the only step where the amplification methods could be directly compared. This is not an ideal comparison as the methods amplified with various amounts of input aDNA that were at different stages of library processing (Supplemental Methods). The isothermal methods consistently produced greater yields for WEA1 than PCR and the discrepancies in aDNA input would not account for the differences in yields.

Table 2A. Enriched library characteristics

				MG-RAST Analysis	
	Barcoded Reads	Filtered Reads	% Filtered Read	% Identified as Prokaryotic	% Identified as Bovidae
Protocol#1 Phusion	723,608	517,320	72	27.9	53.5
Protocol#2 USER+Phusion	1,192,534	829,084	70	28.7	42.7
Protocol#3 RCA Rolling Circle Amplification	1,228,437	812,968	66	27.7	46.2
Protocol#4 RPA Recombinase Polymerase Amplification	755,371	519,914	69	12.6	61.3

Raw Ion Torrent data from the enriched steppe bison libraries were binned according to barcodes and then passed through quality filters with the Cutadapt program to retain sequences 25 to 130 base pairs in length with a quality score ≥ 20. Metagenomics analysis was performed on the filtered reads with the MG-RAST server to determine the fraction of prokaryotic (bacterial contamination) and Bovidae (steppe bison) reads. MG-RAST performed metagenomics analysis by identifying reads through comparison to annotated predicted proteins and ribosomal RNA genes. The different protocols generated TLR8 enriched libraries of similar quality with comparable fractions of filtered reads. In MG-RAST analysis, the filtered reads generated with **Protocol# 4 RPA** produced the lowest fraction of prokaryotic reads suggesting that this library was less biased towards bacterial sequences than the other libraries.

Table 2B. Characteristics of TLR8 mapped reads

	Unique Mapped Reads		Coverage of Gene	Minimum # Reads in Coverage	Maximum # Reads in Coverage
	†Raw	‡Relative			
Protocol#1 Phusion	203	0.04%	86%	0	17
Protocol#2 USER+Phusion	1608	0.19%	100%	3	88
Protocol#3 RCA Rolling Circle Amplification	1571	0.19%	100%	1	173
Protocol#4 RPA Recombinase Polymerase Amplification	1391	0.27%	100%	2	63

Filtered reads from each of the protocols were mapped to the cattle TLR8 reference using the TMAP program (https://github.com/nh13/TMAP) with a minimum mapping score of Phred ≥ 30 and clonal sequences were collapsed to produce unique reads. All protocols except #1 generated unique read pileups that covered the entire TLR8 reference. The low number of reads and minimal coverage of the TLR8 reference produce by **Protocol#1 Phusion** was likely the result of limiting the amplification of damaged aDNA templates by Phusion DNA polymerase.

†Raw - The number of TLR8 unique mapped reads.

‡Relative – The number of TLR8 unique mapped reads as a percentage of the filtered reads for the library.

Table 3. Single nucleotide polymorphism profiles

	SNP Location in Comparison to the Cattle Reference Toll-Like Receptor 8 Gene													
	340	1184	1306	1628	1672	1879	1950	1961	1962	2058	2065	2226	2608	2880
Protocl#1 Phusion	G→A	*	G→A	*	*	*	*	*	*	*	*	T→C	*	C→T
Protocol #2 USER+Phusion	G→A	C→T	G→A	C→T	C→T	G→T	T→G	T→A	C→T	A→G	G→A	T→C	G→T	C→T
Protocol #3 RCA Rolling Circle Amplification (RCA)	G→A	C→T	G→A	C→T	C→T	G→T	T→G	T→A	C→T	A→G	G→A	T→C	G→T	C→T
Protocol #4 RPA Recombinase Polymerase Amplification (RPA)	G→A	C→T	G→A	C→T	C→T	G→T	T→G	T→A	C→T	A→G	G→A	T→C	G→T	C→T

The SNP profiles of the TLR8 unique read pileups were determined using Geneious 6.1.2. A SNP required a minimum 5x read coverage to be called [41]. Protocols #2, #3, and #4 all produced similar profiles containing the identical 14 SNPs in comparison to the cattle TLR8 gene. In the pileup from Protocol #1, coverage was only sufficient to call 4 of the 14 SNPs found in the data from the other protocols.

*Lacked the minimum 5x coverage for SNP calling.

Table S1. Oligonucleotides

Oligo Name	5’ to 3’
IS1_adapter.P5	A*C*A*C*TC TTTCCCTACACGACGCTCTTCCG*A*T*C*T
IS2_adapter.P7	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
IS3_adapter.P5+P7	A*G*A*T*CGGAA*G*A*G*C
IS7_short_amp.P5	ACACTCTTTCCCTACACGAC
IS8_short_amp.P7	GTGACTGGAGTTCAGACGTGT
Ill_TwistDx_For	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Ill_TwistDx_Rev	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Bovid_LR_TLR8_F	GGCAGAGCAGGCCAACTGTCA
Bovid_LR_TLR8_R(T7)	**AATTGTAATACGACTCACTATAGGG**TGATGGACTCGTCTCACCTCTGC
ITF_FOR_BC1	CCATCTCATCCCTGCGTGTCTCCGACTCAGtgacgtgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC2	CCATCTCATCCCTGCGTGTCTCCGACTCAGacagctgACACTCTTTCCTACACGACGCTCTTCCGATCTC
ITF_FOR_BC3	CCATCTCATCCCTGCGTGTCTCCGACTCAGagcactgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_FOR_BC4	CCATCTCATCCCTGCGTGTCTCCGACTCAGtactatgACACTCTTTCCCTACACGACGCTCTTCCGATCT
ITF_REV	CCTCTCTATGGGCAGTCGGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
6R5S	rN*rN*rN*rN*rN*rN
T7-A18B	**GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAG**(A)18[B]
B_bison_TLR8_40_F	AGCTAAGGTCAAAGGCTACAGG
B_bison_TLR8_128_R	TGACAGAAGCGT CTTTGGTG
P5_short_RNAblock	ACACUCUUUCCCUACACGAC
P7_short_RNAblock	GUGACUGGAGUUCAGACGUGU

* = Phosphorothioate bond
Bold Nucleotides – T7 RNA polymerase promoter
Lower Case Nucleotide – Barcodes
B = C or G or T

Isothermal Amplification in Hybridization Capture of Ancient DNA: Supplemental Methods

0.00 aDNA Procedures

All pre-amplification procedures were performed in a dedicated low DNA laboratory at the Australian Centre for Ancient DNA (University of Adelaide, South Australia, Australia) and strict guidelines were followed to ensure the authenticity of the aDNA results [1]. The low DNA laboratory was located in a separate building from where all DNA amplifications were performed. All workspace was cleaned regularly with bleach and exposed to UV light after every procedure. Access to the low DNA laboratory was limited to properly trained staff. Extraction blank controls were performed with every aDNA library construction and all amplifications included no template controls. Extraction blank libraries were found to be negative for bison sequences by quantitative PCR (qPCR) and all no template controls were negative for bands in gel electrophoresis. After construction, all libraries were taken to a post-amplification laboratory for all further procedures.

1.00 TLR8 RNA Probe Synthesis

See Chapter II Supplemental Methods step 8.00.

2.00 Blocking RNA Synthesis

See Chapter II Supplemental Methods step 9.00.

3.00 Sample

The sample used in this study was a Late Pleistocene steppe bison (*Bison priscus*) astragalus bone collected in the Canadian Yukon Territory. The bone (ACAD

sample# A3133) was carbon dated using accelerator mass spectrometry at the Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom). Using a ^{14}C half-life of 5,568 years, the astragalus bone produced an uncalibrated date of 26,360 ± 220 radiocarbon years before present. No permits were required for the described study, which complied with all relevant regulations.

4.00 aDNA Extraction

See Chapter II Supplemental Methods step 1.00.

5.00 Preparation of Truncated Illumina Adapters

See Chapter II Supplemental Methods step 2.00.

6.00 **Protocol#1 Phusion**

6.01 Library Construction: The steppe bison aDNA was initially converted into a truncated version of an Illumina sequencing library [2] as long adapters negatively impact hybridization efficiency [3]. Polishing of the aDNA was performed in a reaction that contained 20 µL bison extract, 1x NEB Buffer 2, 4.5U of T4 DNA polymerase, 20U T4 polynucleotide kinase, 1mM ATP, 0.1 mM dNTPs, 8 µg rabbit serum albumin, and H_2O to 40 µL and was incubated at 25°C for 10 minutes. Polished aDNA was MinElute spin column purified with the PCR cleanup protocol provided and using 20 µL EB + 0.05% Tween-20 for elution. Truncated adapters were ligated to the aDNA by incubating 20 µL polished aDNA, 1 µL 25 µM P5 short adapter working stock, 1 µL 25 µM P7 short adapter working stock, 1x T4 Ligase buffer, 5% (w/v) polyethylene glycol 4000, 6U T4 DNA ligase and H_2O to 40 µL at 22°C for 1 hour. The ligation reaction was MinElute purified as before and used in a

strand displacement reaction to complete the construction of the adapters. The strand displacement reaction contained 20 μL of ligated aDNA, 1x Thermopol buffer, 19.2U *Bst* DNA polymerase large fragment, 250 μM dNTPs and H_2O to 40 μL and was incubated at 37°C for 10 minutes followed by heating to 80°C for 20 minutes to inactivate the *Bst* [4].

6.02 Whole Extract Amplification 1 (WEA1): Amplification was performed with Phusion an archeal DNA polymerase that stalls on uracil and consequently does not amplify templates containing deaminated cytosine efficiently [5]. Amplification with Phusion should minimize misincorporations in the final enriched library [6]. The steppe bison library was initially amplified in five PCRs, each containing: template (5 μL inactivated *Bst* reaction), 1x Phusion HF buffer, 200 μM dNTPs, 200 μM each of primers IS7_short_amp.P5 and IS8_short_amp.P7 (Table S1), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 μL. Amplification was performed in a heated lid thermal cycler programed as follows: 1 cycle: 98°C for 30 seconds; 14 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. After amplification, products in 2 μL of each PCR were gel electrophoresed and produced smears approximately 150 to 300 base pairs in length. The PCRs were pooled and combined with 1.8 volumes of Ampure XP in a 1.5 ml low-bind tube for purification. Ampure and the PCR pool were mixed well and allowed to stand for 5 minutes on the bench top. The tube was placed in a magnetic rack for three minutes to pellet the beads and the beads were then washed three times with 800 μL 70% ethanol. After discarding the last wash, the beads were dried by leaving the uncapped tube in the magnetic rack for 10 minutes. Library was eluted by resuspending the beads in 30 μL 10 mM Tris pH 8.0 + 0.05% Tween-20 and letting

the tube stand on the lab bench for 5 minutes. The beads were then pelleted by placing the tube in the magnetic rack and after 3 minutes the supernate, containing the library, was transferred to a fresh low bind tube. The library was quantified with a NanoDrop 2000 spectrophotometer and then stored at -20°C.

6.03 Whole Extract Amplification 2 (WEA2): WEA2 was performed as in the previous step, except that 3 ng of WEA1 product was used as template.

6.04 Primary TLR8 Hybridization Capture: The primary hybridization capture was performed using three Reagent Tubes prepared as follows:

Reagent Tube 1- 3.5 µL **Phusion**-WEA2 at 41.1 ng/ µL

Reagent Tube 2 - 5 µL TLR8 probe (Step *1.00*), 1 µL HyBloc RNA (Step *2.0*), and 0.5 µL of stock containing 50 µM P5_short_RNAblock and P7_short_RNAblock block (Table S1)

Reagent Tube 3- 30 µL Hybridization Buffer
75% formamide, 75 mM HEPES, pH 7.3,
3 mM EDTA, 0.3% SDS, and 1.2 M NaCl [7]

The P5/P7 short_RNAblock are RNA oligonucleotides that are complementary to the library adapters and were included in the reaction to prevent reannealing of the library adapters. Hybridization capture was carried out in a heated lid thermal cycler with the following program:

Step 1- 94°C for 2 minutes
Step 2- 65°C for 3 minutes
Step 3- 42°C for 2 minutes
Hold 4- 42°C infinite

To prepare for hybridization capture, Reagent Tubes were placed in the thermal cycler at the start of each program Step in the following order:

Step 1- Reagent Tube 1
Step 2- Reagent Tube 2
Step 3- Reagent Tube 3

Once the Hold cycle began, 20 µL of the hybridization buffer from Tube 3 was mixed with the RNA in Tube 2. The entire content of Tube 2 was then mixed with the DNA in Tube 1 to begin the hybridization capture reaction. The hybridization capture was carried out at 42°C for 48 hours.

Just prior to the end of the hybridization incubation, magnetic streptavidin beads were washed and blocked. To wash, 50 µL beads were added to 0.5 mL Wash Buffer 1(2.0x SSC and 0.05% Tween-20) and vortexed briefly. The beads were then centrifuged, pelleted in a magnetic rack, and the supernate discarded. The beads were washed a second time (as above). For blocking, beads were suspended in 0.5 ml Wash Buffer 1 + 100 µg yeast t-RNA and incubated for 30 minutes at room temperature on a rotor. Blocked beads were pelleted in a magnetic rack, washed once with 0.5 mL Wash Buffer 1, and suspended in fresh 0.5 mL Wash Buffer 1.

At the end of the 48 hour incubation, the hybridization reaction was combined with the blocked beads and mixed on a rotor for 30 minutes at room temperature. The beads were then pelleted with a magnetic rack and the supernate discarded. The beads were then taken through a series of washes with increasing stringency as outlined below:

Wash 1 - 0.5 mL Wash Buffer 1 at room temperature for 10 minutes

Wash 2 - 0.5 mL Wash Buffer 2 (0.75x SSC and 0.05% Tween-20) at 50°C for 10 minutes

Wash 3 - 0.5 mL Wash Buffer 2 at 50°C for 10 minutes

Wash 4 - 0.5 mL Wash Buffer 3 (0.20x SSC and 0.05% Tween-20) at 50°C for 10 minutes

After the last wash, the captured library was liberated from the probe by incubating the beads at room temperature for 10 minutes in 50 µL Release Buffer (0.1 M NaOH) and then adding 70 µL Neutralization Buffer (1M Tris-HCl pH 7.5). The beads were again pelleted and the supernate, containing the captured library, was transferred to a fresh tube for purification with a modified MinElute protocol. The captured library was combined with 650 µL PB buffer and 10 µL 3 M sodium acetate to adjust the pH for efficient DNA binding to the MinElute column. Purification was performed with the standard PCR protocol except the library was eluted from the spin column with 35 µL EB + 0.05% Tween-20.

6.05 Amplification of Primary TLR8 Library: Library recovered from the primary TLR8 hybridization capture was used as template in the PCR procedure outlined in step *6.02*. The purified product of this amplification was called **Phusion** Primary Library.

6.06 Secondary TLR8 Hybridization Capture: The secondary TLR8 hybridization capture was performed using the methods outlined in step *6.04* except for the following change:

Reagent Tube 1 - 3.5 µL **Phusion** Primary Library at 58.0 ng/ µL

6.07 Amplification of Secondary TLR8 Library: DNA recovered from the secondary TLR8 hybridization capture was converted to an Ion Torrent library by amplification with fusion primers. Amplification was performed in four PCRs containing: template (5 µL library from secondary hybridization TLR8 capture), 200 µM each of primers ITF_FOR_BC1 and ITF_REV (Table S1), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Fusion primer amplification was performed in a heated lid thermal cycler programed as follows: 1 cycle: 98°C for 30 seconds; 11 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. The PCRs were pooled and processed as in step *6.02*.

7.00 **Protocol#2 USER+Phusion**

7.01 Library Construction: During library construction the aDNA was treated with USER, an enzyme cocktail that has been demonstrated to remove uracil from aDNA molecules [4]. Polishing and removal of uracil was performed in a reaction that initially contained 20 µL of bison extract, 1x NEB Buffer 2, 3U USER enzyme cocktail, 20U T4 polynucleotide kinase, 1mM ATP, 0.1 mM dNTPs, 8 µg rabbit serum albumin, H_2O to 38.5 µL, and was incubated at 37°C for 3 hours. Then 4.5U of T4 DNA polymerase was added and the reaction was incubated at 25°C for an additional 30 minutes. The aDNA was then MinElute purified and processed as in step *6.01* from adapter ligation onward.

7.02 Whole Extract Amplification 1 (WEA1): The WEA1 was performed as in step 6.*02*.

7.03 Whole Extract Amplification 2 (WEA2): The WEA2 was performed as in step *6.03*

7.04 Primary TLR8 Hybridization Capture: The primary hybridization capture was performed using the methods given in step *6.04* except for the following change:

Reagent Tube 1 - 3.5 µL **USER+Phusion**-WEA2 at 39.1ng/ µL

7.05 Amplification of Primary TLR8 Library: Library recovered from the primary TLR8 hybridization capture was used as template in the PCR procedure outlined in step *6.02*. The purified product of this amplification was called **USER+Phusion** Primary Library.

7.06 Secondary TLR8 Hybridization Capture: The secondary TLR8 hybridization capture was performed with the procedures given in step *6.04* except:

Reagent Tube 1 - 3.5 µL of **USER+Phusion** Primary Library at 53 ng/ µL

7.07 Amplification of Secondary TLR8 Enrichment: Library recovered from the secondary TLR8 hybridization capture was used as template in the PCR procedure outlined in step *6.07* using the fusion primers ITF_FOR_BC2 and ITF_REV (Table S1).

8.00 **Protocol#3 Rolling Circle Amplification (RCA)**

8.01 RCA Library Construction and WEA1: Ancient DNA was circularized with CircLigase II and then taken through rolling circle amplification (RCA) using random RNA hexamer primers, which have been reported to reduce non-specific

amplification [8]. The aDNA was first polished in a reaction that contained 20 µL of bison extract, 1x NEB Buffer 2, 20U T4 polynucleotide kinase, 1mM ATP, 0.1 mM dNTPs, 8 µg rat serum albumin, and H_2O to 40 µL. Polishing was carried out at 25°C for 10 minutes followed by MinElute purification as in step *6.01*. Polished aDNA was denatured for circularization by heating at 80°C for 5 minutes then cooling at 4°C for 5 minutes [9]. Circularization was performed in a 30 µL reaction containing 18 µL denatured polished aDNA, 1x CircLigase Buffer, 2.5 mM $MnCl_2$, 20% betaine, and 150U CircLigase II. The circularization reaction was incubated at 60°C for 3 hours and then treated with 20U exonuclease I and 100U exonuclease III at 37°C for 45 minutes to destroy any linear aDNA. The exonucleases were inactivated by heating the reaction at 80°C for 20 minutes and the circular aDNA purified with a MinElute spin column as in step 6.*01*. Next, the circular DNA was primed for RCA by annealing random RNA hexamer primers in a reaction containing 20 µL circular aDNA, 2.5 µL 10x Oligo hybridization buffer (500 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA) [2] and 2.5 µL of 500 µM primer 6R5S random RNA hexamer (Table S1) that was heated to 80°C for 2 minutes then cooled to 25 °C at a rate of - 0.1°C/second [8]. Once cooled, the following were added to the priming reaction: 5 µL 10x phi29 buffer, 4 µL 10 mM dNTPs, 2.5 µL of 500 µM primer 6R5S random RNA hexamer, 0.1 U yeast inorganic pyrophosphatase [10] and H_2O to 46 µL. RCA was started by mixing 4 µL 10 U/ µL phi29 DNA polymerase to the priming reaction and then incubating at 30°C for 16 hours followed by heating at 65°C for 10 minutes to inactivate the phi29 enzyme. For purification, the RCA reaction was combined with 50 µL H_2O and 70 µL Ampure XP beads in a 1.5 mL tube and then incubated for 5 minutes on the bench top. The tube was then placed in a magnetic rack and the supernate discarded after pelleting the beads for three minutes. The beads were then

resuspended in 500 µL of 70% ethanol and pelleted as before. Keeping the tube in the rack, the supernate was discarded and the beads washed twice more with 500 µL of 70% ethanol. After the final wash was discarded, the open tube was left in the magnetic rack for 10 minutes to dry the beads. The dried beads were suspended in 30 µL 10 mM Tris pH 8.0+ 0.05% Tween-20 and allowed to stand for 5 minutes. The beads were then precipitated and the supernate, containing the RCA product, was transferred to a fresh 1.5 mL tube and quantify with a NanoDrop 2000. The RCA product was visualized on a agarose gel and produced a high molecular weight smear (Figure S3)

8.02 Fragmentation and Size-Selection of the RCA Product: The high molecular weight RCA product required fragmentation into a smaller size for use in hybridization capture. To minimize the introduction of bias by Fragmentase the enzyme cocktail used for fragmentation, products from a time series of reactions were pooled to act as template for library construction. Five fragmentation reactions were assembled each containing: 2 µg RCA product, 1x Fragmentase buffer, and H_2O to 36 µL. The reactions were vortexed and then incubated at 4°C for 5 minutes. To start fragmentation, 4 µL of Fragmentase was added to each tube and the tubes vortex and then incubated at 37°C for varying amounts of time. Reactions were stopped at 30, 45, 60, 75, and 90 minutes by the addition of 5 µL of 0.5 M EDTA. Reactions were electrophoresised on a 2% agarose gel and fragments 100-200 base pairs in length excised from all time points (Figure S3). The excised fragments were purified together with a QIAquick Gel Extraction kit, eluted in 30 µL EB + 0.05% Tween-20, and quantified with a NanoDrop 2000.

8.03 Conversion of RCA Product to Truncated Illumina Library: Twenty nanograms of the purified RCA fragments were constructed into a truncated Illumina library using the procedure outline in step *6.01*.

8.04 WEA2: The inactivated *Bst* reaction from ligation of truncated adapters to RCA fragments (step above) was used as template in the PCR procedure outline in step 6.02. The purified product of this amplification was called **RCA**-WEA2.

8.05 Primary TLR8 Hybridization Capture: The primary hybridization capture was performed as in step *6.04* except for the following modifications to these Reagent Tubes:

Reagent Tube 1- 3.0 µL **RCA**-WEA2 at 119 ng/ µL

Reagent Tube 2 - 5 µL TLR8 probe, 1 µL HyBloc RNA, and
1.0 µL of stock containing 50 µM
P5_short_RNAblock and
P7_short_RNAblock block

8.06 Amplification of Primary TLR8 Capture: Library recovered from the primary TLR8 hybridization capture was used as template in the PCR procedure outlined in step *6.02*. The purified product of this amplification was called **RCA** Primary Library.

8.07 Secondary TLR8 Capture: The secondary hybridization capture was performed with the procedures described in step *6.04* except with the following modifications to these Reagent Tubes:

Reagent Tube 1- 3.0 µL of **RCA** Primary Library
at 76.8 ng/ µL

Reagent Tube 2- 5 µL TLR8 probe, 1 µL HyBloc RNA, and
1.0 µL of stock containing 50 µM
P5_short_RNAblock and
P7_short_RNAblock block

8.10 Amplification of Secondary TLR8 Capture: Library recovered from the secondary TLR8 hybridization capture was used as template in the PCR procedure outlined in step *6.07* using the fusion primers ITF_FOR_BC3 and ITF_REV (Table S1).

9.00 **Protocol#4 Recombinase Polymerase Amplification (RPA)**

9.01 Library Construction: A truncated library was constructed as in step *6.01*.

9.02 Whole Extract Amplification 1 (WEA1): RPA was carried out on the ancient bison library using a TwistDx Basic kit. The amplification reaction contained: template (13.2 µL inactivated *Bst* reaction), 480 mM of each primer Ill_TwistDx_For and Ill_TwistDx_Rev (Table S1), 14 mM $Mg(CH_3COO)_2$, and the provided lyophilized reagents. The RPA reaction was incubated at 38°C for 40 minutes and then transferred to a 1.5 mL tube. The reaction was mixed with 50 µL H_2O and 180 µL Ampure XP beads, and allowed to stand on the bench top for 5 minutes. The tube was then placed in a magnetic rack and the supernate discarded after pelleting the beads for three minutes. The beads were then resuspended in 500 µL of 70% ethanol and pelleted as before. Keeping the tube in the rack, the supernate was discarded and the beads washed twice more with 500 µL of 70% ethanol. After the final wash was discarded, the open tube was left in the magnetic rack for 10 minutes to dry the beads. The dried beads were suspended in 30 µL 10 mM Tris pH 8.0+ 0.05% Tween-20 and allowed to stand for 5 minutes. The beads were then precipitated and the supernate,

containing the RPA product, was transferred to a fresh 1.5 mL tube and quantify with a NanoDrop 2000. Two microliters of the purified RPA product was gel electrophoresed and produced a smear approximately 100 – 500 base pairs in length.

9.03 Primary TLR8 Hybridization Capture: The primary hybridization capture was performed using the protocol given in *6.04* except the contents of the following Reagent Tubes were altered as follows:

Reagent Tube 1- 3.0 µL **RPA**-WEA1 at 83.0 ng/ µL

Reagent Tube 2 - 5 µL TLR8 probe, 1 µL HyBloc RNA, and
1.0 µL of stock containing 50 µM
P5_short_RNAblock and
P7_short_RNAblock block

9.05 Amplification of Primary TLR8 Capture: Library from the primary hybridization capture was used as template in the TwistDx procedure outlined in step 9.02. The purified product of this amplification was called **RPA** Primary Library.

9.06 Secondary TLR8 Capture: The Secondary hybridization capture was performed using the procedure given in step *6.04* except these Reagent Tubes were modified as follows:

Reagent Tube 1- 3.0 µL of **RPA** Primary Library
at 82 .0 ng/µL

Reagent Tube 2 - 5 µL TLR8 probe, 1 µL HyBloc RNA, and
1.0 µL of stock containing 50 µM
P5_short_RNAblock and
P7_short_RNAblock block

9.07 Amplification of Secondary TLR8 Enrichment: Amplification of library from the secondary hybridization capture was performed with a TwistDx kit and Ion Torrent fusion primers. The RPA reaction contained: 13.2 µL library from Secondary TLR8 hybridization capture, 480 mM of each primer ITF_FOR_BC4 and ITF_REV, 14 mM $Mg(CH_3COO)_2$, and the provided lyophilized reagents. The RPA reaction was incubated at 38°C for 40 minutes then purified with Ampure beads as in step *9.02*.

10.00 Extraction Blank Controls

Extraction blanks were taken through similar library preparation and WEA1s used with bison aDNA. Products from aDNA and extraction blank WEA1s were used in quantitative PCR (qPCR) to test for contamination. qPCR was performed in triplicate using tubes containing 3 ng WEA1,1x Brilliant Green II SYBR Green Master Mix, 500 nM of each of the primers B_bison_TLR8_40_F and B_bison_TLR8_128_R, 0.5 µg BSA, and H_2O to 10 µL. The samples were amplified on a Roto-Gene 6000 thermal cycler running Roto-Gene v1.7 [Build 87] software and scanning on the SYBR green channel. The amplification program was 1 cycle: 95°C for 5 minutes; 50 cycles: 95°C for 10 seconds, 58°C for 20 seconds, 72°C for 15 seconds, and a melt curve that ramped from 55°C to 95°C. Extraction blank WEA1s produced fluorescence curves similar to no template controls and were considered free of contamination. In contrast, the WEA1s produced with aDNA, generated curves that passed above a background threshold after approximately 38 cycles (Figure S5).

11.00 Ion Torrent Sequencing

Fusion primer library molarity and size were determined with an Agilent 2200 TapeStation running a High Sensitivity D1K ScreenTape. TapeStation molarities

were used to pool the libraries in equal amounts with the Template Dilution Factor (Ion OneTouch System User Guide, page 97, 4472430 Rev. E). Fusion primer libraries were emulsion amplified using an OneTouch DL System running reagents from an Ion OneTouch 200 Template Kits v2 DL and following the protocol provided in Ion OneTouch Quick Guide (publication# Man006959 Rev 5.0). Qualities of the emulsion amplified libraries were verified with a Qubit 2.0 fluorometer using an Ion Sphere Quality Control Kit. Sequencing was performed on an Ion Torrent Personal Genome Machine with 316 chips running reagents from an Ion PGM Sequencing 200 Kits v2. Sequencing was performed following the protocol provided in the Sequencing 200 Kits v2 (publication# Man0007273 Rev 1.0).

12.00 Data Analysis

Sequence data was obtained from the PGM in BAM format and analyzed through a bioinformatics pipeline implementing several available software packages. FastX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/) was used to demultiplex reads by fusion primer barcodes, using a strict zero mismatches threshold (--bol --mismatches 0). Cutadapt v1.2 [11] was used to trim adapter sequences using a maximum error rate of 0.33 (-e 0.3333), and to remove short (-m 25 bp), long (-M 130 bp) and low quality sequences (-q 20), with a total of five passes (-n 5). The characteristics of the filtered reads were checked with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Captured reads were mapped to the cattle TLR8 reference (AC_000187.1) using TMAP v3.2.1 (https://github.com/nh13/TMAP) with the following options (-g 3 -M 3 -n 7 -v stage1 --stage-keep-all map1 --seed-length 12 --seed-max-diff 4 stage2 map2 --z-best 5 map3 --max-seed-hits 10). Reads with mapping quality below Phred 30 and read

duplicates were removed with SAMtools [12] and the MarkDuplicates tool of Picard Tools v1.79 (http://picard.sourceforge.net). GC content of mapped reads was analyzed using the CollectGcBiasMetrics tool of Picard Tools v1.79 and misincorporation patterns were assessed using MapDamage v0.3.6 [13]. The resulting unique read pileups and the cattle TLR8 gene were imported into Biomatters Geneious Pro v R6.1 software (http://www.geneious.com/) and the SNP profiles of each of the enriched libraries in comparison to the reference were generated using a minimum of 5x coverage to call a variant. Mapped data and the TLR8 reference were also imported into the Tablet sequence browser to generate the coverage graphics [14]. Boxplots for filtered read GC content and length were generated using the R statistical software (http://CRAN.R-project.org/package=tweedie). Filtered reads were also imported in to the MG-RAST server to perform metagenomics analysis on the sequence data [15]. Metagenomics analysis was performed on the filtered reads using the default settings except that the filtering option was deactivated. MG-RAST performs analysis through read similarity to known predicted proteins and ribosomal RNA genes. In this study, annotated reads classified as prokaryotic were considered to be bacterial contamination. Since there is no bison reference genome for comparison, reads that were classified as belonging to the family Bovidae, which includes bison, were deemed to be endogenous steppe bison sequences.

Equipment and Materials

Hot Start Phire II DNA Polymerase	Thermo Fisher Scientific, Vic, AU
Calf Thymus DNA	Affymetrix, CA, USA
Oligonucleotides	Intergraded DNA Techn., IA, USA
dNTPs mix	New England Biolabs, MA, USA
QIAquick PCR Purification kit	Qiagen, Vic, AU
NanoDrop 2000 spectrophotometer	Thermo Fisher Scientific, Vic, AU
T7 High Yield RNA Synthesis kit	New England Biolabs, MA, USA
Turbo DNase	Life Technologies, Vic, AU
MEGAclear spin columns	Life Technologies, Vic, AU
NEBNext Mg^{++} RNA Fragm. Buffer	New England Biolabs, MA, USA
RNeasy MinElute spin columns	Qiagen, Vic, AU
Photoprobe (Long Arm) Biotin	Vector Lab. LTD, CA, USA
Hybloc DNA	Applied Genetics Lab., FL,USA
T4 polynucleotide kinase	New England Biolabs, MA, USA
T4 DNA polymerase	New England Biolabs, MA, USA
MinElute Spin columns	Qiagen, Vic, AU
Terminal transferase	New England Biolabs, MA, USA
dTTP	New England Biolabs, MA, USA
ddCTP	Affymetrix, CA, USA
DNA poly. I Klenow fragment	New England Biolabs, MA, USA
Dremel tool	Dremel, CA, USA
Braun micro-dismembrator	Braun, Hesse, DE
Silica	Sigma-Aldrich, NSW, AU
QG Buffer	Qiagen, Vic, AU
MinElute Spin columns	Qiagen, Vic, AU
Triton x100	Sigma-Aldrich, NSW, AU
NaCl	Sigma-Aldrich, NSW, AU
Sodium acetate	Sigma-Aldrich, NSW, AU
Ethanol	Sigma-Aldrich, NSW, AU
1M Tris-HCl pH 8.0	Life Technologies, Vic, AU
1M Tris-HCl pH 7.5	Sigma-Aldrich, NSW, AU
O.5 M EDTA pH 8.0	Life Technologies, Vic, AU
NEB Buffer 2	New England Biolabs, MA, USA
10 mM ATP	New England Biolabs, MA, USA
Tween-20	Thermo Fisher Scientific, Vic, AU
T4 DNA ligase	Thermo Fisher Scientific, Vic, AU
5% (w/v) polyethylene glycol 4000	Thermo Fisher Scientific, Vic, AU
Bst DNA polymerase large fragment	Thermo Fisher Scientific, Vic, AU
1x Thermopol buffer	Thermo Fisher Scientific, Vic, AU
Phusion Hot Start II DNA poly.	New England Biolabs, MA, USA
Ampure XP beads	Beckman Coulter, NSW, AU
Formamide	Sigma-Aldrich, NSW, AU
1M HEPES, pH 7.3	Affymetrix, CA, USA
20x SSC buffer	Sigma-Aldrich, NSW, AU
Magnetic streptavidin beads	New England Biolabs, MA, USA
Yeast t-RNA	Life Technologies, Vic, AU
USER enzyme cocktail	New England Biolabs, MA, USA
CircLigase II	Epicentre Biotechn., WI, USA

Rat serum albumin	Life Technologies, Vic, AU
Exonuclease I	New England Biolabs, MA, USA
Exonuclease III	New England Biolabs, MA, USA
Yeast inorganic pyrophosphatase	New England Biolabs, MA, USA
Phi29 DNA polymerase	New England Biolabs, MA, US
Fragmentase	New England Biolabs, MA, US
QIAquick Gel Extraction kit	Qiagen, Vic, AU
TwistDx Basic kit	TwistDx, Cambridge, UK
2x Brilliant Green II SYBR Green	Life Technologies, Vic, AU
Roto-Gene 6000 thermal cycler	Qiagen, Vic, AU
2200 TapeStation	Agilent, Vic, AU
High Sensitivity D1K ScreenTape	Agilent, Vic, AU
OneTouch DL System	Life Technologies, Vic, AU
Ion OneTouch 200 Templ Kits v2 DL	Life Technologies, Vic, AU
Qubit 2.0 fluorometer	Life Technologies, Vic, AU
Ion Sphere Quality Control Kit	Life Technologies, Vic, AU
Ion Torrent PGM	Life Technologies, Vic, AU
Ion PGM Sequencing 200 Kits v2	Life Technologies, Vic, AU
Ion 316 chip	Life Technologies, Vic, AU
H_2O (Molecular Biology Grade)	Life Technologies, Vic, AU

Supplemental References

1. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
2. Knapp M, Stiller M, Meyer M (2012) Generating barcoded libraries for multiplex high-throughput sequencing. Methods in Molecular Biology 840: 155-170.
3. Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Research 22: 939-946.
4. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.
5. Greagg MA, Fogg AM, Panayotou G, Evans SJ, Connolly BA, et al. (1999) A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. Proceedings of the National Academy of Sciences of the United States of America 96: 9045-9050.
6. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, et al. (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463: 757-762.
7. Konietzko U, Kuhl D (1998) A subtractive hybridisation method for the enrichment of moderately induced sequences. Nucleic Acids Research 26: 1359-1361.
8. Takahashi H, Yamamoto K, Ohtani T, Sugiyama S (2009) Cell-free cloning using multiply-primed rolling circle amplification with modified RNA primers. Biotechniques 47: 609-615.
9. Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, et al. (2011) True single-molecule DNA sequencing of a pleistocene horse bone. Genome Research 21: 1705-1719.
10. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 99: 5261-5266.
11. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200. EMBnetjournal.
12. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
13. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153-2155.
14. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, et al. (2013) Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14: 193-202.
15. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9: 386.

Statement of Authorship

Title of Paper	Detection of Altered Bases in Ancient DNA Using SMRT Sequencing
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards		
Contribution to the Paper	Helped conceive study design, performed all experiments, helped analyze data, wrote paper		
Signature		Date	June 18, 2015

Name of Co-Author	Oliver Lomax-James Wooley		
Contribution to the Paper	Helped analyze data, helped edit paper		
Signature		Date	20/6/2015

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Helped conceive study design, helped edit paper		
Signature		Date	24/06/2015

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Detection of Altered Bases in Ancient DNA Using SMRT Sequencing

Stephen M. Richards*, Oliver Lomax-James Wooley, and Alan Cooper

Australian Centre for Ancient DNA, University of Adelaide, South Australia, Australia

*Corresponding author
E-mail: steve.richards@adelaide.edu.au

Abstract

Recently developed single strand high throughput sequencing (SS-HTS) technologies have the potential to expand our understanding of the modified nucleotides present in ancient DNA (aDNA). Presently, aDNA is known to contain bases modified through decay and epigenetic mechanisms, however the knowledge of the abundances and categories of these altered bases remains rudimentary. Previous high throughput sequencing (HTS) platforms were not ideal for studying base modification because these technologies require amplification of study DNA, which erases information on altered bases from the final sequencing data. In this study, aDNA from a 26,000 year old steppe bison bone was analyzed with Single Molecule Real-Time (SMRT) sequencing in an attempt to identify the types and abundances of modified bases in the sample. SMRT technology sequences from a single strand of unamplified DNA and can identify modified bases present in the molecule. The steppe bison aDNA was sequenced in three SMRT Cell runs and the data processed and mapped to the cattle reference genome (*Bos taurus,* UMD 3.1*)* using the SMRT Analysis software package. A total of 282 modified nucleotides were called, exhibiting an apparent location bias as 49.3% of the altered nucleotides were found in sequences mapping to chromosomes 1 and 2 of the cattle genome. Despite Pacific Biosciences Analysis software being able to call bases as modified, the current algorithms were not able to identify the specific types of alterations present. We show that although SMRT sequencing is likely to become a useful tool in the analysis of aDNA, additional development of data analysis software will be crucial for the identification of true base modification.

Introduction

Since being introduced in 2005, high throughput sequencing (HTS), such as Illumina's sequencing by synthesis, has transformed the field of ancient DNA (aDNA). HTS sequences from thousands of clonal DNA molecules in parallel, generating orders of magnitude more data from an ancient DNA extract than possible with Sanger sequencing. The massive increase in sequencing capacity of HTS has revolutionized aDNA research, allowing investigators to generate genomes of extinct

organisms such as a Neanderthal and an archaic horse [1,2]. Single strand high throughput sequencing (SS-HTS) technologies, such as Pacific Bioscience's Single Molecule Real-Time (SMRT), now bring additional promise to aDNA research by allowing the sequencing of single unamplified DNA molecules. With the elimination of an amplification step in library preparation, SMRT sequencing may have the capacity to detect bases modified through decay or epigenetic mechanisms in raw aDNA [3].

Post-mortem damage is a universal characteristic of aDNA. A living organism maintains a complex network of repair proteins to ensure the integrity of DNA. After death, this repair network ceases to function and DNA starts to decay, first through biological mechanisms and then by long-term chemical modification [4,5]. As a consequence of these biological and chemical attacks, aDNA is fragmented into short sequences that contain various forms of damage such as intra- or inter-strand cross-links and abasic sites. Modified bases are another type of damage commonly found in aDNA that can be a serious concern in aDNA research [6]. Gas chromatography/mass spectrometry analysis has detected damage-modified bases in aDNA including 5-hydroxy-5-methylhydantoin (5-OH-5-MeHyd), 5-hydroxyhydantoin (5-OH-Hyd), and 8-hydroxyguanine (8-OH-Gua) whilst enzymatic methods have also shown the presence of uracil produced by deamination of cytosine [7,8]. Some damaged bases including 5-OH-5-MeHyd and 5-OH-Hyd act as blocking lesions that stop synthesis by DNA polymerases and consequently do not appear in sequencing data [7]. In contrast, 8-OH-Gua and uracil are miscoding lesions that cause certain DNA polymerases to misincorporate an incorrect base during DNA synthesis, which can be mistaken for genetic variation [9,10]. Current knowledge of the types and abundances

of damaged bases in aDNA is still limited. Identification of the types of damaged bases present in aDNA would help elucidate the factors influencing the decay of DNA over time. Furthermore, additional knowledge on *post-mortem* damage could contribute to the development of aDNA repair protocols or analytical software that may improve sequencing data accuracy [6] and maximize library yield [11].

Epigenetic modifications are a group of heritable biochemical changes to DNA or chromatin-associated proteins that influence phenotype whilst leaving the underlying nucleotide sequence of DNA unchanged [12,13]. In mammals, 5-methylcytosine (5-mC) is the most common epigenetic modification that can alter phenotype through regulating gene activity [14]. Epigenetic modifications are thought to be a mechanism that permits rapid adaptation to environmental stresses and these biochemical alterations have therefore become a great interest to biologists studying evolution [15]. In the past, there have been very abrupt climate changes, such as the rapid and severe cooling of the Younger Dryas event which occurred 12.8-11.5 thousand years before present (kyr BP), which would have placed extreme adaptive pressure on organisms in time frames as short as a decade [16]. Standard mutation of DNA and selection processes are unlikely to be able to allow populations to adapt to these abrupt climatic changes and instead, swift epigenetic modifications may have permitted organisms to adapt to these new environmental conditions [13]. Obtaining the epigenetic patterns from aDNA in parallel with genomic sequences, will allow researchers to better understand how past organisms adapted to environmental pressures.

Sequencing technologies that require amplification, such as Illumina's sequencing by synthesis, are not ideal methods for identifying modified bases in DNA. Miscoding lesions can be difficult to identify because PCR amplifies from both strands of a DNA molecule thus preventing the strand of origin of a particular base change from being determined [9,17]. Additionally, DNA polymerases are incapable of transmitting methylation information into the newly synthesized DNA strand during PCR, leading to the erasure of epigenetic modifications [18]. However, there are some technologies that permit detection of modified bases with amplified DNA. Bisulfite treatment can be used to produce cytosine methylation patterns with single nucleotide resolution from amplified modern DNA or aDNA [12,19]. Single cytosine methylation patterns of amplified aDNA can also be determined using an enzyme treatment protocol [20]. However, at this time, only a limited number of base modifications can be identified using amplified DNA and in some cases similar alterations cannot be resolved. For example, bisulfite treatment is not able to discriminate between 5-mC and 5-hydroxy-methylcytosine (5-hmC), another epigenetic modification present in mammalian genomes [21,22].

Presently, there are several platforms that can detect modified bases through single strand sequencing. These technologies can identify a wider range of alterations than is currently possible with other methods and have the potential to greatly expand our understanding of base modification in aDNA [3]. Previously, a Helicos Genetic Analysis System was used to sequence single strands of unamplified horse aDNA [23], but this platform did not have the capability to identify modified bases. All current SS-HTS platforms that are capable of identifying modified bases use one of two methodological principles: nanopore sequencing or SMRT technology.

In nanopore sequencing, a single strand of DNA is threaded through a pore in a linear fashion. As each base passes through the pore a signal, such as current change, is generated in a nucleotide specific manner. Nanopore sequencing has been reported to resolve different methylation states of cytosine [24,25]. At the time of this writing, all nanopore sequencers were in the prototype stage of development.

SMRT technology produced by Pacific Biosciences is currently the only commercially available SS-HTS platform that can detect base modification through single strand sequencing. SMRT is an optically based technology that uses sequencing by synthesis to generate data [26]. In SMRT sequencing, DNA molecules are first converted into circular templates called SMRTbells through ligation of hairpin adapters. Primers are annealed to the loops of the hairpin adapter and the SMRTbells are then bound to DNA polymerases that are subsequently immobilized on a solid support. DNA synthesis is initiated using fluorescently labeled canonical bases that produce a procession of fluorescent pulses during polymerization, which are recorded as sequencing data. Because the SMRTbell is a circular template it can be sequenced repeatedly; each repeated sequencing of an individual strand of the original DNA template is called a subread. As well as the identity of the nucleotides, the time taken for the DNA polymerase to add one nucleotide to the next is also measured and is called the interpulse duration (IPD). A modified base will produce a different IPD in comparison to a native nucleotide by changing the kinetics of the DNA polymerase. For each base in a SMRTbell molecule, an average IPD is calculated using the data from all the subreads generated. The averaged IPD is then compared to a standard or a control IPD generated *in silico* from a reference sequence to produce an IPD ratio, which is used in conjunction with other parameters to call a

base as modified. Depending on the number of subreads generated, certain modified bases can be identified. Some modifications produce subtle changes in IPD ratio and require deep subread coverage for identification whilst other alterations have substantial impacts on DNA polymerase kinetics and need only moderate subread coverage [27,28]. For example, 5-mC requires 250x subread coverage for identification whilst 8-OH-Gua needs only 25x [3].

In the current study, unamplified aDNA from a 26,000 year old steppe bison (*Bison priscus*) bone was converted to a SMRTbell library and shotgun sequenced on a Pacific Biosciences RS sequencer. The goal of the study was to identify the types and prevalence of modified bases in the steppe bison aDNA.

Methods

Sample

A steppe bison astragalus bone from a Late Pleistocene deposit in the Canadian Yukon Territory was the source of aDNA for this study. The bone was carbon dated at the Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom). Using a ^{14}C half-life of 5,568 years, accelerator mass spectrometry produced an uncalibrated age of 26,360 ± 220 radiocarbon kyr BP. No permits were required for the described study, which complied with all relevant regulations.

Reagents and Materials

Dremel (CA, USA): Dremel Tool, carborundum cutting wheel; Braun (Hesse, DE): mikro-dismembrator; Sigma-Aldrich (NSW, AU): N-lauroylsarcosine, Triton x-100, silica, ethanol, sodium acetate; Life Technologies (Vic, AU): proteinase K, 0.5M EDTA pH 8.0, H_2O molecular biology grade, 1 M Tris pH 8.0; Qiagen (Vic, AU): QG buffer, MinElute columns; Intergraded DNA Technology (IA, USA): Pacific Biosciences hairpin adapter (CA, USA).

aDNA Extraction

All library preparation was performed in a cleanroom designated for aDNA work at the Australian Centre for Ancient DNA (University of Adelaide, South Australia, Australia). Using a Dremel tool with a carborundum cutting disk, several sections of the astragalus bone were removed and subsequently pulverized to powder in a Braun mikro-dismembrator. aDNA was extracted from the bone powder using a protocol based on previously published methods [29,30]. In brief, two grams of bone powder were divided equally among 8 x 15 mL conical test tubes each containing 4.44 mL of extraction buffer (0.5M EDTA pH 8.0; 0.05% N-lauroylsarcosine; and 0.25mg/mL proteinase K) and the tubes were incubated overnight at 37°C on a rotor. Undissolved bone powder was then pelleted with centrifugation at 4,600 rpm for 5 minutes and each supernate was transferred to a separate 50 ml conical tube containing 125 µL silica particles and 16 mL binding buffer (13.5 mL QG buffer, 1.0% (v/v) Triton x-100, 20mM NaCl, 0.2M sodium acetate). aDNA was bound to the silica by rotating the tubes at room temperature overnight. The silica particles were then pelleted by centrifugation at 4,600 rpm for 5 minutes and the supernate was discarded. Each silica pellet was transferred to a 1.5 mL micro-centrifuge tube and washed 3 times by

resuspending the particles in 1 mL 80% ethanol followed by centrifugation at 15,000 rpm for 1 minute. After the last wash was discarded, the open tubes were placed in an incubator set to 37°C for 30 minutes to dry the silica. Silica from each tube was suspended in 200 µL TE buffer and then incubated at 37°C for 15 minutes to release the aDNA. The silica was then pelleted and the supernates, containing aDNA, were transferred to O-ring 1.5 mL tubes for storage at -20°C.

SMRTbell Preparation and Sequencing

The extracted aDNA was combined into four pools of 400 µL. Each pool was concentrated with a MinElute spin column following the provided PCR cleanup protocol and eluted with 20 µL EB + 0.05% Tween-20. Each pool was subsequently taken through a ligation protocol to attached Pacific Biosciences hairpin adapters to the aDNA [30]. The completed ligation reactions were subsequently combined to form a single SMRTbell library. Portions of the library were run in three different SMRT Cells on a RS Sequencer using the standard sequencing mode.

Data Analysis

Analysis was performed using SMRT Analysis software v2.0 (https://github.com/PacificBiosciences) with quality filters applied to retain sequences with a minimum polymerase read length of 50 and minimum polymerase read quality of 0.75. Filtered reads were then mapped to the cattle (*Bos taurus)* reference genome UMD 3.1 as subread clusters with the BLASR algorithm that maps to genomes by finding the highest scoring local alignment. After mapping, subread clusters were examined for single nucleotide polymorphisms (SNPs) and indels using the Quiver algorithm. Mapped subread clusters were also scanned for modified bases using the

Base Modification Detection Algorithm that identified putative sites of base alteration through analysis of DNA polymerase kinetics or IPD ratio and a minimum score of 2.0 for IPD ratio was required to mark a base as possibly modified. The IPD ratio was calculated by comparing the IPD produced by an individual base to an *in silico* control IPD generated from the local reference sequence to which the subread cluster mapped. Mapped data was imported into a stand-alone version of the SMRT View software v2.0.0 to visualize the subread clusters on the reference genome. SMRT View also assigned a modification quality score (Qv) to each base and a Qv score in the range of 65 to 256 was necessary to call a residue as modified. A consensus sequence was generated from each of the individual subread clusters and the number of bases between the modification and the 5' end of the consensus sequence was determined.

Results

Raw sequence data consisted of 1,302,375,455 bases from 901,752 polymerase reads with an average polymerase read length of 1,444 bases and an average polymerase read quality of 0.405. After the application of quality filters the sequence data was comprised of 1,218,430,427 bases from 418,632 polymerase reads with an average polymerase read length of 2,911 bases and an average polymerase read quality of 0.853. Mapping of the filtered data produced 15,534 subread clusters (321,751 individual subreads) with an average subread length of 82 base pairs and average quality of 0.92. Mapped steppe bison subread clusters exhibited no SNPs or indels in comparison to the cattle reference genome.

After visualization of the processed data, subread clusters were found to have mapped to the cattle reference in three different manners:

- Individual cluster generated with subreads from both strands of a single SMRTbell (Figure 1A)
- Individual cluster generated with subreads from only one strand of a single SMRTbell (Figure 1B)
- Stacks of subread clusters generated from multiple SMRTbells (Figure 2)

In total, 282 bases were positively called as modified in the steppe bison sequence data. The number of modifications for each nucleotide was as follows: A= 20, C=72, G=61, and T=129. Of the total, 254 modified bases were called in the stacks of subread clusters, which all mapped to repetitive regions of the reference (Table 1). In contrast, modifications found in subread clusters generated from a single SMRTbell did not map to repetitive regions of the reference. Although, the number of subread clusters mapping to the cattle genome was evenly distributed amongst the different chromosomes, base modification showed a large location bias with clusters mapping to chromosomes 1 and 2 containing 49.3% of the total number of altered residues. Base modifications found in data from single SMRTbell molecules did not appear to have any chromosome bias but this may be a stochastic effect as there were so few sequences of this type (Table 1).

Modified bases were called in a large range of local nucleotide contexts but a few motifs produced multiple modified bases. In particular, a large number of poly-T subread clusters mapped to a long AT stretch of chromosome 2 and contained 30.5% of the total modified bases (Table 2). Although only a few are shown, the data

presented in Table 2 is typical of the modified bases called in the stacks of subread clusters: moderate Qv scores, large but variable subread coverage, and low IPD ratios.

Twenty-eight base modifications were detected in subreads from single SMRTbell molecules that mapped to the reference and these alterations were found on 18 different chromosomes. Each chromosome had one to three modified bases and only five of the 28 modified bases were found in subread data from both strands of bison aDNA. Modified bases from these single SMRTbell molecules produced IPD ratios in the range of 3.8 to 25.6 and low modification quality scores ranging from 65 to 97 Qv. Subread coverage for the modifications found in single SMRTbell data was between 12 and 44. The local sequence context was different for all single SMRTbell modifications and there appeared to be a bias against A nucleotides as only three modifications were detected for this base (Table 3). Modifications from single SMRTbell data were randomly distributed along the lengths of consensus sequences with no apparent location preference (Table 3).

Although the SMRT Analysis programs were able to call bases as modified in the steppe bison aDNA data, for unclear reasons the software was not able to determine the specific types of alterations present.

Discussion

The current study applied a new SS-HTS technology, specifically Pacific Bioscience's SMRT sequencing, to an unamplified steppe bison aDNA sample.

Certain SS-HTS platforms, including SMRT sequencing, have the potential to revolutionize the study of aDNA because these technologies can identify base modification. While the sequencing of this sample proved successful, the mapping of the SMRT data generated with steppe bison aDNA was problematic for several reasons. For some SMRTbells, subreads from just one strand of the aDNA mapped to the reference (Figure 1B) and this loss of data was likely the result of nucleotide damage. The SMRT Analysis software actually maps the subreads from each strand of a SMRTbell independently. For subread clusters with missing data, one strand of the aDNA molecule may have been so badly damaged that it generated subreads that could not be mapped and the SMRT Analysis software subsequently discarded these data (Marco Milletti, Pacific Biosciences, personal communication). An additional mapping problem was the large stacks of subread clusters aligned to repetitive regions of the reference (Figure 2). SMRT software is geared to processing the long reads (mean length > 3,000 base pairs) that can be produced with the RS Sequencer and the mapping algorithm may not have been able to accurately process the short low-complexity subreads generated from the steppe bison aDNA. This inaccurate processing by the SMRT software may have caused the short low-complexity subreads to be mapped together and contributed to the location bias of modified bases towards chromosomes 1 and 2 (Marco Milletti, Pacific Biosciences, personal communication).

For the SMRT steppe bison data the average read size was 82 base pairs, contrasting with a previous study that reported that the majority of unamplified aDNA from this sample was molecules < 50 base pairs in length [9]. Several factors likely contributed to the longer read lengths in the SMRT data. First, the minimum read length allowed

in the SMRT analysis package is 50 base pairs; consequently many short sequences were lost during data filtering. Second, smaller SMRTbell/DNA polymerase complexes may have loaded with less efficiency on to the solid support for sequencing. Lastly, the SMRT mapping software was developed for long sequences and may not have been able to effectively map data from many of the small aDNA SMRTbells.

Unexpectedly no SNPs or indels were detected in the steppe bison sequences and it is not clear why these variations are missing from the data. These variations may have rendered short aDNA sequences unmappable and were discarded by the SMRT software.

Identification of the types of base modifications present in the steppe bison aDNA was likely unsuccessful for several reasons. The SMRT software in some cases requires deep subread coverage to identify modified bases and the modest coverage produced in this study may not have met this requirement [31]. Additionally, identification of modified bases using an *in silico* control IPD, as done in the current study, is dependent on the certainty that the subread cluster has been accurately mapped within the reference. In the steppe bison data, this certainty was low because the subreads were short (average 82 base pairs) whilst the cattle reference was over 3 billion base pairs (Marco Milletti, Pacific Biosciences, personal communication). Although there was deep coverage in the stacks of subread clusters, the probability that these low-complexity sequences had been correctly placed in the reference was extremely low. Despite the lack of identification, the modifications in at least the

single SMRTbell data appear to be true residue alterations, as they show strong influence on DNA polymerase kinetics producing IPD ratios up to 25.6 (Table 3).

Damage, particularly deaminated cytosine, in aDNA is more frequent at, though not exclusive to, the ends of molecules [32]. The predominance of damage at the ends of aDNA is because these regions tend to be single stranded [33] and more vulnerable to chemical attack [34]. In the steppe bison SMRT data, base modifications were distributed along the length of the consensus sequences and did not appear to have any location preference. In the single SMRTbell data only four of the 28 modifications were found within five base pairs of the ends of the aDNA consensus sequences (Table 3), but this number may not be representative of the alterations present in single stranded regions of the unprocessed steppe bison aDNA. The SMRT library construction protocol used in this study, similar to other HTS methods, included a repair step that used T4 DNA polymerase to blunt end the aDNA prior to ligation of the hairpin adapters. T4 DNA polymerase has an endonuclease activity that removes 3' single stranded overhang and a polymerase activity that fills in 5' single-stranded overhang. Consequently, only modified bases found near the 5' ends of the consensus sequence data were likely from single stranded regions of the unprocessed aDNA[1]. In the single SMRTbell data, only one modification was found within five base pairs of the 5' end of a consensus sequence. Since Phi29, the polymerase used in SMRT sequencing [26] can read through uracil [35], an increased frequency of base modification should be apparent in the ends of the steppe bison data. However, such

[1] In ancient HTS data such as the sequencing produced by the Illumina platform, uracil retained by the 5' overhang fill-in activity of T4 DNA polymerase will cause miscoding at both ends of reads in the final sequencing data. After amplification, incorrect bases will be present in both strands of the amplicons produced from templates containing uracil. Depending on which strands of the amplicons are sequenced, the original uracil will cause a 3' C → T or 5' G→A transition in the final sequencing data [9].

an increase was not observed and may have been lost in the subreads discarded by the SMRT analysis software.

Although unsuccessful in the current study, identification of altered bases in aDNA with SMRT sequencing may be possible with additional experiments. Pacific Biosciences has recently released an upgraded SMRT system that increases the number of SMRTbells sequenced during a run and triples read length, which would aid in identification of modified bases in aDNA. Although, costly and time consuming, the types of modification could also be identified using synthetic standards. In this study, IPD ratios were determined by comparing a measured change in enzyme kinetics to a control IPD calculated by the SMRT software using the local sequence of the reference. Synthetic copies of aDNA SMRTbell sequences could be manufactured with known modifications in the place of the altered bases detected in the aDNA. These synthetic oligonucleotides would then be sequenced with SMRT technology and the IPD generated by the known modifications matched against the corresponding altered base in the aDNA [36].

Ultimately, new analytical pipelines will likely have to be developed if SMRT sequencing is to be used to identify modified bases in aDNA. SMRT Analysis will need to be tailored so that the subreads of each strand of a SMRTbell can be easily tracked and that stacks of consensus sequences can be separated into data from individual SMRTbells. This study highlights the need for the development of alternative software for mapping short aDNA reads instead of the long sequences for which the SMRT software was designed. It may also be necessary to create an alternative algorithm to generate *in silico* IPDs that does not rely on the certainty of

placing small subreads within a large reference. Given the relatively short period of time that SS-HTS has been available, the utility of sequencing true aDNA base composition, and the current open-source development strategy of the SMRT analysis software (hosted on github), a dramatic expansion in the range of tools available for this type of sequencing will likely occur.

Although this study only examined aDNA, it can serve as a model for applying SMRT sequencing to other areas of research that involve degraded DNA. For example, many medical facilities have collections of normal and diseased formaldehyde-fixed and paraffin-embedded (FFPE) tissues, which have huge potentials for genetic studies. DNA from FFPE tissues contains damage similar to that found in aDNA including fragmentation, crosslinks, and miscoding lesions [37,38]. For genetic studies of FFPE tissues to reach full potential, DNA from these specimens will have to be analyzed with technologies like SMRT sequencing to help differentiate damage caused by the fixing and embedding processes from pathogenic mutations. Furthermore, identification of methylated bases in FFPE tissue will be important as altered epigenetic modification of DNA may have a role in the etiology of diseases such as cancer [39].

Acknowledgements:
The authors would like to thank Jonas Korlach and Marco Milletti at Pacific Biosciences for sequencing the ancient bison DNA and for their helpful discussion on data analysis. We also thank the miners of the Yukon Territory, especially the Johnson family, for assistance in collecting ancient vertebrate bones. Lastly, we acknowledge the Yukon Heritage Branch, especially Grant Zazula for field assistance.

Author Contributions:

Conceived and designed experiments: SMR AC. Performed experiments: SMR. Analyzed data: SMR OLJW. Wrote paper SMR. Edited paper: OLJW AC. Provided ancient sample: AC.

References

1. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft Sequence of the Neandertal Genome. Science 328: 710-722.
2. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, et al. (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499: 74-78.
3. Korlach J, Turner SW (2012) Going beyond five bases in DNA sequencing. Current Opinion in Structural Biology 22: 251-261.
4. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.
5. Paabo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, et al. (2004) Genetic analyses from ancient DNA. Annual Review of Genetics 38: 645-679.
6. Dabney J, Meyer M, Pääbo S (2013) Ancient DNA Damage. Cold Spring Harbor Perspectives in Biology 5: 1-6.
7. Hoss M, Jaruga P, Zastawny TH, Dizdaroglu M, Paabo S (1996) DNA Damage and DNA Sequence Retrieval from Ancient Tissues. Nucleic Acids Research 24: 1304-1307.
8. Paabo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 1939-1943.
9. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, et al. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35: 5717-5728.
10. Parlanti E, Fortini P, Macpherson P, Laval J, Dogliotti E (2002) Base excision repair of adenine/8-oxoguanine mispairs by an aphidicolin-sensitive DNA polymerase in human cell extracts. Oncogene 21: 5204-5212.
11. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.
12. Llamas B, Holland ML, Chen K, Cropley JE, Cooper A, et al. (2012) High-Resolution Analysis of Cytosine Methylation in Ancient DNA. PLoS ONE 7: 1-6.
13. Klironomos FD, Berg J, Collins S (2013) How epigenetic mutations can affect genetic evolution: Model and mechanism. BioEssays 35: 571-578.
14. Bell AC, Felsenfeld G (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature 405: 482-485.
15. Franks SJ, Hoffmann AA (2012) Genetics of Climate Change Adaptation. Annual Review of Genetics 46: 185-208.
16. Alley RB, Marotzke J, Nordhaus WD, Overpeck JT, Peteet DM, et al. (2003) Abrupt Climate Change. Science 299: 2005-2010.
17. Hansen A, Willerslev E, Wiuf C, Mourier T, Arctander P (2001) Statistical evidence for miscoding lesions in ancient DNA templates. Molecular Biology and Evolution 18: 262-265.
18. Bart A, van Passel MW, van Amsterdam K, van der Ende A (2005) Direct detection of methylation in genomic DNA. Nucleic Acids Research 33: 1-6.
19. Bird A (2007) Perceptions of epigenetics. Nature 447: 396-398.

20. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, et al. (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research 38: 1-12.
21. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, et al. (2010) The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. PLoS ONE 5: e8888.
22. Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, et al. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324: 930-935.
23. Orlando L, Ginolhac A, Raghavan M, Vilstrup J, Rasmussen M, et al. (2011) True single-molecule DNA sequencing of a pleistocene horse bone. Genome Research 21: 1705-1719.
24. Manrao EA, Derrington IM, Pavlenok M, Niederweis M, Gundlach JH (2011) Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore. PloS ONE 6: e25723.
25. Wanunu M, Cohen-Karni D, Johnson RR, Fields L, Benner J, et al. (2010) Discrimination of Methylcytosine from Hydroxy-methylcytosine in DNA Molecules. Journal of the American Chemical Society 133: 486-492.
26. Eid J, Fehr A, Gray J, Luong K, Lyle J, et al. (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. Science 323: 133-138.
27. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, et al. (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proceedings of the National Academy of Sciences of the United States of America 105: 1176-1181.
28. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, et al. (2010) Real-Time DNA Sequencing from Single Polymerase Molecules. Methods in Enzymology: Academic Press. pp. 431-455.
29. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. Biotechniques 42: 343-352.
30. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
31. Clark TA, Lu X, Luong K, Dai Q, Boitano M, et al. (2013) Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biology 11: 4.
32. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, et al. (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.
33. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.
34. Lindahl T, Nyberg B (1974) Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry 13: 3405-3410.
35. Serrano-Heras G, Bravo A, Salas M (2008) Phage phi 29 protein p56 prevents viral DNA replication impairment caused by uracil excision activity of uracil-DNA glycosylase. Proceedings of the National Academy of Sciences of the United States of America 105: 19044-19049.

36. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature Methods 7: 461-465.
37. Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, et al. (2009) Genome-Wide Massively Parallel Sequencing of Formaldehyde Fixed-Paraffin Embedded (FFPE) Tumor Tissues for Copy-Number- and Mutation-Analysis. PLoS ONE 4: e5548.
38. Do H, Dobrovic A (2012) Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase. Oncotarget 3: 546-558.
39. Korah R, Healy JM, Kunstman JW, Fonseca AL, Ameri A, et al. (2013) Epigenetic silencing of RASSF1A deregulates cytoskeleton and promotes malignant behavior of adrenocortical carcinoma. Molecular Cancer 12: 87.

Figure 1A.

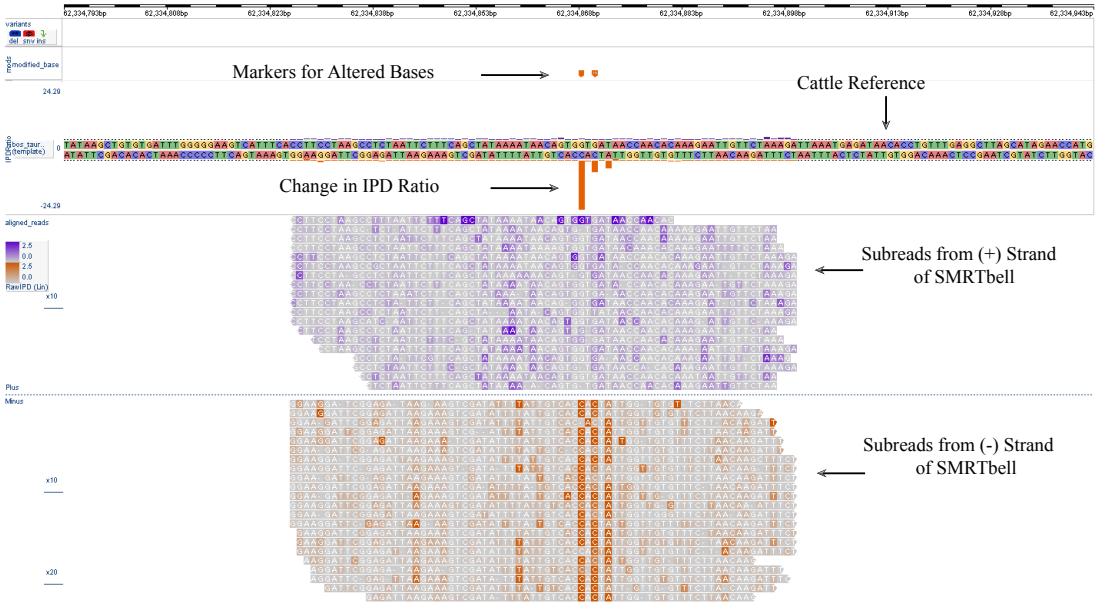


Figure 1B.

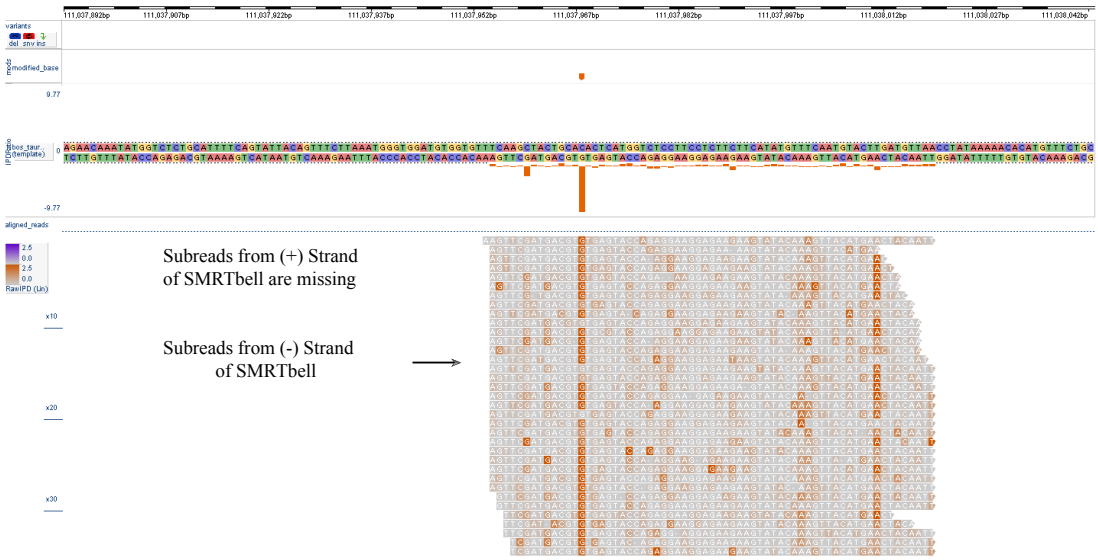


Figure 1. Images generated with SMRT View showing modified bases detected in the subreads clusters of single SMRTbells

Figure 1A: Modified bases detected in sequence data where the subreads of both strands of a single SMRTbell mapped to the reference. The subreads from each strand of the SMRTbell are depicted in orange and purple. The orange bars descending from the reference sequence and the dark orange highlight of bases in the subreads show the increased IPD ratios of bases. In this image, three bases show increased IPD ratios but only two met all the requirements to be called modified by the SMRT Analysis software.

Figure 1B: A modified base detected in sequenced data where subreads from one strand of the SMRTbell did not mapped to the reference and may have been discarded by the analysis software.

Figure 2. An image generated in SMRT View showing modified bases detected in the subreads of multiple SMRTbells

This image is an example where the subreads from multiple SMRTbells mapped to the cattle reference in stacks. In the shown subread stacks, two modified bases were called and the dark highlight shows that the modifications were detected in various locations in the subreads from the different SMRTbells.

Table 1. Distribution of subread clusters and modified bases

Cattle Reference Genome UMD 3.1	[1]Number of SMRTbell Clusters Mapping to Cattle Reference Chromosomes	[2] Total Number of Modified Bases Called Using a Modification Quality Score of 65 to 256 Qv	[3] Total Number of modified Bases Called using Subreads from a Single SMRTbell
Chr1	316	42	1
Chr2	685	97	2
Chr3	607	16	0
Chr4	604	4	2
Chr5	605	3	1
Chr6	597	13	2
Chr7	563	14	3
Chr8	560	5	1
Chr9	528	4	0
Chr10	521	4	2
Chr11	535	4	1
Chr12	455	1	1
Chr13	421	13	3
Chr14	423	2	1
Chr15	427	1	1
Chr16	408	0	0
Chr17	375	0	0
Chr18	660	11	2
Chr19	640	0	0
Chr20	360	3	0
Chr21	357	2	2
Chr22	614	0	0
Chr23	525	1	1
Chr24	627	26	1
Chr25	429	1	1
Chr26	516	0	0
Chr27	454	3	0
Chr28	463	7	0
Chr29	515	5	0
ChrX	744	0	0

Chromosomal location of steppe bison subread clusters and modified bases in SMRT shotgun sequencing data mapped to the cattle genome reference (UMD 3.1). The vast majority of modified bases were called in stacks of subreads produced from multiple SMRTbells (Figure 2).

1- Subreads are sequences generated from a strand of DNA in a SMRTbell molecule. In a normal sequencing run subreads are generated from both strands of DNA in a SMRTbell and the SMRT analysis pipeline maps the subreads as clusters to a reference.

2- In the mapped subread clusters, the SMRT analysis pipeline generates a modification (Qv) score to call bases as modified. In the current study, bases with a Qv score of 65 to 256 were called as modified.

3- In this study, a minority of bases called as modified were found in subread data produced from a single SMRTbell molecule.

Table 2. Examples of local sequence context that produced multiple altered bases in steppe bison aDNA

[1]Local Sequence Context of Modification	Chromosome Location Of Modifications	Number of Modified Modified Bases in Chromosome	[2]Average Modification Qv Score	[3]Average Subread Coverage	[4]Average IPD Ratio
AGC**T**GGC	chr1	5			
	chr6	2			
	chr9	1	115.1 ± 53.2	566.6 ± 480.5	2.5 ± 0.6
	chr24	4			
	chr28	2			
ATG**A**GCA	chr1	6	133.6 ± 101.1	325.2 ± 313.7	2.5 ± 0.2
	chr18	2			
GGT**G**TCG	chr1	6			
	chr6	1			
	chr7	4	88.5 ± 15.4	146.5 ± 107.6	3.3 ± 1.1
	chr11	2			
	chr24	6			
	chr28	3			
TTT**T**TTT	chr2	86	98.4 ± 16.0	430.7 ± 72.3	2.0 ± 0.2

In the steppe bison SMRT sequencing data, low complexity subread clusters are mapped in stacks to repetitive regions of the cattle reference genome. The local sequence contexts for most of the modified bases found in subread stacks were unique but a few motifs produced multiple modified bases.

1- The local sequence context containing modified bases as determined by the reference genome to which the subreads mapped. Modified bases are indicated by the bold and underlined nucleotides.
2- Qv is a base modification quality score generated by the SMRT Analysis software to call bases as modified. In this study a Qv range of 65 to 256 was used to call bases as positively modified.
3- Subread coverage is the number of subreads that contained a modified base.
4- Interpulse duration (IPD) ratio is a metric of the degree to which a modified base has changed the kinetics of the DNA polymerase during sequencing. An IPD ratio of 2.0 or more was necessary to call a nucleotide as modified.

Table 3. Characteristics of modified bases called in subreads from single SMRTbells

	[1]Both Strands of SMRTbell Mapped	[2]Modification Qv Score	[3]Subread Coverage	[4]IPD Ratio	[5]Local Sequence Context	[6]Consensus Sequence Length	[7]Distance of Modified Base from 5' End of Consensus Sequence
Chr1	No	74	24	5.8	TGT**T**CCT	65 bp	28 bp
Chr2	No	65	30	4.1	TCT**G**TGC	75 bp	42 bp
Chr2	No	97	34	9.8	AGT**G**TGC	65 bp	13 bp
Chr4	No	76	34	5.5	AGG**C**ACG	99 bp	66 bp
Chr4	No	86	26	7.9	TTT**C**AGC	107 bp	55 bp
Chr5	Yes	66	18	9.2	CAC**T**TCA	68 bp	**67 bp**
Chr6	No	66	27	9.7	AAA**G**TAT	67 bp	16 bp
Chr6	Yes	69	20	24.3	TCA**C**CAC	73 bp	43 bp
Chr7	Yes	66	20	3.6	TCG**A**CTG	78 bp	14 bp
Chr7	No	69	31	17.5	ACA**G**GAC	102 bp	71 bp
Chr7	Yes	66	17	16.0	TAG**A**TTT	97 bp	37 bp
Chr8	Yes	67	12	21.4	ACC**A**AAT	73 bp	64 bp
Chr10	No	71	37	3.8	TTT**C**CCG	81 bp	37 bp
Chr10	No	69	18	25.6	TCA**C**TGA	110 bp	35 bp
Chr11	No	65	30	4.1	ACT**C**AGT	75 bp	26 bp
Chr12	No	85	44	7.4	AGT**T**GGT	58 bp	9 bp
Chr13	No	65	32	4.6	GCT**G**CAG	110 bp	39 bp
Chr13	No	70	38	4.4	AAA**G**TGA	63 bp	**62 bp**
Chr13	No	66	26	5.0	TCG**C**AGG	118 bp	56 bp
Chr14	No	74	37	8.6	TCA**T**CAA	60 bp	40 bp
Chr15	No	67	18	12.2	ATA**G**AGC	86 bp	16 bp
Chr18	No	65	20	12.7	TAT**C**CGT	97 bp	51 bp
Chr18	No	67	29	11.4	GCA**G**AAC	88 bp	**87 bp**
Chr21	No	69	29	4.8	TGT**T**TCC	87 bp	**2 bp**
Chr21	No	66	18	11.7	GAG**C**GTG	87 bp	54 bp
Chr23	No	86	20	15.9	TCA**C**AGA	83 bp	62 bp
Chr24	No	65	20	12.0	ATA**C**CTG	80 bp	56 bp
Chr25	No	67	19	5.7	CCC**T**CAT	76 bp	42 bp

In the steppe bison SMRT sequencing data, 28 modified bases were called in subread clusters from a single SMRTbell molecule. Only five of the modified bases identified in single SMRTbell data were called using subreads from both strands of the aDNA molecules.

1- This column indicates whether the subread cluster containing the modified base included subreads from both strands of the aDNA in the SMRTbell (Figure 1A and 1B).
2- Qv is a base modification quality score generated by the SMRT Analysis software to call bases as modified. In this study a Qv range of 65 to 256 was used to call bases as positively modified.
3- Subread coverage is the number of subreads that contained a modified base.
4- Interpulse duration (IPD) ratio is a metric of the degree to which a modified base has changed the kinetics of the DNA polymerase during sequencing. An IDP ratio of 2.0 or more was necessary to call a base as modified.
5- The local sequence context containing modified bases as determined by the reference genome to which the subreads mapped. Modified bases are indicated by the bold and highlighted nucleotides.
6- The length in base pairs (bp) of the consensus sequence generated from all the subreads produced by a SMRTbell.
7- The location of the modification as determined by the number of base pairs from the 5' end of the subread consensus sequence. Bases in red were found within 5 bases of the 5' or 3' ends of the consensus sequence.

Statement of Authorship

Title of Paper	Paleoclimatic Impacts on European Bovid Megafauna in the Late Pleistocene
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to Molecular Ecology

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards
Contribution to the Paper	Helped conceive study design, helped perform experiments, helped analyze data, helped write paper
Signature	Date: June 18, 2015

Name of Co-Author	Julien Soubrier
Contribution to the Paper	Helped conceive study design, helped perform experiments, helped analyze data, helped write paper
Signature	Date: 22.06.15

Name of Co-Author	Michael S. Y. Lee
Contribution to the Paper	Helped analyze data, helped edit paper
Signature	Date: 22-06-15

Name of Co-Author	Alan Cooper
Contribution to the Paper	Helped conceive study design, helped write paper
Signature	Date: 24/06/2015

Statement of Authorship

Title of Paper	Paleoclimatic impacts on European bovid megafauna in the Late Pleistocene
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ● Publication style
Publication Details	Written for submission to Molecular Ecology

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards		
Contribution to the Paper			
Signature		Date	

Name of Co-Author	Simon Y.W. Ho		
Contribution to the Paper	Helped analyzed data, helped edit paper		
Signature		Date	20-JUL-14

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Paleoclimatic Impacts on European Bovid Megafauna in the Late Pleistocene

Stephen M. Richards [†]*, Julien Soubrier[†]*, Simon Y.W. Ho[‡], Michael S. Y. Lee[¥], Alan Cooper[†]

[†]*Australian Centre for Ancient DNA, University of Adelaide, Adelaide, SA 5005, Australia*
[‡] *The University of Sydney, NSW 2006, Australia*
[¥] *South Australian Museum, North Terrace, South Australia 5000, Australia*
**These authors contributed equally to this manuscript*
Corresponding author: Stephen M. Richards
email: steve.richards@adelaide.edu.au

ABSTRACT: A wide diversity of fossil bison taxa are known from Eurasia and North America during the late Pleistocene, although only two species remain alive today: the American bison, *Bison bison*, and the European bison, *Bison bonasus*. Ancient DNA has previously been used to study the evolutionary history of North American and Beringian bison in the late Pleistocene, revealing dramatic population changes around the Last Glacial Maximum. In contrast, the genetic diversity of European bison throughout this period is yet to be explored. Here, we extract and analyze ancient DNA from 48 European late Pleistocene bovid samples, collected from the North Sea, the Caucasus, Urals, and Austria. Phylogenetic analyses reveal a previously unknown species of bison (here termed *Bison X*), which is a sister taxon to the European bison. In addition, rapid changes in ecological dominance of the new species and steppe bison across Europe identified major environmental shifts dating back to beyond Oxygen Isotope Stage 3, ca. 55 kyr BP.

Keywords: bison, bovids, ancient DNA, mitochondria, climate change, phylogeography

Introduction

The late Pleistocene (126-11 kyr BP) climate record contains a series of large-scale oscillations with pronounced environmental effects (Martin 1984; Wolff *et al*. 2010). There is great interest in the potential role of these environmental changes in the widespread extinction of mammalian megafauna (>42kg) around the world during this time (Martin 1984; Graham *et al*. 1996; Stuart & Lister 2007). A key issue is how to separate these impacts from those caused by the arrival of human populations as they expanded around the world towards the end of the Pleistocene (50-11 kyr BP). The nature and extent of these impacts, and the degree to which they were synergistic, remains debated (Graham *et al*. 1996; Barnosky *et al*. 2004, 2011; Barnosky & Lindsey 2010; Lorenzen *et al*. 2011; Stewart & Stringer 2012; Rule *et al*. 2012).

The paleontological and archeological record of Europe provides a rich source of data for examining the timing and nature of megafaunal extinctions. Large datasets of radiocarbon-dated fossil mammal bones are available (Stuart & Lister 2007, 2012), and multiple cave environments exist that are suitable for DNA preservation. Detailed climate records are available, ranging from Greenland ice cores (GISP2) to lake, pollen, and paleovegetation records (Sirocko *et al*. 2013; Helmens 2014). Ancient mitochondrial DNA (mtDNA) studies have identified large-scale genetic changes in European megafaunal populations over time, ranging from cave lions (Barnett *et al*. 2009), cave bears (Orlando *et al*. 2002; Stiller *et al*. 2010), cave hyenas (Rohland *et al*. 2005; Sheng *et al*. 2014), horses (Lorenzen *et al*. 2011), mammoths (Barnes *et al*. 2007; Nyström *et al*. 2010), saiga antelope (Campos *et al*. 2010), bison (Shapiro *et al*. 2004) and Neanderthals (Dalén *et al*. 2012).

Late Pleistocene bovid fossils form one of the most complete European megafaunal records, with several species currently recognised. These include the extinct aurochs (*Bos primigenius*) and two species of bison: the extinct steppe bison (*Bison priscus*), which ranged from Alaska to western Europe and the extant but endangered wisent, or European bison (*Bison bonasus*), whose fossils appear around the Pleistocene/Holocene boundary (ca. 11 kyr BP) or later. The bovid fossil record is somewhat complicated by the similarity between the bison species and by the difficulty in identifying the taxonomic affinities of bovid post-cranial elements.

These taxa have widely varying histories. The aurochs is generally accepted as the ancestor of modern cattle, which were independently domesticated from ancient populations of aurochs in different parts of the world. Evidence from mitochondrial

63 DNA (mtDNA) suggests that late Pleistocene populations of aurochs in northern
64 Europe did not contribute to the genetic diversity of European cattle (*Bos taurus*). In
65 contrast, populations in southern Europe shared mtDNA lineages with modern *B.*
66 *taurus*, although it is unclear whether this was due to a direct genetic contribution to
67 the domestication process, or shared ancestry (Lari *et al*. 2011). While the history of
68 the aurochs has been well studied genetically (Beja-Pereira *et al*. 2006; Edwards *et al*.
69 2007, 2010; Lari *et al*. 2011; Bollongino *et al*. 2012), there is little known about that
70 of the bison.
71
72 Studies of ancient DNA have revealed that late Pleistocene populations of steppe
73 bison in the Old World constitute a small subset of the total mtDNA diversity
74 observed in eastern Beringia (Alaska and Yukon) at this time (Shapiro *et al*. 2004).
75 This is surprising, given that the fossil record indicates that steppe bison originated in
76 Asia and only recently colonised the New World, ca. 200-400 kyr BP (McDonald
77 1980). To explain this, it has been suggested that the earliest populations of the Old
78 World underwent at least one major extinction phase, or were replaced by the re-
79 invading Alaskan population by some other means prior to ca. 80 kyr BP (Shapiro *et*
80 *al*. 2004). The last steppe bison in the Old World are thought to have gone extinct in
81 the late Holocene, although the exact date is unknown (Lazarev *et al*. 1998).
82
83 All living European bison are descended from 17 animals originating from two small
84 late-19th/early-20th Century populations (Slatis 1960). Modern populations now
85 exceed more than 2000 individuals, of which all pure-bred European bison represent
86 the recombination of only 12 diploid sets of genes (Slatis 1960). Surviving
87 populations continue to be threatened by habitat destruction, inbreeding, disease, and

88 human hunting. The recent population bottleneck in European bison has made it
89 difficult to reconstruct their genetic history. Separate Caucasus and Carpathian
90 subspecies of the European bison (*B. b. caucasicus and B. b. hungarorum*) have been
91 recognised from the fossil record, but went extinct in the wild in 1927 and 1790,
92 respectively and are of uncertain status (Krasińska & Krasiński 2013). A male *B. b.*
93 *caucasicus* was also used as a founding member of the *B. bonasus* breeding program.
94
95 A close evolutionary relationship between the European bison and American bison
96 (*Bison bison*) has been suggested by morphological similarities, autosomal data, and
97 the ability to interbreed to produce fertile female offspring. However, phylogenetic
98 analyses of mitochondrial data depict a surprisingly different evolutionary history, in
99 which European bison are more closely related to cattle than to American bison
100 (Janecek *et al*. 1996; Verkaar *et al*. 2004). Two hypotheses have been proposed to
101 explain the difference between the autosomal and mitochondrial phylogenies (Verkaar
102 *et al*. 2004): incomplete lineage sorting, or sex-biased genetic introgression between
103 steppe bison males and an ancestral bovid in the ox-zebu lineage, resulting in the
104 appearance of a new species with bison-like morphology and autosomal DNA, but
105 ox/zebu-like mitochondrial DNA.
106
107 The paucity of early, clearly identifiable *B. bonasus* fossil remains further complicates
108 this issue. The oldest *B. bonasus* specimens are from the late Pleistocene to early
109 Holocene (Flerov 1979; Kahlke 1999), or even the late Holocene (Pucek 1986), and
110 there has been little agreement about ancestral forms (Stuart 1991; Kahlke 1999;
111 Bauer 2001). It has been suggested that the direct ancestor of the European bison
112 might be the Pleistocene forest bison *Bison schoetensacki* (*B. schoetensacki*) (Kurtén

1968; Geist & Karsten 1977), although some have also suggested *B. priscus* (Flerov 1979; Bauer 2001).

To clarify the evolutionary history of bovids in Europe and to investigate the role of environmental change in megafaunal extinctions, we analyse ancient DNA sequences from fossil remains of bison collected from four separate locations across Europe (the Caucasus, the North Sea, the Urals, and Austria) spanning a period of more than 60 kyr.

Methods

Samples

Samples from a total of 68 late Pleistocene bison bones were collected from four regions across Europe (Table 1). Northeastern Europe was the region that provided the main set of samples and represents isolated bones excavated from a wide variety of cave deposits throughout the Ural Mountains and surrounding areas. These samples were held in collections at the Zoological Museum of the Institute of Plant and Animal Ecology (ZMIPAE) in Ekaterinburg, Russia. The second region where samples were collected was western Europe in late Pleistocene deposits on the North Sea bed. These North Sea specimens have little stratigraphic information and were recovered by trawling operations and are curated by the North Sea Network (NSN) in the Netherlands. The third region where samples were obtained was southeastern Europe, exclusively from excavations at Mezmaiskaya Cave in the Caucasus Mountains. This high-altitude site contains both Neanderthal and early *Homo sapiens* remains (Skinner *et al*. 2005), and samples were acquired from the Laboratory of

Prehistory, in St Petersburg, Russia. The last collection region was central Europe, which consisted of several Holocene sites and these specimens were held in collections of the Natural History Museum, Vienna, Austria. In all cases, samples were identified as bison or bovid post-cranial bone because cranial material is rare for this time period.

Ancient DNA Extraction and Amplification

Ancient DNA was extracted from 200-500 mg powdered bone using phenol/chloroform/centrifugal filtration methods (Shapiro *et al*. 2004) in a geographically isolated ancient DNA laboratory. An approximate 600 bp fragment of the mitochondrial control region was amplified in one to four (overlapping) fragments, depending on the quality of the DNA preserved in the specimen. Either single (Shapiro *et al*. 2004) or two-step multiplex PCR amplifications were performed using primers designed for the bovid mitochondrial control region. Multiplex primer sets A and B were set up separately (Table S3). Multiplex PCR was performed in a final volume of 25 µL containing 2 µL of aDNA extract, 1 mg/mL rabbit serum albumin fraction V (RSA; Sigma-Aldrich, Dorset, UK), 6 mM MgSO4, 0.2 µM of each primer, 500 µM of each dNTP, 2 U Platinum *Taq* Hi-Fidelity and 1 × PCR buffer (Life Technologies, Paisley, UK). Multiplex PCR conditions were initial denaturation at 95 °C for 2 minutes, followed by 35 cycles of 94 °C for 15 seconds, 55 °C for 20 seconds and 68 °C for 30 seconds, and a final extension at 68 °C for 10 minutes at the end of the 35 cycles. Multiplex PCR products were then diluted to 1:10 as template for the second step of simplex PCR.

The simplex PCR, using Amplitaq Gold (Life Technologies) or Hotmaster *Taq* DNA

polymerase (5Prime, Milton, QLD) was conducted in a final volume of 25 µL containing 1 µL of diluted multiplex PCR product, 2.5 mM MgCl2, 0.4 µM of each primer, 200 µM of each dNTP, 1 U Amplitaq Gold/ Hotmaster *Taq* polymerase and 1 × PCR buffer. The PCR conditions were initial denaturation at 95 °C for 2 minutes, followed by 35 cycles of 94 °C for 20 seconds, 55 °C for 15 seconds and 72 °C for 30 seconds, and a final extension at 72 °C for 10 minutes at the end of the 35 cycles. Multiple PCR fragments were cloned to evaluate the extent of DNA damage and within-PCR template diversity. Multiple samples were also independently replicated in both the Henry Wellcome Ancient Biomolecules Centre at the University of Oxford and at the University of Adelaide.

One-step simplex PCR amplifications using Platinum *Taq* Hi-Fidelity polymerase were performed on a heated lid thermal cycler in a final volume of 25 µl containing 1 µL of aDNA extract, 1mg/ml RSA, 2 mM MgSO4, 0.6 µM of each primer, 250 µM of each dNTP, 1.25 U Platinum *Taq* Hi-Fidelity and 1 × PCR buffer (Life Technologies). The conditions of PCR amplification were initial denaturation at 95 °C for 2 minutes, followed by 50 cycles of 94 °C for 20 seconds, 55 °C for 20 seconds and 68 °C for 30 seconds, and a final extension at 68 °C for 10 minutes at the end of the 50 cycles. Negative extraction controls and non-template PCR controls were used in all experiments. PCR products were then checked by electrophoresis on 3.5-4.0% agarose TBE gels, and visualized after ethidium bromide staining on a UV transilluminator. PCR amplicons were purified using AMPure magnetic beads (Beckman Coulter, Gladesville, NSW) according to the manufacturer's instructions.

Amplicon Sequencing

All purified PCR products were bi-directionally sequenced with the ABI Prism BigDye Terminator Cycle Sequencing Kit version 3.1 (Life Technologies). The sequencing reactions were performed in a final volume of 10 µl containing 3.2 pmol of primer, 0.25 µl BigDye terminator premixture, and 1.875 µl of 5 × sequencing buffer. The reaction conditions included initial denaturation at 95 °C for 2 minutes, 25 cycles with 95 °C for 10 seconds, 55 °C for 15 seconds, and 60 °C for 2 minutes 30 seconds. Sequencing products were purified using CleanSEQ magnetic beads (Beckman Coulter) according to the manufacturer's protocol. All sequencing reactions were analysed on an ABI 3130 DNA capillary sequencer (Life Technologies).

Amplification of Template for In Vitro Transcription (IVT) of Probe for Whole Mitochondrial Genome Enrichment

To provide deeper phylogenetic resolution and further examine the apparent closer relationship between *B. taurus* and *B. bonasus* mitochondria, full genome sequences of four *Bison X* specimens were generated using hybridisation capture with RNA probes transcribed from long range PCR products of *B. taurus* mitochondrial DNA. The Primer-Blast program at NCBI (http://www.ncbi.nlm.nih.gov/tools/ primer-blast/) was used to design primers to amplify the *B. taurus* mitochondrial genome (NC_006853.1) in three overlapping sections: mito-1 (6568 bp), mito-2 (6467 bp), and mito-3 (5390 bp). Primer pairs were designed with a high melting temperature to permit amplification with 2-stage PCR and the T7 RNA promoter was attached to the 5' end of one primer from each pair (Cone & Schlaepfer 1997; Gnirke *et al*. 2009; Ichijo *et al*. 2008). Amplification of each mitochondrial section was performed using a heated lid thermal cycler in multiple PCRs containing 1x Phire Buffer, 25 ng calf

thymus DNA, 200 µM dNTPs, 500 nM forwards and reverse primers (Table S4), 0.5 µL Phire Hot Start II DNA polymerase (Thermo Fisher, Scoresby, VIC), and H2O to 25 µl. The mito-1 and mito-2 sections were amplified with a thermal cycler program of 1 cycle: 98°C for 30 seconds; 26 cycles: 98°C for 10 seconds and 72°C for 70 seconds; and 1 cycle: 72°C for 180 seconds whilst the program for mito-3 was 1 cycle: 98°C for 30 seconds, 28 cycles: 98°C for 10 seconds and 72°C for 60 seconds, and 1 cycle: 72°C for 180 seconds. After amplification, 2 µL of each PCR was agarose gel electrophoresed and the product visualized with Gel-Red (Biotium, Hayward, CA) staining and UV illumination. Amplification of mito-1 and mito-2 produced a single band and the PCRs for these mitochondrial sections were separately pooled and then purified with QiaQuick columns (Qiagen, Chadstone, VIC) following the provided PCR cleanup protocol. Amplification of mito-3 produced additional product and the correct size amplicon was size selected using gel excision followed by purification with QiaQuick columns using the gel extraction protocol. Purified amplicons from each mitochondrial section were quantified using a NanoDrop 2000 Spectrophotometer (Thermo Fisher).

Transcription of B. taurus Mitochondrial IVT Templates

Each of the three mitochondrial *IVT* templates were transcribed using a T7 High Yield RNA Synthesis Kit (New England Biolabs, MA, USA) in multiple reactions containing 150-200 ng purified amplicon, 1x Reaction Buffer, 10 mM rNTPs, 2 µL T7 enzyme mix, and H2O to 20 µL. The *IVT* reactions were incubated for 16 hours at 37°C and then the DNA template was destroyed by incubating for an additional 15 minutes at 37°C with 2U Turbo DNase (Life Technologies). *IVT* reactions for each mitochondrial section were separately pooled and purified with Megaclear spin

columns (Life Technologies) except that H2O was used to elute the RNA instead of the provided Elution buffer because the Elution buffer was found to inhibit RNA fragmentation in the next step. The integrity of the RNA was verified on an acrylamide gel and the mass quantified with a NanoDrop 2000 Spectrophotometer.

Fragmentation of Mitochondrial IVT RNA

RNAs from the *IVT* transcription were fragmented with a NEBNext Magnesium RNA Fragmentation Module (New England Biolabs) in reactions that contained 1x Fragmentation buffer, 45 µg RNA, and H2O to 20 µL. Reactions were incubated at 94°C for 10 minutes and fragmentation stopped with the addition of 2 µL Stop Buffer. After fragmentation, each reaction was purified with a RNeasy MinElute spin column (Qiagen) by following the provided cleanup protocol except for the final elution. To elute, 20 µL H2O was pipetted into the column and the column was heated at 65°C for 5 minutes and then centrifuged at 15,000 g for 1 minute. The flow-through was transferred to a 1.5 mL tube and stored at -80°C. The fragmented RNA was quantified on a NanoDrop 2000 Spectrophotometer and 100 ng was visualized on an acrylamide gel producing a smear in the range of 80-300 bases.

Biotinylation of Fragmented RNA

Biotinylation was performed in several reactions containing 6.7 µg each of mito-1, mito-2, and mito-3 fragmented RNA, 40 µL Photoprobe Long Arm reagent (Vector Laboratories, Burlingame, CA), and H2O to 80 µL in 200 µL PCR tubes. The tubes were placed in a 4°C gel cooling rack and then incubated under the bulb of a UV sterilization cabinet for 30 minutes. Organic extractions were performed on the labeling reactions by adding 64 µL H2O, 16 µL 1 M Tris buffer, and 160 µL sec-

butanol to each tube and shaking vigorously for 30 seconds followed by centrifugation for 1 minute at 1,000 g. The upper organic layers were discarded and the extraction repeated with an additional 160 µL sec-butanol. After the second organic layers were discarded, the remaining aqueous phases were purified with RNeasy MinElute spin columns following the provided reaction cleanup protocol but with a modified elution procedure described in the previous step. Elutions were pooled and then quantified with a NanoDrop Spectrophotometer 2000 and the RNA, which will now be called probe, was stored at -80°C in 5 µL aliquots at 100 ng/ µL.

Repetitive Sequence Blocking RNA

RNA to block repetitive sequences in bison aDNA was transcribed from Bovine Hybloc DNA (Applied Genetics Laboratories, Melbourne, FL), a commercially produced Cot-1 DNA, using a published linear amplification protocol (Liu *et al*. 2005). Briefly, the Hybloc DNA was polished in a reaction containing T4 polynucleotide kinase (New England Biolabs) and T4 DNA polymerase (New England Biolabs) and purified with MinElute spin columns (Qiagen) following the PCR cleanup protocol provided. Poly-T tailing was performed on the polished DNA with terminal transferase and a tailing solution containing 92 µM dTTP (New England Biolabs) and 8 µM ddCTP (Affymetrix, Santa Clara, CA). After tailing, the Hybloc DNA was purified with MinElute spin columns as before. The Hybloc DNA was then heat denatured and T7-A18B primers (Table S4), containing the T7 RNA polymerase promoter, were allowed to anneal to the poly-T tails on the Hybloc DNA with slow cooling. A second-strand synthesis reaction was then performed on the Hybloc DNA using DNA polymerase I Klenow fragment (New England Biolabs) and the product was purified with MinElute spin columns using the provided PCR cleanup

protocol. The double stranded Hybloc DNA was transcribed using a T7 High Yield RNA Synthesis Kit in multiple reactions containing 75 ng DNA, 1x Reaction Buffer, 10 mM rNTPs, 2 µL T7 enzyme mix, and H2O to 20 µL. *IVT* reactions were incubated for 16 hours at 37°C and then the DNA template was destroyed by adding 2U Turbo DNase and incubating for an additional 15 minutes at 37°C. The RNA was purified with RNeasy MinElute spin columns as in the RNA fragmentation step above. Purified RNA was quantified on a NanoDrop 2000 and 100 ng visualized on an acrylamide gel, which produced a smear 80 to 500 bases in length. This blocking RNA will hereafter be called Hyblock RNA.

Bison X Sequencing Library Construction and Amplification

See Supplemental Methods: Whole Mitochondrial Genome Hybridization Capture.

Primary Mitochondria Hybridization Capture

Truncated versions of Illumina sequencing libraries were used for hybridization capture because full-length adapters reduce enrichment efficiency (Rohland & Reich 2012). For the primary hybridization capture, three Reagent Tubes were prepared for each bison library with the following materials: Reagent Tube #1- 3.5 µL of 35-55 ng/µL WEA2 library (WEA2 in Supplemental Methods Step 4); Reagent Tube #2 - 5 µL probe, 1 µL HyBloc RNA, and 0.5 µL of a stock containing 50 µM of both P5_short_RNAblock and P7_short_RNAblock (Table S4); Reagent Tube #3 - 30 µL Hybridization Buffer: 75% formamide, 75 mM HEPES, pH 7.3, 3 mM EDTA, 0.3% SDS, and 1.2 M NaCl (Konietzko & Kuhl, 1998). The P5/P7 short_RNAblock are RNA oligonucleotides that are complementary to the library adapters and were included in the incubation to prevent reannealing of the adapters. Hybridization

capture was performed in a heated lid thermal cycler programed as follows: Step 1- 94°C for 2 minutes, Step 2- 65°C for 3 minutes, Step 3- 42°C for 2 minutes, Hold 4- 42°C hold. To pre-warm reagents for hybridization capture, the Reagent Tubes were placed in the thermal cycler at the start of each program Step in the following order: Step 1- Reagent Tube #1; Step 2- Reagent Tube #2; Step 3- Reagent Tube #3. For each library, once the Hold cycle started 20 µL of hybridization buffer from Reagent Tube #3 was mixed with the RNA in Reagent Tube #2. The entire content of Reagent Tube #2 was then pipetted into Reagent Tube #1 and mixed with the bison library to begin the hybridization capture. Hybridization capture was carried out at 42°C for 48 hours. Magnetic streptavidin beads (New England Biolabs) were washed and blocked, just prior to the end of the hybridization capture incubation. For each library, 50 µL of beads were washed twice using 0.5 mL Wash Buffer 1 (2x SSC+0.05% Tween-20) and a magnetic rack. The beads were then blocked by incubation in 0.5 mL Wash Buffer 1+ 100 µg yeast tRNA (New England Biolabs) for 30 minutes on a rotor. Blocked beads were washed once as before and then suspended in 0.5 mL Wash Buffer. At the end of the hybridization capture, each reaction was added to a tube of blocked beads and incubated at room temperature for 30 minutes on a rotor. The beads were then taken through a series of stringency washes as follows: Wash 1 - 0.5 mL Wash Buffer 1 at room temperature for 10 minutes; Wash 2 - 0.5 mL Wash Buffer 2 (0.75x SSC + 0.05% Tween-20) at 50°C for 10 minutes; Wash 3 - 0.5 mL Wash Buffer 2 at 50°C for 10 minutes; Wash 4 - 0.5 mL Wash Buffer 3 (0.2x SSC + 0.05% Tween-20) at 50°C for 10 minutes. After the last wash, the captured library was released from probe by suspending the beads in 50 µL of Release buffer (0.1 M NaOH) and incubating at room temperature for 10 minutes. The Release buffer was then neutralized with the addition of 70 µL Neutralization buffer (1 M Tris-HCl pH

7.5). The captured library was then purified with MinElute columns by first adding 650 µL PB buffer and 10 µL 3 M sodium acetate to adjust the pH for efficient DNA binding. The library was then purified using the provided PCR cleanup protocol and eluting with 35 µL EB+0.05% Tween-20.

Primary Hybridization Capture Amplification

Amplification of each primary hybridization capture was performed in five PCRs containing 5 µL of library from the primary hybridization capture, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM each of primers IS7_short_amp.P5 and IS8_short_amp.P7 (Table S4), 0.25 U Phusion Hot Start II DNA polymerase (New England Biolabs), and H2O to 25 µL. The captured libraries were amplified and processed in the same manner as for WEA1 in Supplemental Methods Step 4.

Secondary Mitochondria Hybridization Capture

Library from the Primary Hybridization Capture Amplification step was taken through a second round of enrichment using the methods outlined in the Primary Mitochondria Hybridization Capture step.

Secondary Hybridization Capture Amplification

Indexed primers were used to convert the DNA recovered from the Secondary Mitochondria Hybridization Capture to full length Illumina sequencing libraries. Each library was amplified in three PCRs containing 5 µL library from the secondary hybridization capture, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM each of primers GAII_Indexing_*x* (library specific index) and IS4 (Table S4), 0.25 U Phusion Hot Start II DNA polymerase, and H2O to 25 µL. Amplification was performed in a

heated lid thermal cycler programed as follows 1 cycle: 98°C for 30 seconds; 10 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. Libraries were purified and processed in the same manner as WEA1 in Supplemental Methods Step 4.

Sequencing of Enriched Mitochondrial Libraries

The purified full length Illumina libraries were sent to the ACRF South Australia Cancer Genomics Facility for sequencing on an Illumina HiSeq.

NGS Data Analysis

Sequence data were analysed through a bioinformatics pipeline combining various publicly available software packages: FastX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/) was used to sort reads by barcodes, using a strict zero mismatches threshold (--bol --mismatches 0). Cutadapt v1.2.1 (Martin 2011) was used to trim adapter sequences using a maximum error rate of 0.33 (-e 0.3333), and to remove short (-m 25 bp), long (-M 130 bp) and low quality sequences (-q 20), with a total of five passes (-n 5). The filtered reads were checked with FastQC (http://www.bioinformatics.bbsrc.ac.uk/ projects/fastqc). The reads were mapped against the *B*. *bonasus* mitochondrial genome (GenBank: Q223450) using BWA 0.6.2 (Li & Durbin 2009). Mapped reads with mapping quality below Phred 30 and read duplicates were removed using Samtools v0.1.18 (Li *et al*. 2009) and the MarkDuplicates tool of Picard Tools v1.79 (http://picard.sourceforge.net). GC content of mapped reads was analysed using the CollectGcBiasMetrics tool of Picard Tools v1.79.

Radiocarbon Dating

All samples from which bison mitochondrial control region sequences were successfully amplified were sent for accelerator mass spectrometry (AMS) radiocarbon dating (except for seven samples from level 3 of the Mezmaiskaya cave, which were expected to be older than AMS dating capabilities). The dating was performed by the AMS facility at the Oxford Radiocarbon Accelerator Unit at the University of Oxford, the INSTAAR Laboratory for AMS Radiocarbon Preparation and Research (NSRL) at the University of Colorado at Boulder, and the Keck-Carbon Cycle AMS facility (KCCAMS) at the University of California, Irvine and the results are shown in Table 1. The calibration of radiocarbon dates was performed using OxCal v4.1 with the IntCal09 curve (Reimer *et al*. 2011). Dates reported here are given in ca. kyr BP unless otherwise stated.

Phylogenetic Analysis

To provide comparative data, we assembled 302 published sequences of the 628 bp region of the mitochondrial control region (Table S1), representing the following bovid mitochondrial lineages: European bison, American bison, steppe bison, zebu (*Bos indicus*), and cattle. Among these published sequences, 5 were from steppe bison collected in the Urals (Shapiro et al. 2004, Table 1). Whole mitochondrial genome sequences were also available for 28 bovid specimens: 2 European bison, 8 American bison, 4 yaks (*Bos grunniens*), 3 zebus, 8 cattle, and 3 buffalo (*Bubalus bubalis*) (Table S2).

Genetic Identification of the New Specimens

A phylogenetic tree of the 302 published, and 48 new, control region sequences was constructed using both maximum-likelihood and Bayesian methods: the HKY+G model of nucleotide substitution was selected by comparison of Bayesian information criterion (BIC) scores in ModelGenerator v0.85 (Keane *et al*. 2006). A maximum-likelihood analysis was performed with the program PhyML v3 (Guindon *et al*. 2010), using NNI and SPR rearrangements to search for the tree topology and using approximate likelihood-ratio tests to establish the statistical support of internal branches. Two independent Bayesian analyses were performed using the program MrBayes v3.2.1 (Ronquist *et al*. 2012). Posterior estimates of parameters were obtained by Markov chain Monte Carlo sampling, with samples drawn every 1000 steps. We used four Markov chains, comprising one cold and three heated chains, each of 10 million steps. The first 50% of samples were discarded as burn-in before the majority-rule consensus tree was calculated.

The same methods were used to estimate the corresponding bovid phylogeny using the aligned 28 published and 4 newly sequenced *Bison X* mitochondrial genomes. The HKY+I model of nucleotide substitution was selected through comparison of BIC scores.

Estimation of Evolutionary Timescale

In order to estimate the evolutionary timescale, we used the program BEAST v1.6.2 (Drummond & Rambaut 2007) to conduct a Bayesian phylogenetic analysis of all radiocarbon-dated *Bison X* samples. The GMRF skyride model (Minin *et al*. 2008) was used to account for the population demographic history and a strict clock was

440 assumed. Replicate analyses, using a relaxed uncorrelated lognormal clock to account
441 for potential rate variation among lineages (Drummond *et al*. 2006), provided support
442 for a strict molecular clock.
443
444 Calibrated radiocarbon dates associated with the sequences were used as calibration
445 points. Some samples appear to be older than 55 kyr: one from the Urals, four from
446 the North Sea, and four from the Caucasus (Table 1). Because these dates have
447 infinite error margins, we allowed them to vary in the analysis by treating them as
448 distinct parameters in the model (Shapiro *et al*. 2011). The dated samples from
449 Mezmaiskaya Cave are from stratigraphic layer 2B, which lies on top of layer 3-4,
450 and the latter has been dated at a maximum age of ca. 78 kyr BP (including error
451 margins). Consequently, for each Caucasian sample, we specified a lognormal prior
452 age distribution (mean = 8,000) with a minimum bound of 40 kyr and with 95% of the
453 prior probability less than 80 kyr. A similar prior distribution (mean = 26,000) was
454 used for the five remaining samples that had infinite radiocarbon dates, with a 95%
455 prior probability lower than 150 kyr.
456
457 All parameters showed sufficient sampling (indicated by effective sample sizes above
458 200) after 50,000,000 steps, with the first 10% of samples discarded as burn-in. In
459 addition, a date-randomization test was conducted to check whether the temporal
460 signal from the radiocarbon dates associated with the ancient sequences was sufficient
461 to calibrate the analysis (Figure S1, Ho *et al*. 2011). This test randomizes all dates and
462 determines whether the 95% HPD interval of the rate estimated from the date-
463 randomized analysis includes the mean rate estimated from the original data set.
464

Survey of Temperature and Paleovegetation Record

The calibrated AMS radiocarbon dates and posterior distributions of sample ages were compared to paleotemperature records from the North Greenland Ice Core Project (NGRIP; Wolff *et al*. 2010), and to paleovegetation reconstructions for the Eifel region of SW Germany (Sirocko *et al*. 2013) and northern and central Europe (Helmens 2014).

Morphological Comparisons

Morphological measurements (greatest length of shaft, and width of shaft at narrowest point) of metacarpal bones were collected for specimens from: the Urals (4 *B*. *priscus*; 4 *Bison X* and 1 *B*. *bonasus* (Drees & Post, 2007)), the North Sea (18 *B*. *priscus* and 14 small bison morphologically similar to the genetically identified *Bison X* samples), Germany (40 *B*. *bonasus* (Koch 1932)), Latvia (10 *B*. *bonasus* (Zalkin 1958)), and Belarus (8 *B*. *bonasus*).

To study the morphological variability and allometry between *Bison X*, Steppe bison and European bison, lengths were plotted against width of shaft and differences between taxa assessed via standardised major axis regression using SmatR version 2 (https://github.com/dfalster/smatr/). First, regression lines for taxa were compared for slope differences; for taxa with no significant slope differences, lines were regressed using a common slope to allow differences in intercept (elevation) to be assessed.

Results and Discussion

Ancient DNA Mitochondrial Typing

Mitochondrial control region sequences (>600 bp) were successfully amplified from 53 out of 68 samples analysed. Three samples produced a mixture of cattle and bison amplification products (Table 1), these were identified as contaminated and removed from all analyses. Sequences from two individuals did not match bovid haplotypes and were identified as brown bear and elk in BLAST searches (Table 1). This appears to be due to the morphological misidentification of postcranial elements.

Position of New Samples in the Bovid Mitochondrial Phylogeny

The phylogenetic analysis of the 48 new bovid control region sequences (Figure 1) allowed us to characterize mitochondrial haplotypes from five European bison, four cattle, eight steppe bison, and 31 individuals belonging to a new clade we have designated *Bison X*. The four cattle sequences came from one Caucasus and three North Sea samples. Repetition and cloning experiments suggested that these are genuine rather than recent contaminants, and indeed the haplotypes were rare or previously unknown (Figure 1). The *Bison X* clade forms a monophyletic sister clade to European bison and has high statistical support from both likelihood and Bayesian analyses. A deep genetic split is observed between European bison and *Bison X*, indeed deeper than that observed between yak and steppe/American bison, and between Zebu and cattle. Such a deep genetic split is likely to reflect a species-level separation, although this is difficult to test in an extinct species.

Our study confirms that the European bison was rare in Europe prior to the Holocene, raising the possibility that *Bison X* might represent ancestral genetic diversity within

515 the *B*. *bonasus* lineage, or perhaps the short-horned Pleistocene European forest
516 wisent *B*. *schoetensacki*, a putative ancestor of *B*. *bonasus*. However, ancient *B*.
517 *bonasus* sequences are clearly contemporaneous and genetically distinct from *Bison X*
518 (Figure 1). For instance, specimens were identified from the Urals at 20 kyr BP, and
519 in Mezmaiskaya cave in the Caucasus with a date >56.3 kyr BP. The latter sample is
520 the oldest European bison specimen described, and demonstrates that *Bison X* and
521 European bison were contemporaneous in the Caucasus around 50-60 kyr BP.
522 Furthermore, while the ancient *B*. *bonasus* sequences fall outside of the modern
523 mtDNA variation in European bison, the clade remains monophyletic and well
524 supported, and clearly distinct from *Bison X*.
525
526 A key issue is that the larger clade formed by European bison and *Bison X* appears to
527 be more closely related to the zebu, aurochs and cattle than the other bison species
528 (Figure 1). This topology is in direct contrast with the close morphological and
529 nuclear genomic relationship between European and American bison. Previous
530 analyses of short mitochondrial sequences have also found a grouping of European
531 bison with cattle and zebu (Janecek *et al*. 1996; Verkaar *et al*. 2004). To further
532 evaluate support for the contrasting mitochondrial and nuclear topologies, we
533 analysed 28 whole mitochondrial genome sequences from American bison, European
534 bison, cattle, zebu, yak, buffalo and four newly sequenced *Bison X* (Figure 2a). This
535 reveals strong support for the placement of European bison and *Bison X* as a sister
536 clade to zebu and cattle, confirming the control region based phylogeny.
537 Incomplete lineage sorting seems an unlikely explanation for the discordant
538 mitochondrial and nuclear topologies of the European bison (Verkaar *et al*. 2004),
539 given the number of speciation events between European bison and cattle on the

nuclear tree (gaur, banteng, yak, and American bison). On the other hand, one (or multiple) introgression event(s) between the ancestral lineages of cattle/zebu and *Bison X*/European bison also seems unlikely (Verkaar *et al*. 2004). The hybridisation event would have had to be ancient to explain the branch lengths leading to European bison and *Bison X*, and most likely with repeated sex-biased (male *Bison*, female cattle) events in order for the ancestral cattle mtDNA lineage to become fixed. However, modern cattle-European bison hybrids are not viable without medical intervention, and F1 males are infertile. Given the current evidence, neither of the proposed explanations for the discordant phylogenies seems likely. However, the uncertainty over higher-level bovid relationships does not alter the strong evidence for the distinctness of *Bison X*, and the sister-group relationship of these animals to European bison. Further support is provided by morphological analysis of metacarpal measurements (Figure 2b) with steppe bison, European bison and *Bison X* specimens showing a clear hierarchical relationship. The regression of allometrically scaled (width/length ratio) steppe bison metacarpals is clearly different to that of European bison (from several Eastern-European locations) and *Bison X*, which are similar but have quite different intercept values. These observations suggest that the steppe bison forms an outgroup to the *Bison X* and European bison pair, which could clearly be separated morphologically. *Bison X* appears to have represented a more squat and robust subspecies or species. Unfortunately, dietary isotopes of the differing bison species overlap, and provide no evidence for different behaviours (Figure S2).

Late Pleistocene Movements of Bison in Europe

The ecological dominance of *Bison X* and steppe bison appears to be strongly correlated with changes in climate and paleovegetation patterns, with *Bison X* more

565 common during forested periods. The chronological distribution (Figure 3) reveals
566 two periods of abrupt turnover and rapid replacement between the two species.
567 During the milder conditions at the beginning of MIS 3 (ca. 60-50 kyr BP, Figure 3),
568 *Bison X* individuals are observed in three locations (Urals, North Sea, Caucasus) to
569 the near-exclusion of other bovid taxa. After Greenland Interstadial 13 (GI13) ca. 50
570 kyr BP, during a period of rapid climate oscillations, steppe bison specimens
571 dominate the fossil record of the Urals and North Sea until around 32 kyr BP, when
572 they disappear permanently from the records of all locations. The transition date of 32
573 kyr BP (dotted line on Figure 3) corresponds to a second major change in climate, at
574 GI 5, and marks the beginning of an extended cold period, which is only very briefly
575 interrupted by GI 4 and GI 3, and extends through the Last Glacial Maximum (LGM).
576 *Bison X* fossils are observed in the Urals until the end of the LGM, *ca*. 14.5 kyr BP
577 (GI 1), and then disappear from the records of both the Urals and Caucasus. The
578 widespread temporal and geographic sampling suggests this pronounced pattern of
579 alternating dominance between populations of steppe bison and *Bison X* is not a
580 taphonomic artefact.
581
582 Reconstructed paleovegetation histories for different areas of Europe (Figure 3)
583 (Lapteva 2009; Helmens 2014) suggest the presence of a classic steppe flora (grass
584 steppe with scattered birch, pine and spruce) when the steppe bison is present in the
585 North Sea and Urals. In contrast, *Bison X* appears to have been present during periods
586 that were more characterised by pine and spruce/birch forests. Regional differences in
587 paleovegetation conditions are likely to have played a significant role, making the
588 apparent temporal correlation across widely different geographic areas surprising.
589 Although the record is far less complete in the North Sea and the Caucasus (owing to

590 taphonomy and ancient DNA preservation), the sampled area spans the western coast
591 of Europe to the border of Asia and the Near East.
592
593 When the *Bison X* populations reappear after a hiatus during the steppe conditions
594 between 50 and 32 kyr BP, they form a monophyletic group with moderate diversity,
595 consistent with a relatively recent common genetic origin. This suggests the
596 population had been constrained to a small size prior to this point, potentially in a
597 single refugium. To investigate this scenario further, we constructed a temporally
598 calibrated phylogeny of *Bison X* samples (Figure 4). Our results show that the genetic
599 diversity of *Bison X* at the end of MIS 4 (60 kyr BP) can be divided into two clades,
600 with the North Sea individuals forming a sister group to the rest, suggesting some
601 degree of phylogeographic structure. However, when *Bison X* reappears after the
602 hiatus at 32 kyr BP it has comparable genetic diversity but no clear phylogeographic
603 patterns suggesting a recent geographic expansion. *Bison X* is then well represented in
604 the Urals until 14.6 kyr BP and appears in the Caucasus at 14 kyr BP immediately
605 before GI 1 and the subsequent Younger Dryas. The time to the most recent common
606 ancestor of the MIS 3 *Bison X* diversity is estimated to be around 58 (45.8 – 74.9) kyr
607 BP (grey area in Figure 4). This date closely matches the ages of the last observed
608 MIS 4 *Bison X* individuals across all three locations, supporting the idea of a
609 population movement and contraction of forest-adapted *Bison X* towards a refugium
610 during the steppe dominated periods of MIS 3 in Europe.
611
612 It is interesting that none of the 44 bison specimens detected in the Urals and North
613 Sea is dated between 46 and 38 kyr BP. This gap occurs immediately after a
614 particularly cold stadial corresponding to Heinrich event 5a (50-48 kyr BP) and this

615 period has already been correlated with major losses in genetic diversity in several
616 diverse Eurasian mammals, including cave lions (Barnett *et al*. 2009), Neanderthals
617 (Dalén *et al*. 2012) and mammoths (Barnes *et al*. 2007; Gilbert *et al*. 2008). The
618 paleovegetation record from Germany (Figure 3) and from a separate study in Greece
619 (Müller *et al*. 2011) also show important alterations of the flora close to this date. The
620 synchrony of drastic perturbations in the temperature record, the vegetation record,
621 and the genetic diversity of large mammals (including hominids), suggests that
622 important climate change remodelled the megafaunal diversity of Eurasia around 50
623 kyr BP, well before the arrival of anatomically modern humans or the late Pleistocene
624 extinctions following the LGM (Barnosky *et al*. 2004; Lorenzen *et al*. 2011).
625
626 While it is not currently possible to assign a firm taxonomic identity to *Bison X* it is
627 clearly genetically and morphologically distinct from all living species. It is important
628 to note that morphological studies of specimens in Dmanisi, Georgia, and in
629 Voigtstedt and Ungermassfeld, Germany, have identified possible new bison species
630 from the early to mid Pleistocene, intermediate in size between *B*. *priscus* and *B*.
631 *schoetensacki* (van der Made 1999, Sher 1997). Consequently, both of these bison,
632 designated *Bison* cf. *voigtstedtensis* and *Bison menneri* respectively, appear to be
633 potential ancestors of *Bison X*.
634
635 **Acknowledgements**
636 We thank numerous museum curators and colleagues for assistance with samples and morphological
637 data. This work was funded by Australian Research Council grant DP0773602.
638

References

Barnes I, Shapiro B, Lister A *et al*. (2007) Genetic Structure and Extinction of the Woolly Mammoth, *Mammuthus primigenius*. *Current Biology*, **17**, 1072–1075.

Barnett R, Shapiro B, Barnes I *et al*. (2009) Phylogeography of lions (Panthera leo ssp.) reveals three distinct taxa and a late Pleistocene reduction in genetic diversity. *Molecular Ecology*, **18**, 1668–1677.

Barnosky AD, Koch PL, Feranec RS *et al*. (2004) Assessing the Causes of Late Pleistocene Extinctions on the Continents. *Science*, **306**, 70–75.

Barnosky AD, Lindsey EL (2010) Timing of Quaternary megafaunal extinction in South America in relation to human arrival and climate change. *Quaternary International*, **217**, 10–29.

Barnosky AD, Matzke N, Tomiya S *et al*. (2011) Has the Earth's sixth mass extinction already arrived? *Nature*, **471**, 51–57.

Bauer K (2001) *Wisent Bison bonasus (Linnaeus, 1758)*. Bundesministerium für Land- und Forstwirtschaft Umwelt und Wasserwirtschaft, Graz, Austria.

Beja-Pereira A, Caramelli D, Lalueza-Fox C *et al*. (2006) The origin of European cattle: Evidence from modern and ancient DNA. *Proceedings of the National Academy of Sciences*, **103**, 8113–8118.

Bollongino R, Burger J, Powell A *et al*. (2012) Modern Taurine Cattle Descended from Small Number of Near-Eastern Founders. *Molecular Biology and Evolution*, **29**, 2101–2104.

Campos PF, Kristensen T, Orlando L *et al*. (2010) Ancient DNA sequences point to a large loss of mitochondrial genetic diversity in the saiga antelope (*Saiga tatarica*) since the Pleistocene. *Molecular Ecology*, **19**, 4863–4875.

Cone RW, Schlaepfer E (1997). Improved in situ hybridization to HIV with RNA probes derived from PCR products. *Journal of Histochemistry & Cytochemistry*, **45**, 721-727.

Dalén L, Orlando L, Shapiro B *et al*. (2012) Partial Genetic Turnover in Neandertals: Continuity in the East and Population Replacement in the West. *Molecular Biology and Evolution*, **29**, 1893–1897.

Drees M, Post K. (2007). *Bison bonasus* from the North Sea, the Netherlands. *Cranium*, **24**, 48-52.

Drummond AJ, Ho SYW, Phillips MJ *et al*. (2006) Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, **4**, e88.

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.

Edwards CJ, Bollongino R, Scheu A *et al*. (2007) Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1377–1385.

Edwards CJ, Mage DA, Park SDE *et al*. (2010) A Complete Mitochondrial Genome Sequence from a Mesolithic Wild Aurochs (*Bos primigenius*). *PLoS ONE*, **5**, e9255.

Flerov CC (1979) *European Bison. Morphology, Systematics, Evolution, Ecology*. Nauka Publishers [in Russian], Moscow.

Geist V, Karsten P (1977) The wood bison in relation to hypothesis on the origin of the American bison. *Z. Säugetierk*, **42**, 119–122.

687 Gilbert MTP, Drautz DI, Lesk AM *et al*. (2008) Intraspecific phylogenetic analysis of
688 Siberian woolly mammoths using complete mitochondrial genomes.
689 *Proceedings of the National Academy of Sciences*,**105**, 8327-8332.
690 Gnirke A, Melnikov A, Maguire J *et al*. (2009). Solution hybrid selection with ultra-
691 long oligonucleotides for massively parallel targeted sequencing. Nature
692 Biotechnology, **27**, 182-189.
693 Graham RW, Lundelius EL, Graham MA *et al*. (1996) Spatial Response of Mammals
694 to Late Quaternary Environmental Fluctuations. *Science*, **272**, 1601–1606.
695 Guindon S, Dufayard J-F, Lefort V *et al*. (2010) New Algorithms and Methods to
696 Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of
697 PhyML 3.0. *Systematic Biology*, **59**, 307–321.
698 Helmens KF (2014) The Last Interglacial–Glacial cycle (MIS 5–2) re-examined based
699 on long proxy records from central and northern Europe. *Quaternary Science*
700 *Reviews*, **86**, 115–143.
701 Ho SYW, Lanfear R, Phillips MJ *et al*. (2011) Bayesian Estimation of Substitution
702 Rates from Ancient DNA Sequences with Low Information Content.
703 *Systematic Biology*, **60**, 366–375.
704 Ichijo T, Yamaguchi N, Tani K *et al*. (2008). 16S rRNA sequence-based rapid and
705 sensitive detection of aquatic bacteria by on-chip hybridization following
706 multiplex PCR. *Journal of Health Science*, **54**, 123-128.
707 Janecek LL, Honeycutt RL, Adkins RM *et al*. (1996) Mitochondrial Gene Sequences
708 and the Molecular Systematics of the Artiodactyl Subfamily Bovinae.
709 *Molecular Phylogenetics and Evolution*, **6**, 107–119.
710 Kahlke, R-D (1999). *The history of the origin, evolution and dispersal of the late*
711 *Pleistocene Mammuthus-Coelodonta faunal complex in Eurasia (Large*
712 *Mammals)*. Fenske Media, Rapid City.
713 Keane TM, Creevey CJ, Pentony MM *et al*. (2006). Assessment of methods for amino
714 acid matrix selection and their use on empirical data shows that ad hoc
715 assumptions for choice of matrix are not justified. *BMC Evolutionary Biology*,
716 **6**, 29.
717 Koch W (1932) Über Wachstums und Altersveränderungen am Skelett des Wisents.
718 In: *Supplementbände zu den Abhandlungen der Mathematisch-*
719 *naturwissenschaftlichen Klasse*. Verlag der Bayer, München.
720 Konietzko U, Kuhl D (1998). A subtractive hybridisation method for the enrichment
721 of moderately induced sequences. *Nucleic Acids Research*, **26**, 1359-1361.
722 Krasiński M, Krasiński ZA (2013). *European Bison: The Nature Monograph* (2nd
723 ed.). Mammal Research Institute, Polish Academy of Sciences, Białowieża.
724 Kurtén, B. (1968). *Pleistocene Mammals of Europe*. London: Weidenfeld & Nicolson.
725 Lapteva EG (2009) Landscape-climatic changes on the eastern macroslope of the
726 Northern Urals over the past 50000 years. *Russian Journal of Ecology*, **40**,
727 267–273.
728 Lari M, Rizzi E, Mona S *et al*. (2011) The Complete Mitochondrial Genome of an
729 11,450-year-old Aurochsen (*Bos primigenius*) from Central Italy. *BMC*
730 *Evolutionary Biology*, **11**, 32.
731 Lazarev PA, Boeskorov GG, Tomskaya AI (1998) *Mlekopitauschie Antropogena*
732 *Yakutii* (Editor: Labutin YV). Sibirskoe otdelenie Rossiiskoi Akademii Nauk
733 ,Yakutsk.
734 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler
735 transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Liu C, Bernstein B, Schreiber S (2005) *DNA linear amplification*. Scion Publishing Ltd, Bloxham, Oxfordshire, United Kingdom.
Lorenzen ED, Nogués-Bravo D, Orlando L *et al*. (2011) Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, **479**, 359–364.
Martin, PS (1984). Prehistoric Overkill: The Global Model. In *Quaternary Extinctions: A Prehistoric Revolution* (Editors: Martin PS, Klein RG). Tucson: University of Arizona Press.
Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, pp. 10–12.
McDonald JN (1980) *North American bison. Their classification and evolution*. University of California Press, Berkeley.
Minin VN, Bloomquist EW, Suchard MA (2008) Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Molecular Biology and Evolution*, **25**, 1459–1471.
Müller UC, Pross J, Tzedakis PC *et al*. (2011) The role of climate in the spread of modern humans into Europe. *Quaternary Science Reviews*, **30**, 273–279.
Nyström, V, Dalén, L, Vartanyan, S *et al*. (2010). Temporal genetic change in the last remaining population of woolly mammoth. *Proceedings of the Royal Society B: Biological Sciences*, **277**, 2331-2337.
Orlando L, Bonjean D, Bocherens H *et al*. (2002) Ancient DNA and the Population Genetics of Cave Bears (Ursus spelaeus) Through Space and Time. *Molecular Biology and Evolution*, **19**, 1920–1933.
Pucek Z (1986) Bison bonasus (Linnaeus 1758) - Wisent. In: *Handbuch der Säugetiere Europas* (Editor: Niethammer FK). Aula Verlag, Wiesbaden.
Reimer PJ, Baillie MGL, Bard E *et al*. (2011) IntCal09 and Marine09 Radiocarbon Age Calibration Curves, 0-50,000 Years cal BP. *Radiocarbon*, **51**, 1111–1150.
Rohland N, Pollack JL, Nagel D *et al*. (2005) The Population History of Extant and Extinct Hyenas. *Molecular Biology and Evolution*, **22**, 2435–2443.
Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*, **22**, 939-946.
Ronquist F, Teslenko M, Mark P van der *et al*. (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*, **61**, 539–542.
Rule S, Brook BW, Haberle SG *et al*. (2012) The Aftermath of Megafaunal Extinction: Ecosystem Transformation in Pleistocene Australia. *Science*, **335**, 1483–1486.
Shapiro B, Drummond AJ, Rambaut A *et al*. (2004) Rise and Fall of the Beringian Steppe Bison. *Science*, **306**, 1561–1565.
Shapiro B, Ho SYW, Drummond AJ *et al*. (2011) A Bayesian Phylogenetic Method to Estimate Unknown Sequence Ages. *Molecular Biology and Evolution*, **28**, 879–887.
Sheng G-L, Soubrier J, Liu J-Y *et al*. (2014) Pleistocene Chinese cave hyenas and the recent Eurasian history of the spotted hyena, *Crocuta crocuta*. *Molecular Ecology*, **23**, 522–533.
Sher, AV (1997). An Early Quaternary bison population from Untermassfeld: *Bison menneri* sp. nov. In: Das Pleistozän von Untermassfeld bei Meiningen (Editor: Kahlke RD). Bonn, Dr Rudolf Habelt GMBH.

Sirocko F, Dietrich S, Veres D *et al*. (2013) Multi-proxy dating of Holocene maar lakes and Pleistocene dry maar sediments in the Eifel, Germany. *Quaternary Science Reviews*, **62**, 56–76.
Skinner AR, Blackwell BAB, Martin S *et al*. (2005) ESR dating at Mezmaiskaya Cave, Russia. *Applied Radiation and Isotopes*, **62**, 219–224.
Slatis HM (1960) An Analysis of Inbreeding in the European Bison. *Genetics*, **45**, 275–287.
Stewart JR, Stringer CB (2012) Human Evolution Out of Africa: The Role of Refugia and Climate Change. *Science*, **335**, 1317–1321.
Stiller M, Baryshnikov G, Bocherens H *et al*. (2010) Withering Away—25,000 Years of Genetic Decline Preceded Cave Bear Extinction. *Molecular Biology and Evolution*, **27**, 975–978.
Stuart AJ (1991) Mammalian Extinctions in the Late Pleistocene of Northern Eurasia and North America. *Biological Reviews*, **66**, 453–562.
Stuart AJ, Lister AM (2007) Patterns of Late Quaternary megafaunal extinctions in Europe and northern Asia. In: *Late Neogene and Quaternary Biodiversity and Evolution: Regional Developments and Interregional Correlations, Vol II* Courier Forschungsinstitut Senckenberg Series. Senckenbergische Naturforschende Gesellschaft, Frankfurt.
Stuart AJ, Lister AM (2012) Extinction chronology of the woolly rhinoceros Coelodonta antiquitatis in the context of late Quaternary megafaunal extinctions in northern Eurasia. *Quaternary Science Reviews*, **51**, 1–17.
van der Made, J (1999). Ungulates from Atapuerca TD6. *Journal of Human Evolution*, **37**, 389-413.
Verkaar ELC, Nijman IJ, Beeke M *et al*. (2004) Maternal and Paternal Lineages in Cross-Breeding Bovine Species. Has Wisent a Hybrid Origin? *Molecular Biology and Evolution*, **21**, 1165–1170.
Wolff EW, Chappellaz J, Blunier T *et al*. (2010) Millennial-scale variability during the last glacial: The ice core record. *Quaternary Science Reviews*, **29**, 2828–2838.
Zalkin V. (1958) The Mammals of ancient Latvia. *Moscow Society of Naturalists Bulletin*, **V. LXIII**, 5–17.

Data Accessibility:

All DNA sequences generated during the course of this study will be deposited in Genbank.

Author Contributions:

Design study: SMR JS AC, preform experiments SMR JS, analyse data: SMR JS SYWH MSYL, wrote paper: SMR JS, Edit paper: SYWH MSYL AC

Table 1. List of all samples from Urals, North Sea, Caucasus and Austria analyzed in this study.

	Sample ID	Sequence > 400bp	Species	AMS date		Calibrated dates			Origin	Field ID	Museum ID	Type
				Oxdate	Oxerr	Mean	Low	High				
Urals	A001	✓	**BisonX**	12565	55	14773	14240	15159	Rasik 1 (ZMIPAE)	ACS110	888/117	Pelvis fragment
	A003	✓	**BisonX**	12505	55	14655	14206	15072	Voronovka (ZMIPAE)	ACS121	1871/01	Humerus
	A004	✓	**BisonX**	19010	80	22698	22320	23269	Rasik 1 (ZMIPAE)	ACS88	888/1705	Metacarpal
	A005	✓	**BisonX**	15310	70	18563	18161	18764	Ladeinyi Kamen (ZMIPAE)	ACS108	929/1	Femur
	A006	✓	**BisonX**	18880	90	22536	22210	23235	Sur'ya 5 (ZMIPAE)	ACS100	994/714	Metatarsal
	A007	✓	**BisonX**	58300	2900	60660	53641	70691	Sur'ya 3 (ZMIPAE)	ACS94	884/19	Metatarsal
	A011	✓	**BisonX**	60900	**INF**	79077	50088	111451	Sur'ya 5 (ZMIPAE)	ACS99	994/715	Metatarsal
	A016	✓	**BisonX**	49600	1200	49890	47459	52529	Gofmana (ZMIPAE)	ACS104	1111/2	Humerus
	A017	✓	**BisonX**	18850	90	22495	22184	23229	Sur'ya 5 (ZMIPAE)	ACS103	994/475	Upper mandible
	A018	✓	**BisonX**	13120	60	15907	15241	16482	Sur'ya 5 (ZMIPAE)	ACS102	994/315	Radius
	BS599	✓	**BisonX**	26330	120	31003	30759	31215	Kholodnyi (ZMIPAE)		816/163	Tibia
	BS604	✓	**BisonX**	55400	1800	56123	52281	60568	Sur'ya 5 (ZMIPAE)		994/37	Astralagus
	BS606	✓	**BisonX**	25000	100	29881	29522	30221	Kholodnyi (ZMIPAE)		816/168	Bone fragment
	A002	✓	***B. priscus***	51800	1300	52145	49495	55047	Sur'ya 5 (ZMIPAE)	ACS101	994/435	Metacarpal
	A008	✓	***B. priscus***	31560	210	35960	35285	36582	Dinamitnaya (ZMIPAE)	ACS107	878/28	Metacarpal
	A013	✓	***B. priscus***	48400	900	48558	46760	50464	Rasik 1 (ZMIPAE)	ACS109	888/2271	Tibia
	A014	✓	***B. priscus***	33820	260	38673	37675	39451	Bobylek (ZMIPAE)	ACS187	528/42256	Tibia
	BS592	AY748756	***B. priscus***	42500	450	45725	44975	46500	Chernye Kosti (ZMIPAE)		887/3	Femur
	BS660	AY748766	***B. priscus***	29500	140	34195	33610	34655	Sur'ya 5 (ZMIPAE)		994/252	Metapodial
	BS674	AY748775	***B. priscus***	29060	140	33765	33220	34469	Kholodnyi (ZMIPAE)		816/166	Phalanx
	BS708	AY748793	***B. priscus***	47050	750	47158	45665	48725	Rasik 1 (ZMIPAE)		888/47	Femur
	BS713	AY748795	***B. priscus***	30970	180	35595	34957	36283	Irtysh River (ZMIPAE)		915/166	Metatarsal
	BS588	✓	***B. bonasus***	16810	65	19964	19584	20271	Sur'ya 5 (ZMIPAE)		994/716	Metapodial
	A012	✗	Contamination						Sur'ya 5 (ZMIPAE)	ACS91	994/1003	Metacarpal
	A015	✗	Contamination						Yurovsk (ZMIPAE)	ACS89	577/7	Femur
North Sea	A2791	✓	**BisonX**	53800	**INF**	74612	50092	107567	North Sea bed deposit (NSN)	JGAC09		-
	A2795	✓	**BisonX**	29010	160	33708	33142	34469	North Sea bed deposit (NSN)	JGAC13		-
	A2798	✓	**BisonX**	29230	150	33936	33365	34530	North Sea bed deposit (NSN)	JGAC16		-
	A2808	✓	**BisonX**	61500	**INF**	55979	50038	67952	North Sea bed deposit (NSN)	JGAC26		-
	A2809	✓	**BisonX**	61300	**INF**	58581	50016	75258	North Sea bed deposit (NSN)	JGAC27		-
	A2811	✓	**BisonX**	62000	**INF**	57640	50035	71248	North Sea bed deposit (NSN)	JGAC29		-
	A2792	✓	***B. priscus***	29100	150	33811	33251	34485	North Sea bed deposit (NSN)	JGAC10		-
	A2793	✓	***B. priscus***	28340	130	32629	32003	33185	North Sea bed deposit (NSN)	JGAC11		-
	A2796	✓	***B. priscus***	43850	650	47125	45672	48890	North Sea bed deposit (NSN)	JGAC14		-
	A2797	✓	*B. taurus*						North Sea bed deposit (NSN)	JGAC15		-
	A2799	✓	*B. taurus*						North Sea bed deposit (NSN)	JGAC17		-
	A2810	✓	*B. taurus*						North Sea bed deposit (NSN)	JGAC28		-
	A2800	✓	Elk						North Sea bed deposit (NSN)	JGAC18		-
	A2801	✗	Contamination						North Sea bed deposit (NSN)	JGAC19		-
	A2794	✗							North Sea bed deposit (NSN)	JGAC12		-
	A2802	✗							North Sea bed deposit (NSN)	JGAC20		-
	A2803	✗							North Sea bed deposit (NSN)	JGAC21		-
	A2804	✗							North Sea bed deposit (NSN)	JGAC22		-
	A2805	✗							North Sea bed deposit (NSN)	JGAC23		-
	A2806	✗							North Sea bed deposit (NSN)	JGAC24		-
	A2807	✗							North Sea bed deposit (NSN)	JGAC25		-
Caucasus	A4081	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N1		Long Bone
	A4082	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N2		Long Bone
	A4083	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N3		Long Bone
	A4084	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N4		Long Bone
	A4085	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N5		Long Bone
	A4087	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N7		Long Bone
	A4088	✓	**BisonX**	N/A					Mezmaiskaya, level 3	M3M N8		Long Bone
	A4089	✓	**BisonX**	59400	**INF**	96533	50080	128118	Mezmaiskaya, level 2B4	M3M N9		Long Bone
	A4091	✓	**BisonX**	59700	**INF**	77885	50037	119808	Mezmaiskaya, level 2B4	M3M N11		Long Bone
	A4092	✓	**BisonX**	56600	**INF**	54627	50020	65094	Mezmaiskaya, level 2B4	M3M N12		Long Bone
	A4094	✓	**BisonX**	56500	**INF**	53962	50010	62790	Mezmaiskaya, level 2B3	M3M N14		Long Bone
	A4104	✓	**BisonX**	12160	40	14008	13852	14163	Mezmaiskaya, level 1-2	M3M N24		Long Bone
	A4090	✓	***B. priscus***	59400	**INF**				Mezmaiskaya, level 2B4	M3M N10		Long Bone
	A4093	✓	***B. bonasus***	56300	**INF**				Mezmaiskaya, level 2B3	M3M N13		Long Bone
	A4103	✓	*B. taurus*						Mezmaiskaya, level 1-2	M3M N23		Long Bone
	A4098	✓	Brown bear						Mezmaiskaya, level 2A	M3M N18		Long Bone
	A4086	✗							Mezmaiskaya, level 3	M3M N6		Long Bone
	A4095	✗							Mezmaiskaya, level 2B2	M3M N15		Long Bone
	A4096	✗							Mezmaiskaya, level 2B2	M3M N16		Long Bone
	A4097	✗							Mezmaiskaya, level 2A	M3M N17		Long Bone
	A4099	✗							Mezmaiskaya, level 2A	M3M N19		Long Bone
	A4100	✗							Mezmaiskaya, level 2A	M3M N20		Long Bone
	A4101	✗							Mezmaiskaya, level 1C	M3M N21		Long Bone
	A4102	✗							Mezmaiskaya, level 1C	M3M N22		Long Bone
Austria	BS593	✓	***B. bonasus***	5090	60	5824	5662	5982	Steiermark (VNHM)			Femur
	BS600	✓	***B. bonasus***	3430	50	3696	3571	3833	Steiermark (VNHM)			Femur
	BS607	✓	***B. bonasus***	1370	50	1287	1179	1371	Oberosterreich (VNHM)			Femur

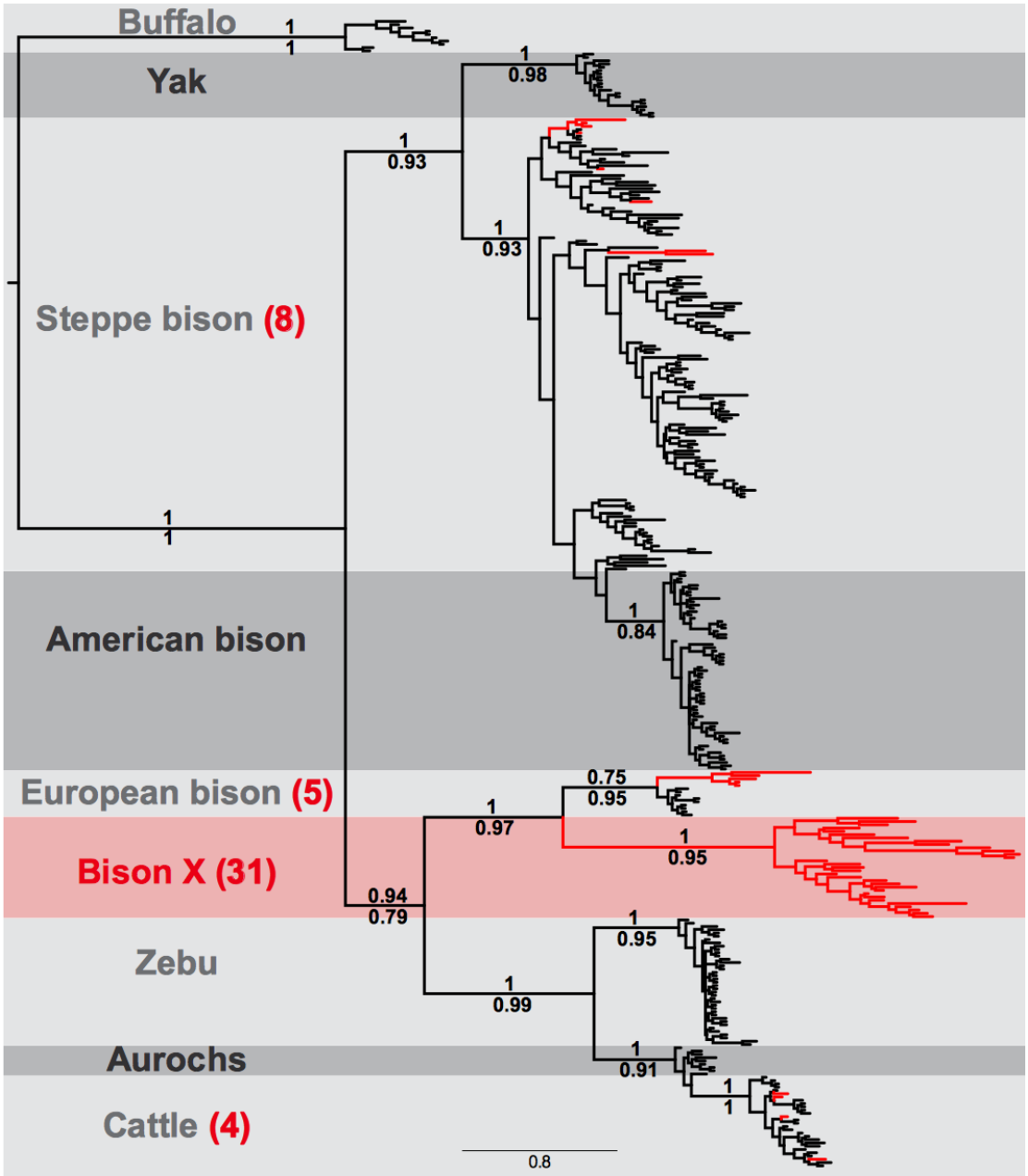


Figure 1. Phylogenetic tree of control region sequences from 350 bovid samples. The position of the 48 newly sequenced individuals are marked in red. Numbers above branches represent posterior probabilities from MrBayes, while those under branches represent aLRT support values from PhyML. Scale bar represents nucleotide substitution per site. The deep split between European bison and *Bison X* is obvious, and is of comparable scale to yak and steppe/American bison.

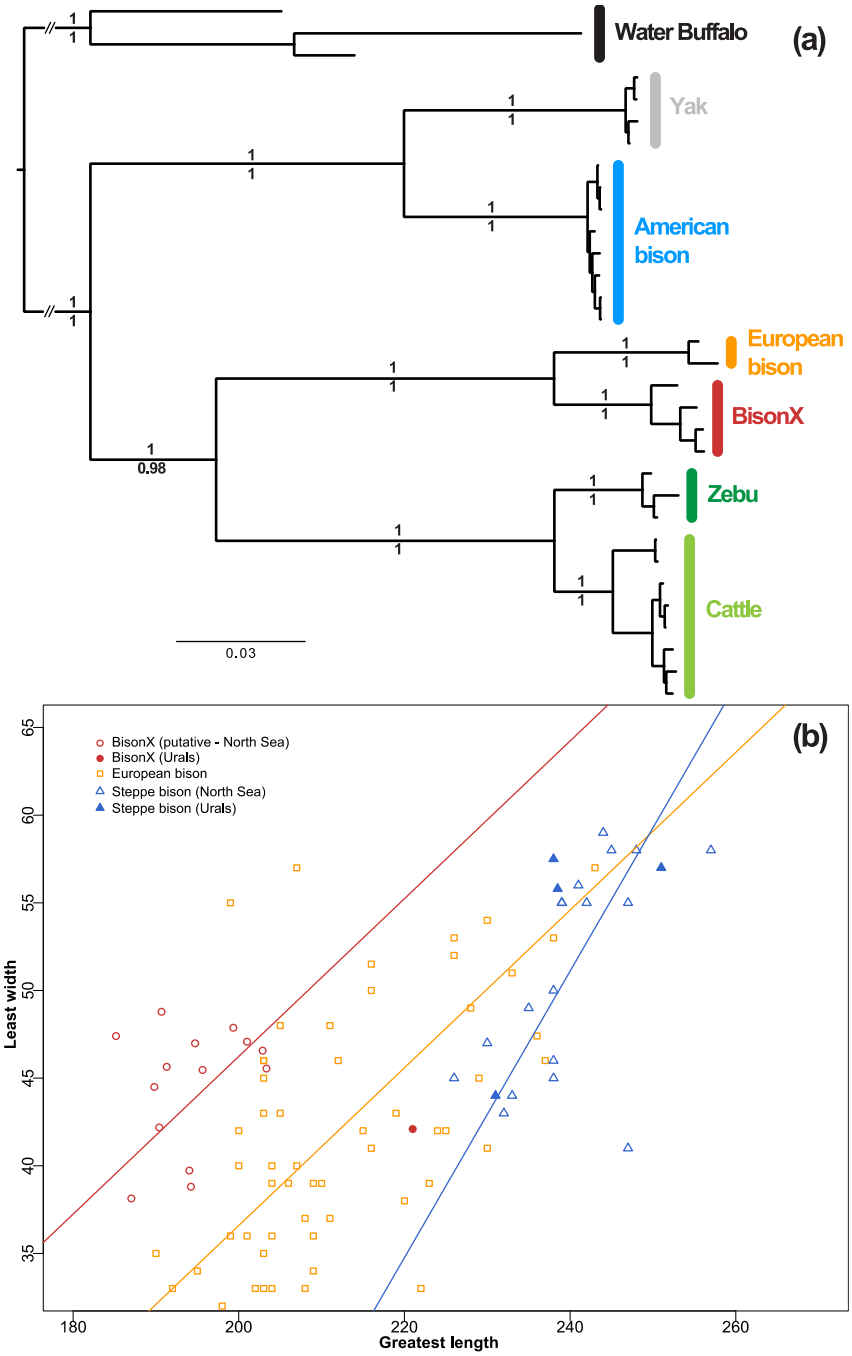
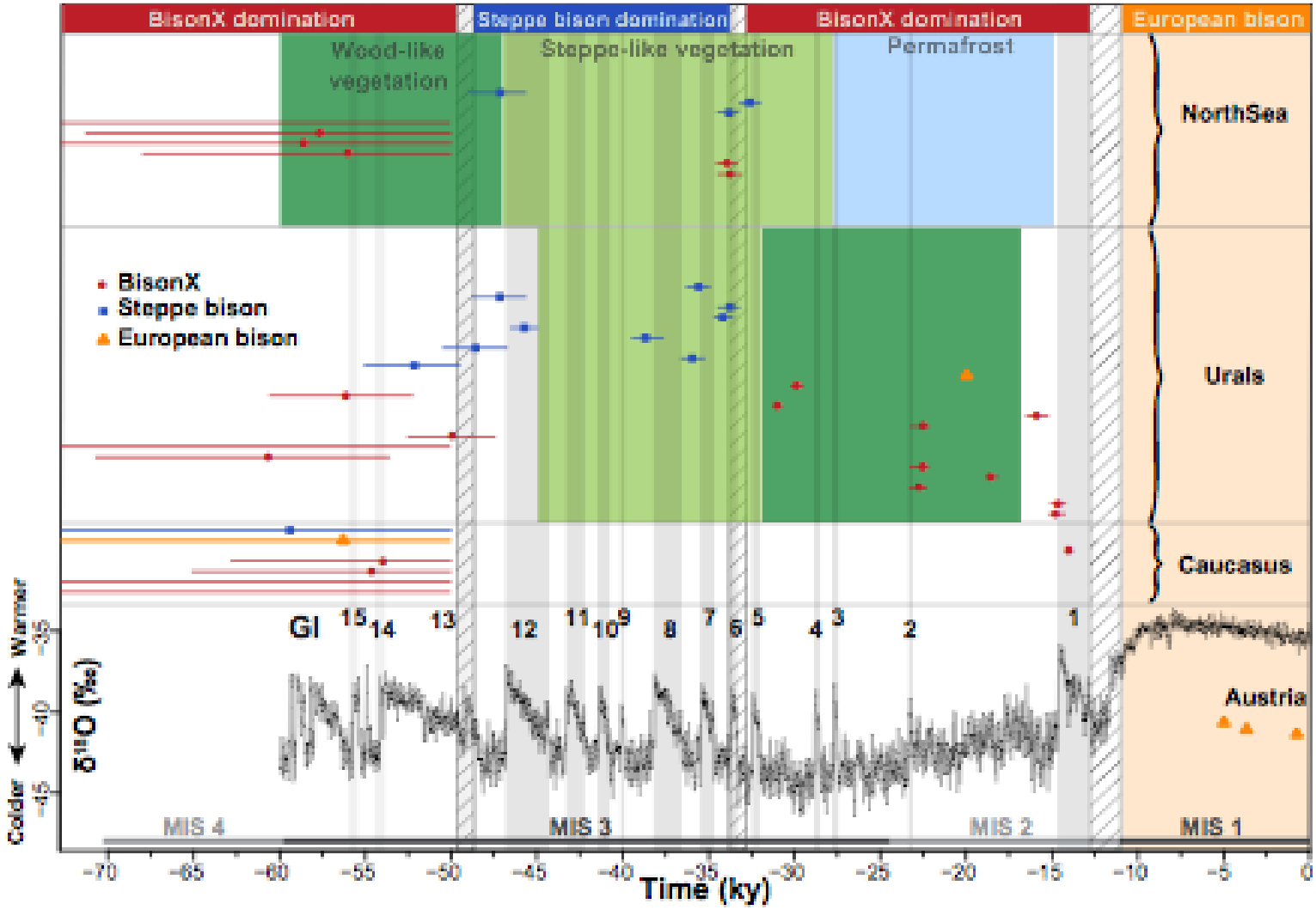


Figure 2. (a) Bovid phylogeny estimated from whole mitochondrial genome sequences. The tree generated from whole mitochondrial genomes confirms the position of *Bison X* and showing strong support for the grouping of European bison with cattle and zebu. Numbers above branches represent the posterior probabilities from MrBayes, while numbers beneath branches represent aLRT support values from PhyML. Scale bar represents lineage-wise substitutional divergence. **(b)** Allometric scaling of metacarpal measurements between three bison groups (see Table 1 for sample details). Coloured lines represent standardised major axis regressions. There is a significant difference in slope between the regression line of the steppe bison and the other two taxa (European bison and *Bison X*), as predicted by taxonomy, which groups the latter two together. However, there is no significant difference in slope between the *Bison X* and European bison regression lines (P=0.10). The intercept of the *Bison X* regression line was significantly higher than the European bison ($P=7.8x10^{-8}$) confirming the taxonomic uniqueness of the *Bison X* clade.


BisonX domination
Steppe bison domination
BisonX domination
European bison
Wood-like vegetation
Steppe-like vegetation
Permafrost
NorthSea
Urals
Caucasus
Austria
BisonX
Steppe bison
European bison
GI 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
MIS 4
MIS 3
MIS 2
MIS 1
Warmer
Colder
δ18O (‰)
−35
−40
−45
−70 −65 −60 −55 −50 −45 −40 −35 −30 −25 −20 −15 −10 −5 0
Time (ky)

851
852

853 **Figure 3.** Geographical origin and chronology of study bison samples. This illustration depicts a series of replacement patterns that correlate with climate/paleovegetation
854 events. Individual calibrated AMS dates for specimens (*Bison X*, red; steppe bison *B. priscus*, blue; European bison *B. bonasus*, orange) are shown above for the four regions
855 sampled (Urals, North Sea, Caucasus and Austria), with the NGRIP δO^{18} record below (Wolff *et al*. 2010). Greenland Interstadials (GI) are numbered in black, and Marine
856 Isotope Stages (MIS) in grey. Paleovegetation reconstructions from Eifel, Germany (Sirocko et al. 2013) and from the eastern macroslope of the Northern Urals (Lapteva
857 2009) are superimposed on the North Sea and the Urals parts of the plot respectively. Towards the end of MIS Stage 4 (60 kyr BP), *Bison X* was common across three regions
858 (Urals, North Sea, Caucasus), but was replaced by a dominance of steppe bison between GI 14 and 13 (greyed area). GI 14 was the warmest MIS 3 interstadial in the
859 paleovegetation reconstruction, followed by a succession of warming periods inducing a change into a grass steppe / meadow landscape with scattered pine and spruce which
860 continued until GI 3 (ca. 28kry BP) for Germany and GI 5 (ca. 32 kry BP) for the Urals. There are no records from the Caucasus during this period. From ca. 32 kry BP (GI
861 5) to ca. 14.5 kry BP (GI 1) *Bison X* was once again the dominant bovid observed in the Urals. These transitions closely match the change from steppe to LGM conditions,
862 which were observed as tundra / birch forest for the Urals and a polar desert in the southwestern German site slightly later. The last *Bison X* specimens (Urals: ca. 14.6 kyr
863 BP; Caucasus: ca. 14.0 C kyr BP) were dated very close to the return to grass-steppe like conditions at Eifel after the LGM (ca. 14.5 kyr BP). Although steppe bison occupied
864 this environment in MIS 3, it was not detected after this stage and indeed was in severe population decline by GI 1 (Shapiro *et al*. 2004). *B. bonasus* was observed only twice
865 prior to MIS 1, once in the MIS 4 layers at Mezmaiskaya and in MIS 2 in the Urals, suggesting these animals were not common in Europe during MIS 4-2, and only became
866 abundant during the Holocene.

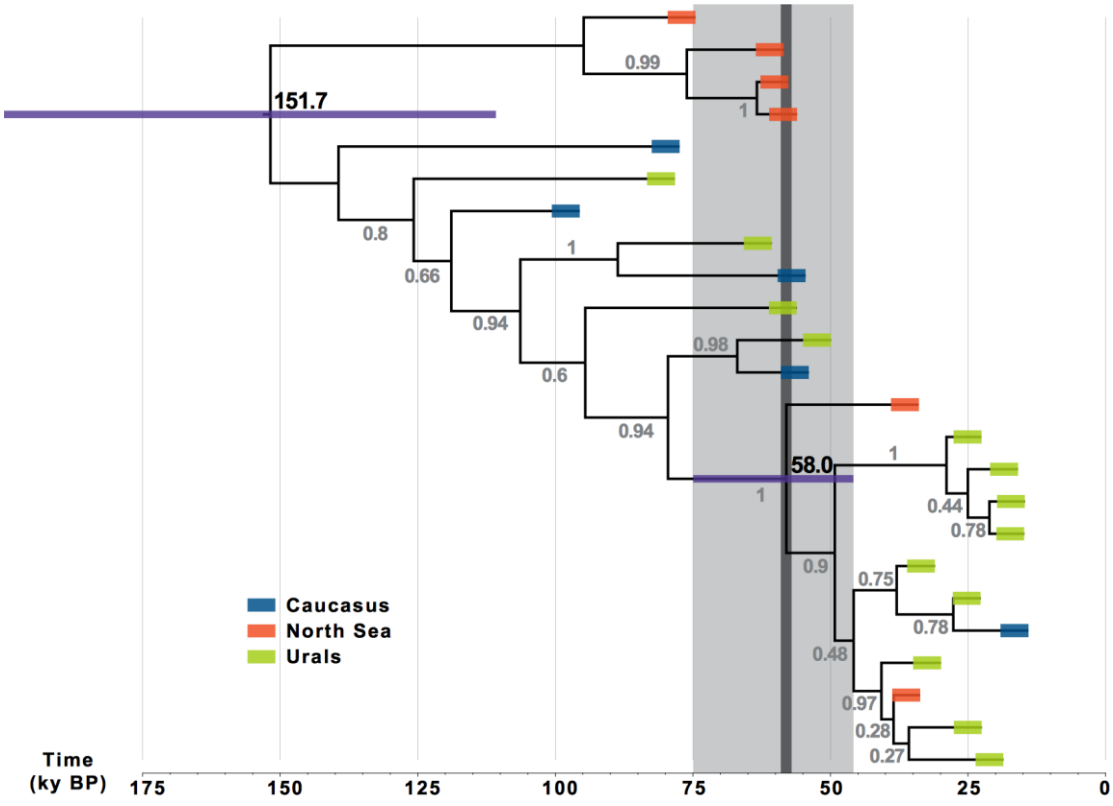


Figure 4. Maximum-clade-credibility tree of *Bison X*. Tree was estimated using Bayesian analysis and calibrated with AMS radiocarbon dates associated with the sequenced bones. Dates of samples older than 50 kyr BP are estimated in the phylogenetic reconstruction. The specimens that reappear in the colder conditions of late MIS 3/early MIS 2 (right of the grey box) form a well-nested clade within the tree, indicating they share a recent common origin. The most recent specimens within the diversity observed during MIS 4 are dated around 55-60 kyr BP, close to the MIS 3-4 boundary, and the time to the most recent common ancestor of the MIS 3 *Bison X* clade (58 kry BP). This is consistent with a scenario in which *Bison X* retreated to a refugium during the warmer phases of MIS 3, and only re-expanded with the return of cold conditions during the earliest phases of the LGM.

Supplementary Information

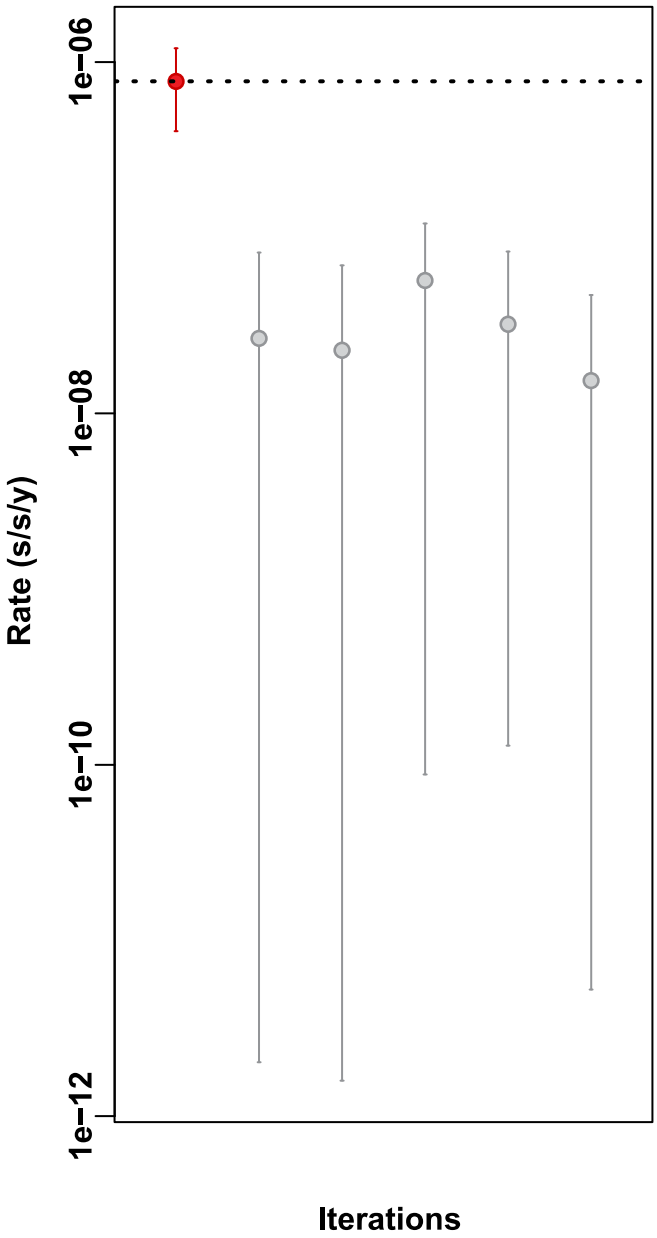


Figure S1. Date-randomization test. The red circle and dotted line represent the mean estimate of the molecular rate obtained in the phylogenetic analysis of *Bison X*, calibrated using the radiocarbon dates associated with the ancient sequences. The grey lines represent the 95% HPD intervals of rates estimated with randomized dates. None of these margins overlaps with the mean rate estimate from the original data set demonstrates that the radiocarbon dates used for this study contain sufficient temporal information for calibrating the molecular clock.

Table S1. List of published mitochondrial control region sequences used for phylogenetic analysis. The Urals steppe bison are highlighted in red. The two last numbers correspond to the radiocarbon date and its standard error, respectively.

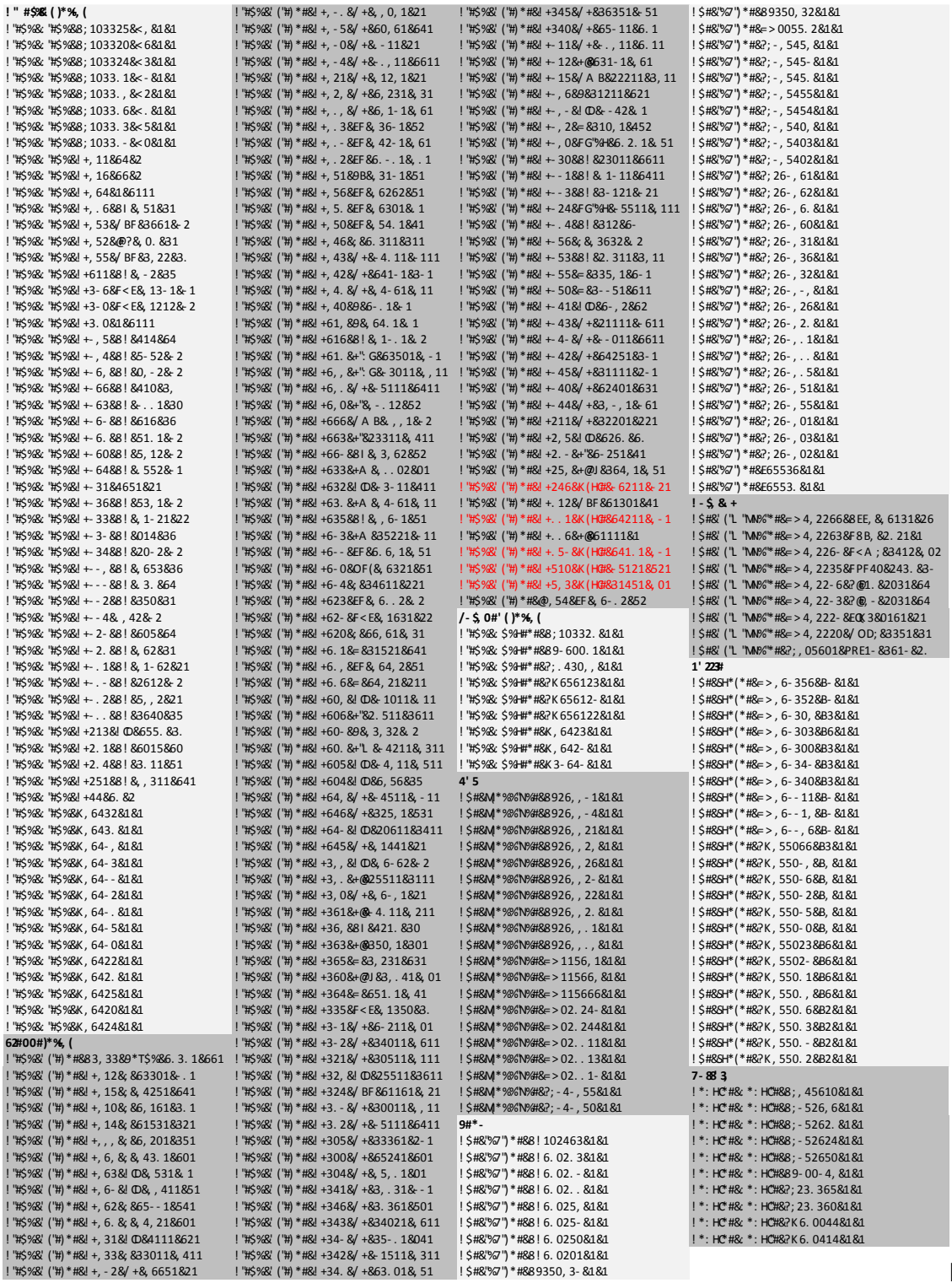


Table S2. List of published whole mitochondrial genome sequences used for phylogenetic analysis.

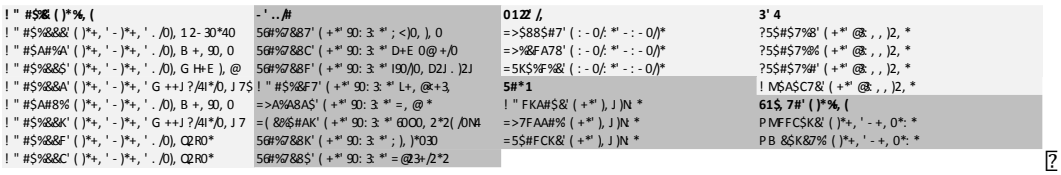


Table S3. Mitochondria control region primers

	Primer	Primer Sequence (5' --- 3')	Length[(c)]
Set_A1	BovCR-16351F	CAACCCCCAAAGCTGAAG	
	BovCR-16457R	TGGTTRGGGTACAAAGTCTGTG	~96bp
Set_B1	BovCR-16420F	CCATAAATGCAAAGAGCCTCAYCAG	
	BovCR-16642R	TGCATGGGGCATATAATTTAATGTA	~172bp
Set_A2	BovCR-16507F	AATGCATTACCCAAACRGGG	
	BovCR-16755R	ATTAAGCTCGTGATCTARTGG	~184bp
Set-B2	BovCR-16633F[(a)]	GCCCCATGCATATAAGCAAG	
	BovCR-16810R[(a)]	GCCTAGCGGGTTGCTGGTTTCACGC	~132bp
Set_A3	BovCR-16765F[(a)]	GAGCTTAAYTACCATGCCG	
	BovCR-16998R	CGAGATGTCTTATTTAAGAGGAAAGAATGG	~125bp
Set_B3	BovCR-16960F	CATCTGGTTCTTTCTTCAGGGCC	
	BovCR-80R[(a)]	CAAGCATCCCCCAAAATAAA	~110bp

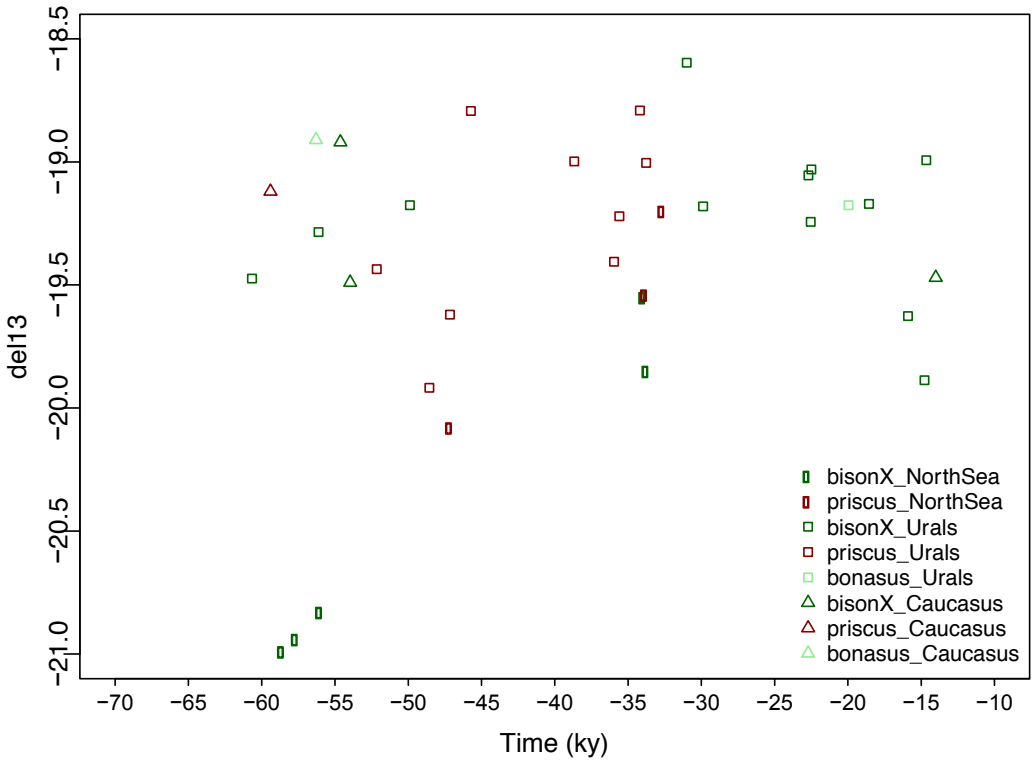
Two pairs of PCR primers derived from the mitochondrial control region and 12S rRNA were used for one-step simplex PCRs.

	Primer	Primer Sequence (5'--- 3')	Length [(c)]
Frag1	BovCR_16738MF[(b)]	***CACGACGTTGTAAAACGAC***ATYGTACATAGYA CATTATGTCAA	67bp
	BovCR_16810TR[(b)]	***TACGACTCACTATAGGGCGA***GCCTAGCGGGTT GCTGGTTTCACGC	
Frag2	Mamm_12SE	CTATAATCGATAAACCCCGATA	96bp
	Mamm_12SH	GCTACACCTTGACCTAAC	

(a): Primers (BovCR-16633F, BovCR-16810R, BovCR-16765F, BovCR-80R) were published in (Shapiro *et al*. 2004).

(b): To obtain good quality sequences for short fragment from directly sequencing, M13 (CAC GAC GTT GTA AAA CGA C) and T7 (TAC GAC TCA CTA TAG GGC GA) primer sequences were tagged at the primers BovCR_16738F and BovCR_16810R, respectively.

(c): Length of PCR amplicon is primer-excluded.



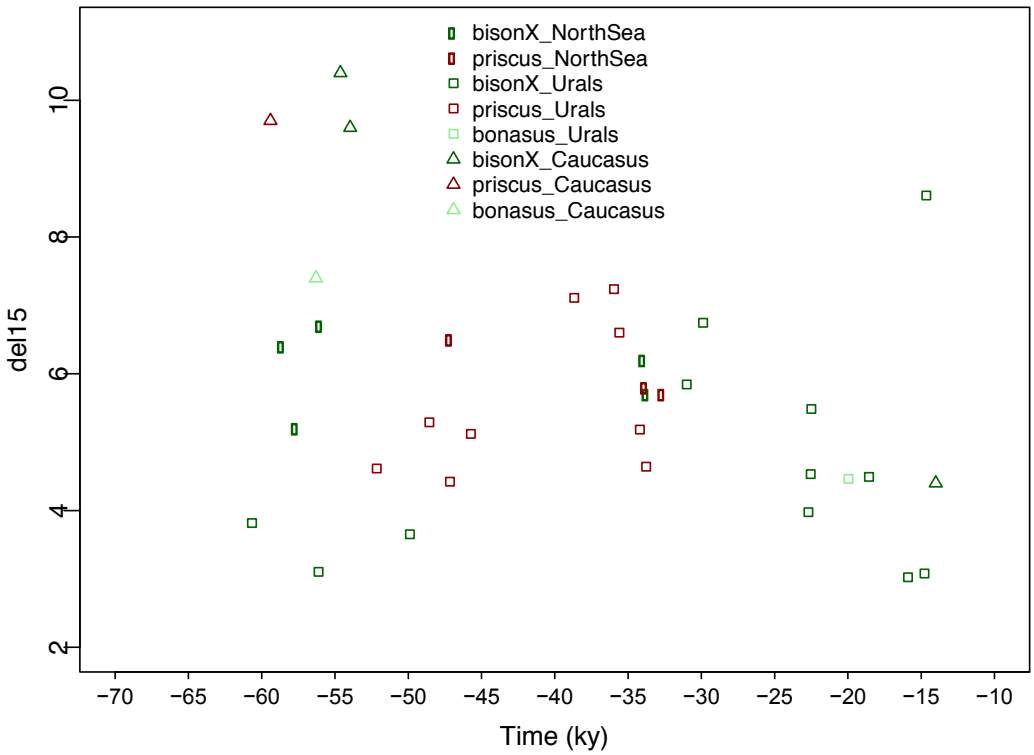


Figure S2. Comparison of Nitrogen 15 and Carbon 13 values from the surveyed samples through time.

Supplemental Methods

Whole Mitochondrial Genome Hybridization Capture

1.0 No Template Controls

All PCRs included a no template control to test for possible contamination and in all cases these negative controls produced no products.

2.0 Preparation of Truncated Adapters Working Stocks

Working stocks of the truncated adapters were prepared using oligos from Table S4 as follow (Knapp *et al*. 2012):

25 µM P5 Working Stock:
10 µL of 250 µM IS1_adapter.P5
10 µL of 250 µM IS3_adapter.P5+P7
10 µL 10× Oligo hybridization buffer
(500 mM NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA)
70 µL Molecular Biology Grade H_2O

25 µM P7 Working Stock:
10 µL of 250 µM IS2_adapter.P7
10 µL of 250 µM IS3_adapter.P5+P7
10 µL 10× Oligo hybridization buffer
70 µL Molecular Biology Grade H_2O

The adapter oligos were annealed in a thermal cycler by heating to 95°C for 10 seconds and then cooling to 14°C at -0.1°C/second. After annealing the stocks were aliquoted and stored at – 20°.

3.0 Library Construction

aDNA from four of the bison samples (ACAD sample#: A001, A004, A018, and A4089) were constructed into truncated versions of the Illumina sequencing library because the shorter adapter sequences improve hybridization capture efficiency in comparison to full length adapters (Rohland & Reich, 2012). Libraries were constructed with a protocol that was

based on previously published methods that were developed for aDNA (Briggs & Heyn, 2012; Knapp et al., 2012). The extracts were first taken through a reaction that removed uracil and polished the ends of the aDNA for ligation of adapters. Initially, these reactions contained 20 µL bison extract, 1x NEB Buffer 2 (New England Biolabs), 3U USER enzyme cocktail (New England Biolabs), 20U T4 polynucleotide kinase, 1mM ATP, 0.1 mM dNTPs, 8 µg rabbit serum albumin, and H_2O to 38.5 µL and were incubated at 37°C for 3 hours. After which, 4.5U of T4 DNA polymerase was added to each tube and the reactions were incubated for an additional 30 minutes at 25°C. Polishing reactions were purified with MinElute spin columns following the PCR cleanup provided by the manufacturer and eluting with 20 µL EB + 0.05% Tween-20. Ligation of the adapters were performed in reactions containing 20 µL polished aDNA, 1 µL 25 µM P5 short adapter working stock, 1 µL 25 µM P7 short adapter working stock, 1x T4 Ligase buffer, 5% (w/v) polyethylene glycol 4000, 6U T4 DNA ligase (Thermo Fisher) and H_2O to 40 µL and incubated at 22°C for 1 hour. Each completed ligation was MinElute purified and then pipetted in to a strand displacement reaction of 20 µL of ligated aDNA, 1x Thermopol buffer, 19.2U *Bst* DNA polymerase large fragment (New England Biolabs), 250 µM dNTPs and H_2O to 40 µL. The displacement reactions were incubated at 37°C for 10 minutes followed by heating to 80°C for 20 minutes to inactivate the *Bst*.

4.0 Whole Extract Amplification

Prior to hybridization capture, libraries were taken through two sequential low cycle PCRs. These amplifications were designed to generate the concentrated

libraries stocks needed for capture whilst minimizing the introduction of PCR artifacts (Dabney & Meyer, 2012). The first amplification was called Whole Extract Amplification 1(WEA1) and the second Whole Extract Amplification 2 (WEA2).

WEA1

Libraries were initially amplified in five PCRs, each containing 5 µL inactivated *Bst* reaction, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM each of primers IS7_short_amp.P5 and IS8_short_amp.P7 (Table S4), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Amplification was performed in a heated lid thermal cycler programed as follows 1 cycle: 98°C for 30 seconds; 14 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. After amplification, 2 µL of each PCR was gel electrophoresed and produced smears approximately 150 to 300 base pairs in length. PCRs were pooled and mixed with 1.8 volumes of Ampure XP (Beckman Coulter, Gladesville, NSW) in 1.5 ml low-bind tubes. The mixtures were allowed to stand for 5 minutes, then the tubes were placed in magnetic rack and the beads pelleted for 3 minutes. Supernates were discarded and the beads were resuspended in 800 µL 70% ethanol then pelleted as before. Supernates were again discarded and keeping the tubes in the magnetic rack, the beads were washed twice with 800 µL 70% ethanol. After the last wash was discarded, the tubes were left open in the magnetic rack to dry the beads. The dried beads were resuspended in 30 µL 10 mM Tris pH 8.0 + 0.05% Tween-20 and allowed to stand for 5 minute. Beads were then pelleted in the magnetic rack for 3 minutes and the supernate containing the amplified library transferred to a fresh low bind tube. Libraries

were visualized on an agarose gel, quantified with a NanoDrop 2000, and then stored at -20°C.

WEA2

WEA2 was performed in five PCRs that contained 3 ng purified WEA1 library, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM each of primers IS7_short_amp.P5 and IS8_short_amp.P7 (Table S4), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Amplification and processing of the library was done as per WEA1.

Supplemental References

Briggs AW, Heyn P (2012). Preparation of next-generation sequencing libraries from damaged DNA. *Methods in Molecular Biology,* **840**, 143-154.

Dabney J, Meyer M (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.

Knapp M, Stiller M, Meyer M (2012). Generating barcoded libraries for multiplex high-throughput sequencing. *Methods in Molecular Biology,* **840**, 155-170.

Rohland N, Reich D (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research,* **22,** 939-946.

Table S4. Oligonucleotides for whole mitochondrial genome hybridization capture

	5’ to 3’ Sequence
IS1_adapter.P5	A*C*A*C*TC TTTCCCTACACGACGCTCTTCCG*A*T*C*T
IS2_adapter.P7	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
IS3_adapter.P5+P7	A*G*A*T*CGGAA*G*A*G*C
Bovid_LR_Mito_F1	GGTTTGGTCCCAGCCTTCCTGT
Bovid_LR_Mito_R1(T7)	**AATTGTAATACGACTCACTATAGGG**CGTCGAGGCATGCCAGATAGTCC
Bovid_LR_Mito_F2(T7)	**AATTGTAATACGACTCACTATAGGG**TTCGACCCGGCAGGAGGAGG
Bovid_LR_Mito_R2	GGGAAGTCACGGGTGGAGGC
Bovid_LR_Mito_F3	CGCCTTCATTACCAGCATAATTCCCA
Bovid_LR_Mito_R3(T7)	**AATTGTAATACGACTCACTATAGGG**TTGAGGAGGGTGACGGGCGG
IS7_short_amp.P5	ACACTCTTTCCCTACACGAC
IS8_short_amp.P7	GTGACTGGAGTTCAGACGTGT
IS4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT
GAII_Indexing_11	CAAGCAGAAGACGGCATACGAGATcggagttGTGACTGGAGTTCAGACGTGT
GAII_Indexing_12	CAAGCAGAAGACGGCATACGAGATacttcaaGTGACTGGAGTTCAGACGTGT
GAII_Indexing_13	CAAGCAGAAGACGGCATACGAGATtgatagtGTGACTGGAGTTCAGACGTGT
GAII_Indexing_14	CAAGCAGAAGACGGCATACGAGATgatccaaGTGACTGGAGTTCAGACGTGT
T7-A18B	**GCATTAGCGGCCGCGAAATTAATACGACTCACTATAGGGAG**(A)18[B]
P5_short_RNAblock	ACACUCUUUCCCUACACGAC
P7_short_RNAblock	GUGACUGGAGUUCAGACGUGU

* = Phosphorothioate Bond
Bold = T7 RNA Polymerase Promoter
Lower Case = Index
B = C or G or T

Statement of Authorship

Title of Paper	Elucidating Bovid Evolution with Genotyping Technologies
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards
Contribution to the Paper	Helped conceive study design, performed all experiments, helped analyze data, helped write paper
Signature	Date: June 18, 2015

Name of Co-Author	Oliver Lomax-James Wooley
Contribution to the Paper	Helped conceive study design, helped analyze data, helped write paper
Signature	Date

Name of Co-Author	Julien Soubrier
Contribution to the Paper	Helped analyze data, helped write paper
Signature	Date: 22.06.15

Name of Co-Author	Michael S. Y. Lee
Contribution to the Paper	Helped analyze data, helped edit paper
Signature	Date: 22.6.15

Statement of Authorship

Title of Paper	Elucidating Bovid Evolution with Genotyping Technologies
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards		
Contribution to the Paper	Helped conceive study design, performed all experiments, helped analyze data, helped write paper		
Signature		Date	

Name of Co-Author	Oliver Lomax-James Wooley		
Contribution to the Paper	Helped conceive study design, helped analyze data, helped write paper		
Signature		Date	20/6/2015

Name of Co-Author	Julien Soubrier		
Contribution to the Paper	Helped analyze data, helped write paper		
Signature		Date	

Name of Co-Author	Michael S. Y. Lee		
Contribution to the Paper	Helped analyze data, helped edit paper		
Signature		Date	

Statement of Authorship

Title of Paper	Elucidating Bovid Evolution with Genotyping Technologies
Publication Status	○ Published, ○ Accepted for Publication, ○ Submitted for Publication, ◉ Publication style
Publication Details	Written for submission to PLOS ONE

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Stephen M. Richards		
Contribution to the Paper			
Signature		Date	

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Helped conceive study design, helped edit paper		
Signature		Date	24/06/2015

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Name of Co-Author			
Contribution to the Paper			
Signature		Date	

Elucidating Bovid Evolution with Genotyping Technologies

Stephen M. Richards[1¶*], Oliver Lomax-James Wooley[1¶], Julien Soubrier[1], Michael S. Y. Lee [2], Alan Cooper[1]

[1]Australian Centre for Ancient DNA, University of Adelaide, Adelaide, South Australia, Australia
[2] South Australian Museum, Adelaide, South Australia, Australia
*Corresponding author
E-mail: steve.richards@adelaide.edu.au
¶These authors contributed equally to this study

Abstract

Genotyping microarrays assay a large number of SNPs across the genome of an organism. Genotyping microarrays are inexpensive and are available for many domesticated animals, making the technology attractive for evolutionary studies. The Illumina BovineSNP50 BeadChip is a microarray for genotyping taurine cattle and has the potential for use in evolutionary studies of bovids. However, there are three main challenges to using the BovineSNP50 BeadChip in classical phylogenetic analyses and evolutionary studies: *1) the* BovineSNP50 *reports SNPs using a three state representation (AA, BB, and AB) system instead of identifying the nucleotides in the SNPs, 2) SNPs found on the microarray have a large ascertainment bias because these mutations only reflect the variability in taurine cattle, 3) the BovineSNP50 is not designed to assay damaged and contaminated samples such as ancient DNA*. In this study, we describe new analytical methods to address these issues: accounting for the three state SNP data and the inherent ascertainment bias in the BovineSNP50. When reanalyzing genotypes generated with modern DNA, our method produced improved phylogenies in comparison to past studies. Furthermore, we examined the reproducibility of the genotyping approach on extinct steppe bison using the BovineSNP50 and found that bison aDNA produced poor quality data and variable placement in a phylogenetic tree. We demonstrated that the BovineSNP50 can be used to accurately produce bovid phylogenies with modern DNA but further work is needed before aDNA can be reliably genotyped using this technology. In addition, we examined the possibility of genotyping ancient bison specimens using a commercially produced hybridization capture kit and high throughput sequencing. In order to maximize the variations assayed, we targeted each SNP with a single probe. Our probe design proved to be ineffective by producing insufficient read depth coverage for SNP calling in a majority of the loci targeted for enrichment.

Introduction

Genotyping Microarrays in Evolutionary Biology

Genotyping microarrays permit the interrogation of tens of thousands of single nucleotide polymorphisms (SNPs) distributed across the genome of an organism in a

single assay. A genotyping microarray is a solid support to which oligonucleotide probes are affixed. These probes bind complimentary molecules in a DNA sample and once captured, the SNPs of interest are determined with various fluorescent reporter systems [1]. Genotyping with microarrays is a quick and inexpensive method to obtain a considerable amount of genomic data from a large number of samples. As a result the technology has been adopted for use in health care [2] and in surveying known polymorphisms in domesticated plants [3] and animals [4]. Additionally, it is now recognized that genotyping microarrays have huge potential as a tool in evolutionary biology. The demographics of human populations have been inferred using genotyping microarrays [5,6] and microarrays designed for genotyping domesticated animals have been successfully used to screen organisms from related taxa including canines [7], chickens [8], and equids [9]. Illumina produces a range of genotyping microarrays known as BeadChips. Recently, the Illumina BovineSNP50 BeadChip has been applied to modern DNA and ancient DNA (aDNA) to construct a large-scale phylogeny of the genetically diverse ruminant group, which ranges from cows to giraffes and antelope [10]. While this impressive taxonomic range is encouraging and suggests that the BovineSNP50 may have considerable potential in evolutionary studies, several analytical and technical challenges remain.

Illumina BovineSNP50 BeadChip

The Illumina BovineSNP50 BeadChip was designed to assay > 54,000 SNPs that span the genome of taurine cattle (*Bos taurus*), with known variation between commercially important breeds [11]. For the BeadChip platform, SNP variation is reported using a two-state fluorescent system; a single state is used to represent two different nucleotides (e.g. A = a or t; B = c or g). This results in three possible character states for biallelic data: homozygote ‘AA’, homozygote ‘BB’, and a

combination of states for heterozygote 'AB'. Whilst, this three-state system is perfectly adequate for breed differentiation within taurine cattle, this data poses analytical challenges for classical phylogenetic approaches. As the nucleotide composition of SNPs assayed by the BovineSNP50 is largely unknown in bovids outside the taurine cattle clade, new analytical methods will need to be developed to allow the three-state data to be used with an explicit substitution model, as implemented in maximum likelihood (ML) or related Bayesian algorithms.

Ascertainment bias

Ascertainment bias is a sampling artefact that is caused by the non-random selection of polymorphisms from the population of interest [12]. In this case, the SNP selection for the BovineSNP50 was focused on variability within taurine cattle breeds. When applied at increasing evolutionary distance from this group, the fraction of SNPs interrogated by the microarray that remain variable will rapidly decrease. This reduction in observed variability is a reflection of ascertainment bias and will affect phylogenetic inferences in two ways:

- *The non-proportionality of estimated genetic distances between taxa*. The ascertainment bias may not necessarily impact the topology (branching order) of inferred phylogenetic trees, but is expected to strongly bias branch length estimates, by generating underestimates with increasing genetic distance from taurine cattle.

- *The number of polymorphic sites is lower for taxa that are evolutionary distant from taurine cattle, limiting phylogenetic signals between closely related taxa*. In the case of American bison (*Bison bison*), the reduced amount of potentially informative characters complicates population level studies and comparisons between the two possible subspecies: plains (*Bison bison bison*) and woods bison (*Bison bison athabascae*) [10].

However, in a single clade where all taxa are equidistant to taurine cattle the ascertainment bias is normalized and the relative genetic distances between taxa are

proportional to evolutionary time. For example, the Gaur (*Bos gaurus*), Banteng (*Bos javanicus*), American bison, European bison (*Bison bonasus*) and Yak (*Bos grunniens*) form a monophyletic group within Bovidae that is sister to taurine and indicine cattle (Zebu or *Bos indicus*). In that regard, SNP data from species within the group should share an equal level of ascertainment bias. Accordingly, genotyping distances can be compared between these species to generate a dated phylogeny, as the branch lengths are proportional to time.

Bison Evolution

Paleontological and morphological records have revealed a vast diversity of extinct bison species and sub-species, many of which appear to have been based on relatively insecure grounds. Paleontological records indicate the genus *Bison* originated in southern Asia roughly 2.3 million years before present (Myr BP) [13]. By the Late Pleistocene, 126-11 thousand years before present (kyr BP), separate bison species were present in the New and Old Worlds and concurrently a single steppe bison species (*Bison priscus*) inhabited the Holarctic, ranging from the United Kingdom to northern Mexico. In North America, the long-horned bison (*Bison latifrons*) became extinct during the Late Pleistocene extinction events and was succeeded by the ancient bison (*Bison antiquus*), which evolved into the extant American bison. Whilst in Europe, the Caucasus bison (*Bison bonasus caucasicus*) and the Carpathian bison (*Bison bonasus hungarorum*) have been recognised [14,15]. Currently, only two extant species of bison are recognised: the European bison and the American bison, which includes the plains and woods bison. The modern American species are thought to be direct descendants of Late Pleistocene *B. antiquus* in America while the origins of *B. bonasus* in Europe remain unclear. American and European bison are

morphologically similar [16,17] and are capable of hybridising, characteristics that have led to their current classification as closely related sister species [13,18].

Evolutionary History of American Bison

Based on the paleontological record, Eurasian bison populations are thought to have reached the American continent through Beringia between 300 and 130 kyr BP [13]. Ancient and modern mitochondrial DNA has revealed that the modern American bison originated from the physical separation of Beringian steppe bison populations by ice sheets forming over North America during the Last Glacial Maximum (25 to 13 kyr BP) [19]. Indeed, the entire mitochondrial genetic diversity observed in modern American bison originates from the population restricted to areas south of these ice barriers.

Genetic Markers

Modern populations of both American and European bison show lower than expected mitochondrial genetic diversity and this is thought to relate to pronounced population bottlenecks caused by human hunting in the recent histories of these groups. The American bison was restricted to less than 100 individuals by the late 19th century [20] and currently there are approximately 500,000 animals of this species (including both plains and woods varieties) in North America [21]. All living European bison are descended from 17 animals originating from two small late 19th century populations [22]. The population of European bison now exceeds more than 2,000 individuals, which represents the recombination of only 12 diploid sets of genes [22]. Interbreeding with modern domestic cattle since the population bottlenecks has complicated the conservation of both American and European bison.

Nuclear and mitochondrial genetic analyses have reconstructed markedly different evolutionary relationships between European and American bison [23]. Autosomal nuclear DNA studies confirm a close relationship between European and American bison [23-25], with both species forming a monophyletic clade closely related to Yak. In contrast, mitochondrial phylogenies suggest the closest relative of the American bison is the Yak, whilst the European bison falls with the taurine and indicine cattle lineage [23,26]. Two hypotheses have been proposed to explain the discrepancy between these phylogenies:

- Incomplete lineage sorting, in which two distinct mitochondrial lineages survived in the bison/yak populations until the recent species-level split with European bison [23,26].
- Genetic introgression, where ancestral European bison mated with cattle/zebu individuals, during or shortly after a population bottleneck or with a male sex-bias so that ox/zebu-like mitochondrial DNA was able to become fixed within the population [23,26].

Neither of the two hypotheses currently appears very likely and the issue remains unresolved.

Sub-species Distinction of Plains and Woods American Bison

The subspecies distinction of American plains and woods bison is also unclear. To date, the taxonomic separation has been largely based on phenotypic characteristics such as body size and coat morphology [13,27]. This division has been supported by a previous study that found plains and woods bison genotypes carried a significant F statistic, suggesting a meaningful genetic differentiation between the two groups [28]. However, the taxonomic separation of plains and woods bison is not universally accepted as mitochondrial genetic data shows no such separation [19,29,30]. This

issue has significant conservation and agricultural implications in North America, and has become an important political consideration given the economic significance of the species, and the interaction of these animals with cattle ranching operations.

Modern Genotyping Data

The modern genotyping data used in the current study came from two sources. First, a majority of the BovineSNP50 genotyping data came from the Decker *et al*. (2009) study, which included cattle, bison, and other bovids. However, after reanalysis of the Decker data, one plains and eight woods bison were found to have produced low quality data (a high fraction of the SNPs were called as missing and heterozygous) and were excluded from the current study. After this exclusion, a total of 24 plains bison and 22 woods bison were used in our analysis. The second source of data was Dr. Ottmar Distl at the University of Veterinary Medicine Hannover, who provided the BovineSNP50 genotypes of seven European bison from Wisentgehege Springe, Germany. These data are unpublished and were included in the current study because there were no European bison analysed in the Decker *et al*. (2009) study. New analytical approaches were applied to these modern BovineSNP50 genotypes to address the issues of the three character state of the data and the inherent ascertainment bias of the microarray in classical phylogenetic analyses and evolutionary studies.

Genotyping aDNA

To fully elucidate the evolutionary history of bovids, phylogenetic analysis of aDNA from fossil remains is required. Genotyping these ancient samples with the BovineSNP50 BeadChip is one possible method to generate the necessary data. In

addition to the previously stated analytical challenges for assaying bovids, there are specific technical challenges for evaluating aDNA with BeadChip platforms. The input requirement of the BovineSNP50 BeadChip is 500 ng, a mass that is impractical to obtain using aDNA without amplification. In the standard BovineSNP50 protocol, genomic DNA is amplified by multiple displacement amplification (MDA) before being fragmented into molecules with an average length of 300 base pairs for genotyping. Unfortunately, aDNA is not suitable for MDA as it is comprised of short molecules [31] and MDA only efficiently amplifies longer templates [32]. Consequently, aDNA mass must be increased through other methods such as conversion to a sequencing library and then amplification using the library adapters [33]. However, the construction and amplification of sequencing libraries [34-36] is known to introduce various biases, and care must be taken to preserve the fidelity of any sequencing library made with aDNA.

Almost all aDNA contains damaged nucleotides, which are further challenges for BovineSNP50 genotyping. After the death of an organism, DNA begins to accumulate damage through biological and chemical decay processes. Consequently, aDNA is prone to cross links, modified bases, abasic sites, and fragmentation into short molecules (generally <150 base pairs in length) [37,38]. Nucleotides damaged through decay can act as miscoding lesions that cause certain DNA polymerases to misincorporate residues during DNA synthesis. Deamination of cytosine to uracil is a type of base damage commonly found in aDNA and causes some DNA polymerases to misincorporate an adenosine in the newly synthesized DNA strand [39]. It is thought that most miscoding lesions occur at the ends of aDNA templates because these regions are single stranded and more prone to chemical decay [37,40]. An

abundance of nucleotide misincorporations in an amplified aDNA library could contribute to poor BovineSNP50 genotyping results in several ways. Incorrect bases near the ends of DNA molecules could cause the termini of the strands to anneal incorrectly to the microarray probe, which could interfere with the single base extension reaction used by the BovineSNP50 for SNP identification. Furthermore, if there is a large fraction of molecules in a sequencing library containing misincorporated nucleotides at the site of an assayed SNP the background signal produced by these incorrect residues could make correct calling of the variation difficult.

In an ancient DNA extract, endogenous molecules typically represent only a small fraction of the total present and in most cases, the majority of aDNA stems from contaminating environmental molecules usually from bacterial or fungal sources [41,42]. The low endogenous and high exogenous molecule concentrations may also complicate accurate genotyping of aDNA with the BovineSNP50.

In this study, we investigated the reproducibility of genotyping aDNA with the BovineSNP50 BeadChip. Six sequencing libraries were constructed from the same ancient steppe bison DNA extract using Phusion DNA polymerase to minimize deaminated cytosine misincorporation [43]. Phusion is an archeal DNA polymerase that does not amplify templates containing uracil efficiently, thus reducing misincorporations in the amplified library. The libraries were amplified to produce sufficient DNA to allow the samples to be loaded directly on to the BovineSNP50 BeadChip. The resulting genotype data were analysed for reproducibility of the SNP character composition calls and placement of the specimens within a bison phylogenetic tree.

Hybridization Capture

Another possible technology that can be used for genotyping aDNA is hybridization capture, which will enrich the fraction of endogenous molecules in an ancient sample. Hybridization capture uses complimentary oligonucleotide probes to immobilize target DNA, allowing unwanted molecules to be washed away. The captured molecules are then released from the probe and sequenced with high throughput sequencing (HTS). Hybridization capture can be performed with the probes attached to a solid support or in solution [44]. To perform hybridization capture on an ancient sample, the aDNA must first be converted into a sequencing library to allow amplification through the library adapters.

The majority of aDNA hybridization capture experiments are performed with in solution protocols. Commercially produced probes for in solution enrichment generally range between 60 to 120 bases in length [45] and are labelled with biotin, which is used to precipitate the captured library molecules using streptavidin-coupled magnetic beads [46]. For commercially produced enrichment kits, the term ‘probe’ refers to a population of identical oligonucleotides that are designed to anneal to a single target of interest. The oligonucleotides that comprise a probe can number in the tens of thousands of molecules depending on the system used. Consequently, ‘probes’ refers to multiple populations of identical oligonucleotides with each population complementary to a different target. In a typical hybridization capture experiment of aDNA, each locus is targeted for enrichment with multiple tiled (overlapping) probes (Figure S1A) that are usually offset between 3 to 5 bases [47,48].

In the current study, an alternative approach to probe design was taken. In order to maximize the number of variations that could be assayed with a commercially

produced hybridization capture kit, informative SNPs were targeted for enrichment with a single probe (i.e. a population of identical nucleotides, Figure S1B). SNPs found on the BovineSNP50 were each targeted with a single 121 base long probe with the SNP residing at the exact centre of the probe sequence. SNP enrichment on three steppe bison sequencing libraries was performed to determine if this single probe design was suitable for the degraded molecules found in aDNA.

Methods

Modern Genotyping Data

Three Illumina BovineSNP50 BeadChip data sub-sets were assembled from the Decker *et al*. (2009) study and the genotyping data from the seven European bison from Wisentgehege Springe:

Dataset #1- Cattle: BovineSNP50 genotypes from 371 cattle individuals, which comprised three Indian indicine cattle breeds, one African, two Asian and 41 European taurine breeds.

Dataset #2 - Bison and relatives: BovineSNP50 data from the Bos/Bison clade, which included 10 Gaur, four Banteng, two Yak, 24 plains bison, 22 woods bison, and seven European bison.

Dataset #3 – American Bison: BovineSNP50 data from 24 plains bison, 22 woods bison, and seven European bison to serve as an outgroup.

SNP Character Analysis

Following Decker *et al*. (2009), genotyping was performed using a 40,843 SNP subset of the 54,693 loci surveyed. This subset was selected for being autosomal, having a call rate of at least 80% in 36 taurine cattle breeds, and being non-monomorphic in all

breeds [10]. As an initial evaluation, two parameters of character composition of each BovineSNP50 assay were examined: the ratio of SNP homozygote characters (AA/BB), and SNP character heterozygosity (AB) in relation to the fraction of missing data.

Phylogenetic Analysis

For the BovineSNP50 data, the program Randomized Accelerated Maximum Likelihood (RaxML) v7.2.8 [49] was used to perform maximum likelihood tree searches. The three character states from the Bovine50SNP chip (AA, BB and AB) were considered as different states in an explicit analogue of the General Time Reversible (GTR) substitution model. This employed a separate substitution parameters for the three possible transformations (AA-BB; AA-AB; BB-AB). For all analyses, 20 maximum likelihood searches were conducted to find the best tree, and branch support was estimated with 500 bootstrap replicates using the rapid bootstrapping algorithm [50].

Additionally, a dated phylogenetic analysis was performed on a data set comprising 54 Yak and American bison individuals to investigate the possibility of creating a temporal timescale for bovid evolution. The program BEAST v1.6.2 [51] was used with the *BEAST option to account for both inter and intra-specific diversity. Based on the RaxML results, plains and woods bison were considered as reciprocally monophyletic populations for the analysis. A three-state general substitution model was implemented to allow for different equilibrium state frequencies and different substitution rates between the three possible substitution types. A gamma distribution with 6 rate categories accounted for rate heterogeneity among sites. The tree was

calibrated by setting the divergence between Yak and bison at around 2.5 Myr, using a lognormal distribution with a mean of 2.5 Myr and 95% of the prior probability between 2 and 3 Myr [52]. An uncorrelated lognormal relaxed-clock accounted for rate heterogeneity among lineages [53]. Considering the large size of the data set, eight Markov chain Monte Carlo (MCMC) sampling of the same analysis were run to check for convergence toward the same likelihood. Each chain was run for 100,000,000 iterations, with posterior samples drawn every 10,000 steps. The convergence of the analysis was checked in Tracer 1.5 and the initial 10% was discarded as burn-in [51].

Extraction of aDNA

Extraction of aDNA was performed in a cleanroom designated for low DNA procedures following established guidelines for aDNA [54]. The cleanroom was in a separate building from where all post-amplification steps were performed and contained filtered positive air pressure to minimize contamination from outside sources. Workspace was cleaned with bleach after each procedure and irradiated with UV light each night. During the course of the study aDNA was extracted from the following three steppe bison bones (Table 1).

A section of each bison bone was removed with a Dremel tool with a carborundum cutting disk and subsequently ground to powder with a Braun mikro-dismembrator. For each bison, aDNA was extracted from 200 mg bone powder using a previously described method [55,56] to produce 200 μL aDNA in 1x TE buffer. An extraction blank comprising 200 μL H_2O instead of bone powder was also included. No permits were required for the described study, which complied with all relevant regulations.

BovineSNP50 BeadChip Genotyping: Steppe Bison aDNA

Three aliquots of bison A3133 aDNA extract (20 μL each) and one aliquot of extraction blank (20 μL) were converted into truncated Illumina sequencing libraries (Chapter II, Supplemental Methods 3.01) and then taken through a low cycle PCR amplification (Chapter II, Supplemental Methods 4.01). Material from this first amplification was tested for bison sequences using qPCR (Chapter II, Supplemental Methods 6.0). In the qPCR assay, libraries made with aDNA were found to be positive for bison DNA while the extraction blank library was negative. DNA from this first amplification was also used in a second low cycle PCR as previously described (Chapter II, Supplemental Methods 5.01), except that for each library 30 PCRs were performed instead of five to ensure that a sufficient mass of library was produced for direct loading of the sample on to the BovineSNP50. After this second amplification, the DNA from each library was separated into two pools of 15 reactions and then purified using QIAquick spin columns with the provided PCR cleanup protocol. After purification, 500 ng of each pool was heated at 70°C until dry and then mailed to GeneSeek, Inc. (Lincoln, NE, USA) for genotyping on a BovineSNP50 BeadChip. The library pools were each loaded directly in to a chamber of a BeadChip, forgoing the standard MDA and fragmentation procedure.

Hybridization Capture of SNPs from Steppe Bison aDNA Libraries

After the poor performance of the steppe bison sequencing libraries in the BovineSNP50 procedure above, we questioned whether the library construction methodology was contributing to the low quality genotyping data produced by the aDNA (Figures 2 and 5). Consequently, several aDNA library preparation methods were tested for generating data from endogenous DNA and a protocol that used the

USER enzyme cocktail to remove uracils during library construction proved to be optimal for producing endogenous data (Chapter II) [33]. Accordingly, this USER protocol was used to prepare steppe bison libraries for these hybridization capture procedures. Truncated Illumina libraries were constructed using 20 μL extract from each of the three steppe bison (Table 1) and an extraction blank using a protocol that included USER treatment (Chapter II, Supplemental Methods 3.02) and then taken through an initial round of low cycle PCR (Chapter II, Supplemental Methods 4.01). DNA from this first round of amplification was tested for bison sequences using qPCR as above, and the libraries made with bison aDNA were found to be positive whilst the library made with extraction blank was negative. Material from this first amplification was also taken through a second round of low cycle PCR to generate a library stock with sufficient concentration for hybridization capture (Chapter II, Supplemental Methods 5.01).

A custom hybridization capture probe set was ordered from MYcroarray, (Ann Arbor, MI, USA) which targeted 39,294 of the 40,843 SNPs used in the Decker *et al*. (2009) study. The hybridization capture probes were designed using the reference sequences for the probes included in the BovineSNP50 BeadChip. The hybridization capture kit targeted fewer SNPs because the sequences surrounding certain polymorphisms were too degenerate for probe synthesis by MYcroarray. In this customized probe set, each SNP was targeted by a single 121 base pair RNA probe with the SNP residing in the exact center of the probe sequence. In contrast, the probes on the BovineSNP50 are only 60 base pairs in length. Each of the bison libraries were taken through two sequential rounds of hybridization capture following the protocol provided by MYcroarray with several minor modifications. Instead of the Block#3 provided in the kit, 1 μM of both the P5/P7 RNA short_RNAblock RNA oligonucleotides (Table S1)

were included in the reaction. These RNA oligonucleotides are complementary to the library adapters and were included in the hybridization capture reaction to prevent the reannealing of the adapters. All hybridization reactions were carried out for 48 hours and the post-hybridization washes were performed at 50°C instead of the recommended 65°C in an attempt to increase the recovery of short aDNA molecules [57]. After each hybridization capture, the libraries were amplified as previously described (Chapter II, Supplemental Methods 12.0 and 14.0). For each bison, DNA from each stage of enrichment (shotgun, first enrichment, and second enrichment) was converted into an indexed full-length Illumina library using fusion primers, in order to trace the efficiency of the enrichment steps. Each library was amplified in four PCRs containing 3 ng library DNA, 1x Phusion HF buffer, 200 µM dNTPs, 200 µM each of primers GAII_Indexing_BC*x* and IS4 (Table S1), 0.25 U Phusion Hot Start II DNA polymerase, and H_2O to 25 µL. Amplification was performed in a heated lid thermal cycler programed as follows 1 cycle: 98°C for 30 seconds; 10 cycles: 98°C for 10 seconds, 60°C for 20 seconds, 72°C for 20 seconds; and 1 cycle: 72°C for 180 seconds. After amplification, 2 µL of each PCR were gel electrophoresed and produced product smears of approximately 200 to 350 base pairs in length. PCRs were pooled and combined with 1.8 volumes of Ampure XP in a 1.5 ml low-bind tube for purification. Ampure and the pooled PCR products were mixed well and allowed to stand for 5 minutes after which the tube was placed in a magnetic rack for 3 minutes to pellet the beads. The supernate was discarded and the beads were washed three times with 800 µL 70% ethanol. After discarding the last wash, the beads were dried by leaving the uncapped tube in a magnetic rack for 10 minutes. Library was eluted by resuspending the beads in 30 µL 10 mM Tris pH 8.0 + 0.05% Tween-20 and incubating at room temperature for 5 minutes. The beads were then

pelleted by placing the tube in a magnetic rack and after 3 minutes, the supernate (containing the library) was transferred to a fresh low-bind tube. The libraries were then quantified using a NanoDrop 2000 spectrophotometer and an Agilent 2200 TapeStation running a High Sensitivity D1K ScreenTape. A fraction of each library (20 μL at 10 nM) was placed in a 96 well microtiter plate then heated at 70°C until dry and subsequently mailed to the University of Missouri for sequencing on an Illumina Hi-Seq.

Genomic Mapping of Steppe Bison SNP Enriched Libraries

Illumina HiSeq reads were filtered and trimmed using AdapterRemoval [58]. Parameters in the program were set such that mate pairs were collapsed, consecutive stretches of low quality bases with a Phred quality score below 4 were trimmed from read ends, and only sequences exceeding 25 base pairs in length were retained. Reads were then mapped to the *B. taurus* reference genome (UMD 3.1) using BWA v0.6.2, with the following parameters: "-l 1024 -n 0.01 -o 2" [59]. Amongst the successfully paired reads, duplicates were identified and removed using the programs FilterUniqueSAMCons.py (http:// bioinf.eva.mpg.de/fastqProcessing/) and picardTools markDuplicates (http://sourceforge.net/projects/picard/files/). The program mpileup from the SAMtools toolkit [60] was used to gather read depth coverage and nucleotide information at the SNP coordinates of interest.

Results/Discussion

Analysis of Modern Genotyping

In the original investigation of the modern genotyping data, Decker *et al*. (2009), maximum parsimony (MP) analysis could only reconstruct the biogeographical history of taurine cattle breeds once heterozygote SNP calls were removed. Unlike maximum likelihood or Bayesian based methods, MP does not incorporate explicit substitution models; the costs (and thus the frequencies) of all possible changes between states AA, BB, and AB are considered identical, which clearly does not reflect the evolutionary path length. The use of probabilistic methods such as ML or Bayesian approaches allow for the implementation of explicit models of evolution that account for different substitution probabilities between the three SNP genotype character states, although these analyses cannot complete phylogenetic inferences on such large data sets (≈ 50,000 characters for up to 600 taxa) in a reasonable amount of time. The RaxML program has been developed for timely phylogenetic analysis using substitution models and is built around an optimized version of the rapid hill-climbing algorithm [49].

In the present study, the program RaxML was used with a GTR-like substitution model specifically developed for multi-state data sets. This method estimates an explicit substitution model empirically from the data, allowing for unequal equilibrium character state frequencies and rates between different types of substitutions.

When the phylogeny of 47 cattle breeds was re-calculated using RaxML most of the biogeographical history described in the original analysis was recovered, with equal

or higher support for the main clades (Figure 1). However, several key differences were apparent:

1. Only three breeds appear paraphyletic (Salers, Corriente and Angus), compared to seven in the original MP tree (Salers, Corriente, Angus, Hanwoo, Texas Longhorn, Limousin, Maine-Anjou).

2. The three Italian breeds fall basal to other European cattle, rather than the New World Spanish breeds observed in the original MP tree. This arrangement is a much better match for the movement of domesticated cattle from the Fertile Crescent along the Mediterranean Coast and into south western Europe during the Neolithic, and removes a phylogeographic inconsistency identified in the original analysis.

3. British cattle breeds form a monophyletic group, with Irish representatives at the base.

These findings improve the match between phylogeny and biogeographic history compared to the Decker *et al*. (2009) analysis and confirms that heterozygote sites can contribute phylogenetic signals when an adequate substitution model is used. This represents a considerable advance, as these analytical tools can be used to reconstruct biogeographic histories for many taxa. In addition, past bison populations have undergone migration, fragmentation, and replacement events (Chapter V) [19], and consequently, reconstructing the biogeographic history of these animals will be important to understanding the evolution of the taxa.

Modern Bison Genotyping Character Composition

In our reanalysis, most modern bison produced a near identical homozygote ratio (AA/BB ≈ 0.63, Figure 2a) and a low fraction of heterozygote calls (AB ≈ 0.8% to 2.0%), potentially reflecting the recent population bottlenecks. Missing data was also minimal for the modern bison (≈ 2.0 to 4.0%, Figure 2b), with the exception of the one plains and eight woods bison that were excluded from phylogenetic analysis in

the current study, where abnormally high proportions of missing data were observed (5% to 36%). Interestingly, the elevated proportion of missing data in these excluded samples was associated with a high proportion of heterozygotes and the homozygote ratio deviated greatly from the values observed for other modern bison samples. It is not clear why these poor quality samples produced a high fraction of heterozygous SNPs. Possibly, minimal binding of target library molecules to probes may have produced a low fluorescent signal near the level of background. The resulting low fluorescence may have then been called as heterozygous by the analysis software. A detailed examination of the raw fluorescence intensities data may help elucidate this unresolved issue.

In comparison to plains bison, woods bison genotyping data exhibited a reduced level of heterozygous SNP calls (Figure 2b). Given the recent history of population bottlenecks in both groups, this suggests that woods bison suffered a more dramatic or prolonged period of restricted population size. Alternatively, some plains bison are known to have levels of introgressed cattle DNA, which would presumably also increase observed heterozygosity [61]. However, the samples analysed here are thought to contain no detectable levels of introgressed cattle DNA [10].

Modern Bison Phylogeny

The European bison was found to be closely related to American bison in the maximum likelihood phylogeny (with 100% bootstrap support, Figure 3a), as previously reported with nuclear sequences [23,24]. This is consistent with morphological and phenotypic analyses, but directly contrasts mitochondrial data [23,26]. However, the BovineSNP50 surveys the species history in more detail

because it draws information from across the entire nuclear genome rather than just the mitochondrion, which represents a single non-recombining locus. Conflicts between mitochondrial and nuclear phylogenies have been reported for other animals such as the African elephant [62], and more recently, ancient human and hominin groups [63].

Surprisingly, within the diversity of American bison, the sampled plains and woods bison animals clearly divided into two reciprocally monophyletic clades (Figure 3a). Within a phylogenetic tree comprising only bison species (Figure 3b), a deep genetic split is observed and is supported by strong bootstrap values. The reciprocal monophyly was also strongly supported when the nine low quality modern bison samples were included in the analysis (Figure S2). As noted in the Decker *et al*. (2009) analysis, when heterozygote characters were excluded reciprocal monophyly was not detected (Figure S3). However, this was because a single plains bison animal fell basal to both groups (with low support, Figure S3). This appears to suggest a relatively deep genetic separation between these two groups, and low levels of introgression at least in the sampled woods bison populations. This is perhaps surprising given the extensive co-habitation between the two groups in some national parks and suggests that a detailed assessment of conservation management aims is urgently required. The degree of genetic separation between the two groups is compelling, and reciprocal monophyly would in this case be compatible with a species-level (or at least sub-species) distinction. These results lead to the interesting question of whether the recent profound population bottlenecks in both groups may have been capable of generating the reciprocal monophyly of the SNP datasets.

Molecular Clock Analysis

To generate a timescale for bison evolution and estimate the timing of the genetic separation between American plains and woods bison, a dated phylogenetic analysis was performed using Yak, American bison, and European bison genotypes. In order to minimize ascertainment bias, dating was performed on taxa within a monophyletic clade where all taxa were equally phylogenetically distant from taurine cattle. Such an arrangement allows the branch lengths to be used to estimate differences in evolutionary time, as each is relative to the taurine cattle outgroup.

Four of eight runs converged toward the same mean likelihood, and the results were consequently combined to build the consensus tree (Figure 4). The combined statistics show convergence of the MCMC for all parameters, with effective sample size (ESS) values above 200. The other chains appeared to stall in a local maximum likelihood, and would potentially require more than 100,000,000 iterations to reach convergence. The time to the most recent common ancestor (tMRCA) of all bison (European and American) was calculated to be 594 kyr BP (279 – 984), suggesting a mid Pleistocene split between the Old and New World species, with a date of 308 kyr BP (158 – 491) for the separation of the plains and woods bison.

Previous studies of mitochondrial sequences have estimated an age for the plains bison tMRCA between 122-136 kyr BP [19,64]. It is possible that the current study produced an older age for the bison tMRCA as the paleontological calibration points are likely to produce slower estimates of evolutionary rates than the much younger carbon dated steppe bone ages used in previous studies, due to the temporal dependency of molecular rates [65,66]. As a result, the slower rates are likely to produce overestimated divergence times [67].

Genotyping Steppe Bison aDNA with the BovineSNP50 BeadChip

In the modern genotyping data, the plains and woods bison produced a reproducible AA/BB ratio of 0.63. In contrast, the ancient steppe bison replicates generated a much larger and more variable AA/BB ratio of 6.60 ± 3.3 (Figure 2a). In the steppe bison genotyping data, a large fraction of the SNPs assayed by the BovineSNP50 BeadChip did not produce data (45.3% to 56.5%, Figure 2a) and of the SNPs that did generate data a majority were called as heterozygous (54.0% to 64.0%, Figure 2b). A similar pattern in missing data and heterozygote SNP calls was seen in the low quality modern bison genotypes that were previously identified and excluded from analysis. In phylogenetic analysis, the steppe bison genotyping data replicates were located at variable positions within the topology, including being placed within *Bos javanicus* and other branches of the tree (Figure 5). This inconsistent performance contrasts with the Decker *et al*. (2009) study, which amplified and genotyped aDNA from a steppe bison in a similar manner and found the specimen to be basal to the American bison clade [10]. However, this analysis did not involve replicates of the ancient specimen, and so it is unknown if this result is reproducible.

Many factors are likely influencing the high levels of heterozygous and missing SNPs in the ancient steppe bison genotyping data. In this study, steppe bison aDNA was converted into a sequencing library and then amplified using the library adapters to produce sufficient quantities for loading on to the BovineSNP50. This may not have been an ideal method because during the hybridization incubation the adapters can re-anneal and form secondary structures that inhibit binding of the library molecules to the microarray probes [68]. If aDNA sequencing libraries are to be used for genotyping, further optimization may minimize the impact of adapter re-annealing. For example, blocking oligonucleotides could be included when the aDNA library is

hybridized to the BovineSNP50 to prevent re-annealing of adapters [46]. A modified library preparation protocol could also be used, which would enzymatically remove the adapters prior to BovineSNP50 hybridization [55]. However, the optimal solution is to use an amplification method that does not require library adapters. Converting aDNA into a circular single stranded template and then amplifying with rolling circle amplification (RCA) is one possible method that does not require adapters (Chapter III). RCA produces large concatemer products [69], which could be treated as genomic DNA and taken through the standard MDA and fragmentation protocol used for genotyping with the BovineSNP50.

The amplification method used to produce the sequencing libraries also likely contributed to the poor quality steppe bison genotyping data. The libraries were amplified with Phusion DNA polymerase as the enzyme has been reported to reduce deaminated cytosine-induced misincorporation and consequently improve library quality [43]. However, after the poor results in the current genotyping data we determined that although amplifying with Phusion does minimize misincorporations in a library, there is also a parallel loss of endogenous DNA (Chapters II and III). Consequently, the use of Phusion in the preparation of the steppe bison libraries likely contributed to the poor quality of the data for these specimens.

It is possible that non-specific binding of environmental contamination to the probes on the BovineSNP50 contributed to the poor quality of the steppe bison data, but this appears to be unlikely. The large number of SNPs called as missing in the genotyping data suggests that minimal non-specific binding occurred with the steppe bison libraries.

Hybridization Capture of SNPs

Previous studies have demonstrated that aDNA is comprised of small molecules generally < 150 base pairs in length [31]. Consequently, in an ancient extraction any locus will be represented by a multitude of aDNA molecules of varying length and sequence. Previous hybridization capture of aDNA has been successful, but the approach in these studies was to use probes that overlapped through random fragmentation of a larger template [70] or through design by tiling of the probes [47]. Contrary to these previous experiments, we targeted individual SNPs for enrichment with a single probe (i.e. population of identical oligonucleotides). Given the fragmented nature of aDNA there was uncertainty if a single probe approach would enrich the selected targets efficiently.

After hybridization capture, HTS is typically used to call any variations that may be present in the enriched DNA molecules. In HTS data, the calling of a SNP is dependent upon the read depth coverage. For diploid organisms, a SNP needs to be covered by a sufficient number of reads so that both chromosomes have been sampled. HTS produces several types of inaccuracies that can confound the accuracy of SNP calling from read depth coverage, including base calling and mapping errors [71]. SNP calling in aDNA is further complicated by contamination and damaged-induced misincorporations [41]. SNP calling can be performed with shallow 5x read depth coverage but there is a significant probability that a locus has not been sampled from both chromosomes, which may lead to inaccurate calling of heterozygous SNPs. Deep coverage at >20x will maximize the accuracy of SNP calling but sequencing at this depth is expensive. Consequently, SNP calling can be performed with medium

coverage 5-20x as this depth is considered an effective balance of study cost and data accuracy [71].

In the steppe bison shotgun data, the number of SNPs with at least 1x coverage varied greatly between the three bison specimens, which is likely a result of different levels of aDNA preservation (Table 2). Bison A3133 appears to be the best preserved of the three, producing 16,309 SNPs with at least 1x coverage in the shotgun sequencing, whilst A885 the worst preserved with only 995 SNPs. The shotgun data also produced few SNPs with at least shallow (≥ 5x) coverage with bison A3133 producing the most at 74 and bison A885 the least with none. Each of the sequential hybridization capture enrichments increased the number of SNPs with at least shallow coverage. However, even after the second enrichment these SNPs represent only a fraction of the total number of variations targeted for enrichment (A885: 1.1%, A860: 9.61%, A3133: 17.22%). Furthermore, in the second hybridization capture, the fraction of SNPs that reached the high ≥ 20x read depth coverage needed for the most accurate variant calling was even lower (A885: 0.005%, A860: 0.37%, A3133: 10.26%).

The low number of SNPs generating at least shallow coverage suggested that the single probe design of this study is not appropriate for genotyping of aDNA. The poor performance of this probe design may stem from several factors. First, the probe concentration in our study was probably insufficient for efficient hybridization capture of loci from aDNA sequencing libraries. Second, using a single probe may not be appropriate for capturing targets from aDNA sequencing libraries, which are primarily comprised of short molecules and tiled probes of different sequences may be more effective in capturing small aDNA fragments. Given the importance of accurate genotype data in phylogenetic analysis, it is recommended that such

approaches are explored in order to test the reliability of placing ancient specimens on phylogenetic tree topologies.

Conclusion

In summary, we demonstrate that with the proper analytical approach genotypes produced with the Illumina BovineSNP50 BeadChip can be used to construct accurate phylogenies of both modern taurine cattle and bison. Using this analytical method we also demonstrate that there is a significant genetic split between the American plains and woods bison, suggesting these animals may be separate species. We also found that assaying steppe bison aDNA libraries with the BovineSNP50 BeadChip produces poor quality genotypes and variable placement in a bison tree. Furthermore, we tested a single probe hybridization capture approach that prove to be ineffective for enriching informative SNPs from sequencing libraries made with steppe bison aDNA. At this time it is not clear if aDNA can be accurately genotyped with the BeadChip platform. To elucidate the evolutionary history of bovids it will be essential to generate genotype data from aDNA by some manner. Using hybridization capture to enrich for loci containing informative SNPs may prove a more viable option than microarrays. Hybridization capture is more expensive and time consuming than microarrays but less costly than shotgun sequencing.

Author Contributions

Conceived and designed experiment: SMR OLJW AC. Performed experiments: SMR. Analyzed the data: SMR OLJW JS. Write paper: SMR JS. Edited the paper: AC OLJW Provided the ancient samples: AC.

Acknowledgements

The authors would like to thank Grant Zazula and the Yukon Heritage Branch for their assistance in our fieldwork. The authors would also like to acknowledge the miners of the Yukon Territory, including the Johnson family, for their help in the collection of ancient vertebrate bones. The authors would also thank Ottmar Distl from the University of Veterinary Medicine Hannover, Germany for the genotyping data from the European bison.

References

1. Dufva M (2009) Introduction to Microarray Technology. In: Dufva M, editor. DNA Microarrays for Biomedical Research: Humana Press. pp. 1-22.
2. Henriquez-Hernandez LA, Valenciano A, Herrera-Ramos E, Lloret M, Riveros-Perez A, et al. (2013) High-throughput genotyping system as a robust and useful tool in oncology: experience from a single institution. Biologicals 41: 424-429.
3. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, et al. (2011) A Large Maize SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. PLoS ONE 6: 1-15.
4. Gheyas AA, Burt DW (2013) Microarray resources for genetic and genomic studies in chicken: A review. genesis 51: 337-356.
5. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. (2012) Ancient Admixture in Human History. Genetics 192: 1065-1093.
6. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, et al. (2014) A genetic atlas of human admixture history. Science 343: 747-751.
7. vonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, et al. (2010) Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464: 898-U109.
8. Elferink MG, Megens H-J, Vereijken A, Hu X, Crooijmans RPMA, et al. (2012) Signatures of Selection in the Genomes of Commercial and Non-Commercial Chicken Breeds. PloS One 7: e32720.
9. McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, et al. (2012) A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. PloS Genetics 8: e1002451.
10. Decker JE, Pires JC, Conant GC, Mckay SD, Heaton MP, et al. (2009) Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proceedings of the National Academy of Sciences of the United States of America 106: 18644-18649.
11. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, et al. (2009) Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4: 1-13.
12. Heslot N, Rutkoski J, Poland J, Jannink J-L, Sorrells ME (2013) Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. PLoS ONE 8: e74612.
13. McDonald JN (1981) North American Bison, Their classification and Evolution Berkeley, Los Angeles, London: University of California Press.
14. Guthrie RD (1970) Bison Evolution and Zoogeography in North America During the Pleistocene. The Quarterly Review of Biology 45: 1-15.
15. Prusak B, Grzybowski G, Zieba G (2004) Taxonomic position of *Bison bison* (Linnaeus 1758) and *Bison bonasus* (Linnaeus 1758) as determined by means of cytb gene sequence. Animal Science Papers and Reports 22: 27-35.
16. Van Zyll de Jong CG (1986) A systematic study of recent bison, with particular consideration of the wood bison (*Bison bison athabascae* Rhoads, 1898). Ottawa: National Museums of Canada : National Museum of Natural Sciences.

17. Guthrie RD (1990) Frozen Fauna of the Mammoth Steppe: The story of Blue Babe. Chicago and London: The University of Chicago Press.
18. Skinner MF, Kaisen OC (1947) The fossil Bison of Alaska and preliminary revision of the genus. Bulletin of the American Museum of Natural History 89: 123-256.
19. Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, et al. (2004) Rise and fall of the Beringian steppe bison. Science 306: 1561-1565.
20. Roe FG (1951) The North American Buffalo-a critical study of the species in its wild state. Toronto: University of Toronto Press.
21. Kohl MT, Krausman PR, Kunkel K, Williams DM (2013) Bison Versus Cattle: Are They Ecologically Synonymous? Rangeland Ecology & Management 66: 721-731.
22. Slatis HM (1960) An analysis of inbreeding in the European bison. Genetics 45: 275-287.
23. Verkaar ELC, Nijman IJ, Beeke M, Hanekamp E, Lenstra JA (2004) Maternal and paternal lineages in cross-breeding bovine species has wisent a hybrid origin? Molecular Biology and Evolution 21: 1165-1170.
24. Buntjer JB, Otsen M, Nijman IJ, Kuiper MTR, Lenstra JA (2002) Phylogeny of bovine species based on AFLP fingerprinting. Heredity 88: 46-51.
25. Vasil'ev VA, Steklenev EP, Morozova EV, Semenova SK (2002) DNA fingerprinting of individual species and intergeneric and interspecific hybrids of the genera Bos and Bison, subfamily bovinae. Genetika 38: 515-520.
26. Janecek LL, Honeycutt RL, Adkins RM, Davis SK (1996) Mitochondrial gene sequences and the molecular systematics of the Artiodactyl subfamily Bovinae. Molecular Phylogenetics and Evolution 6: 107-119.
27. Van Zyll De Jong CG, Gates C, Reynolds H, Olson W (1995) Phenotypic variation in remnant populations of North American bison. Journal of Mammalogy 76: 391-405.
28. Pertoldi C, Tokarska M, Wojcik JM, Kawalko A, Randi E, et al. (2010) Phylogenetic relationships among the European and American bison and seven cattle breeds reconstructed using the BovineSNP50 Illumina Genotyping BeadChip. Acta Theriologica 55: 97-108.
29. Douglas KC, Halbert ND, Kolenda C, Childers C, Hunter DL, et al. (2011) Complete mitochondrial DNA sequence analysis of Bison bison and bison-cattle hybrids: Function and phylogeny. Mitochondrion 11: 166-175.
30. Halbert ND (2005) The utilization of genetic markers to resolve modern management issues in historic bison populations: implications for species conservation. College Station: Texas A&M University.
31. Brotherton P, Endicott P, Sanchez JJ, Beaumont M, Barnett R, et al. (2007) Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. Nucleic Acids Research 35: 5717-5728.
32. Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, et al. (2003) Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. Genome Research 13: 294-307.
33. Briggs AW, Heyn P (2012) Preparation of next-generation sequencing libraries from damaged DNA. Methods in Molecular Biology 840: 143-154.

34. Seguin-Orlando A, Schubert M, Clary J, Stagegaard J, Alberdi MT, et al. (2013) Ligation Bias in Illumina Next-Generation DNA Libraries: Implications for Sequencing Ancient Genomes. PLoS ONE 8: 1-11.
35. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, et al. (2013) Characterizing and measuring bias in sequence data. Genome Biology 14: 1-20.
36. Dabney J, Meyer M (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.
37. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362: 709-715.
38. Dabney J, Meyer M, Pääbo S (2013) Ancient DNA Damage. Cold Spring Harbor Perspectives in Biology 5: 1-6.
39. Brotherton P, Endicott P, Beaumont M, Barnett R, Austin J, et al. (2008) Single primer extension (SPEX) amplification to accurately genotype highly damaged DNA templates. Forensic Science International: Genetics Supplement Series 1: 19-21.
40. Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, et al. (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.
41. Pääbo S (1989) Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences of the United States of America 86: 1939-1943.
42. Pääbo S (1985) Molecular cloning of Ancient Egyptian mummy DNA. Nature 314: 644-645.
43. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L (2011) mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 27: 2153-2155.
44. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, et al. (2010) Target-enrichment strategies for next-generation sequencing. Nature Methods 7: 111-118.
45. Sulonen AM, Ellonen P, Almusa H, Lepisto M, Eldfors S, et al. (2011) Comparison of solution-based exome capture methods for next generation sequencing. Genome Biology 12: R94.
46. Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, et al. (2013) Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. American Journal of Human Genetics 93: 852-864.
47. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, et al. (2011) A draft genome of *Yersinia pestis* from victims of the Black Death. Nature 478: 506-510.
48. Fu QM, Meyer M, Gao X, Stenzel U, Burbano HA, et al. (2013) DNA analysis of an early modern human from Tianyuan Cave, China. Proceedings of the National Academy of Sciences of the United States of America 110: 2223-2227.
49. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.
50. Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology 57: 758-771.

51. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7: 214.
52. Wang Z, Shen X, Liu B, Su J, Yonezawa T, et al. (2010) Phylogeographical analyses of domestic and wild yaks based on mitochondrial DNA: new data and reappraisal. Journal of Biogeography 37: 2332-2344.
53. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. Plos Biology 4: 699-710.
54. Cooper A, Poinar HN (2000) Ancient DNA: Do it right or not at ALL. Science 289: 1139-1139.
55. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
56. Rohland N, Hofreiter M (2007) Comparison and optimization of ancient DNA extraction. Biotechniques 42: 343-352.
57. Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ (2013) Capturing protein-coding genes across highly divergent species. Biotechniques 54: 321-326.
58. Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Research Notes 5: 337.
59. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25: 1754-1760.
60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
61. Hedrick PW (2009) Conservation Genetics and North American Bison (Bison bison). Journal of Heredity 100: 411-420.
62. Ishida Y, Oleksyk TK, Georgiadis NJ, David VA, Zhao K, et al. (2011) Reconciling Apparent Conflicts between Mitochondrial and Nuclear Phylogenies in African Elephants. PLoS ONE 6: e20642.
63. Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, et al. (2014) A mitochondrial genome sequence of a hominin from Sima de los Huesos. Nature 505: 403-406.
64. Ho SYW, Larson G, Edwards CJ, Heupink TH, Lakin KE, et al. (2008) Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. Biology Letters 4: 370-374.
65. Ho SYW, Phillips MJ, Cooper A, Drummond AJ (2005) Time Dependency of Molecular Rate Estimates and Systematic Overestimation of Recent Divergence Times. Molecular Biology and Evolution 22: 1561-1568.
66. Ho SYW, Lanfear R, Bromham L, Phillips MJ, Soubrier J, et al. (2011) Time-dependent rates of molecular evolution. Molecular Ecology 20: 3087-3101.
67. Sheng G-L, Soubrier J, Liu J-Y, Werdelin L, Llamas B, et al. (2014) Pleistocene Chinese cave hyenas and the recent Eurasian history of the spotted hyena, *Crocuta crocuta*. Molecular Ecology 23: 522-533.
68. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, et al. (2012) Targeted enrichment strategies for next-generation plant biology. American Journal of Botany 99: 291-311.
69. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. Genome Research 11: 1095-1099.
70. Maricic T, Whitten M, Pääbo S (2010) Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. PLoS ONE 5: e14004.

71. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12: 443-451.

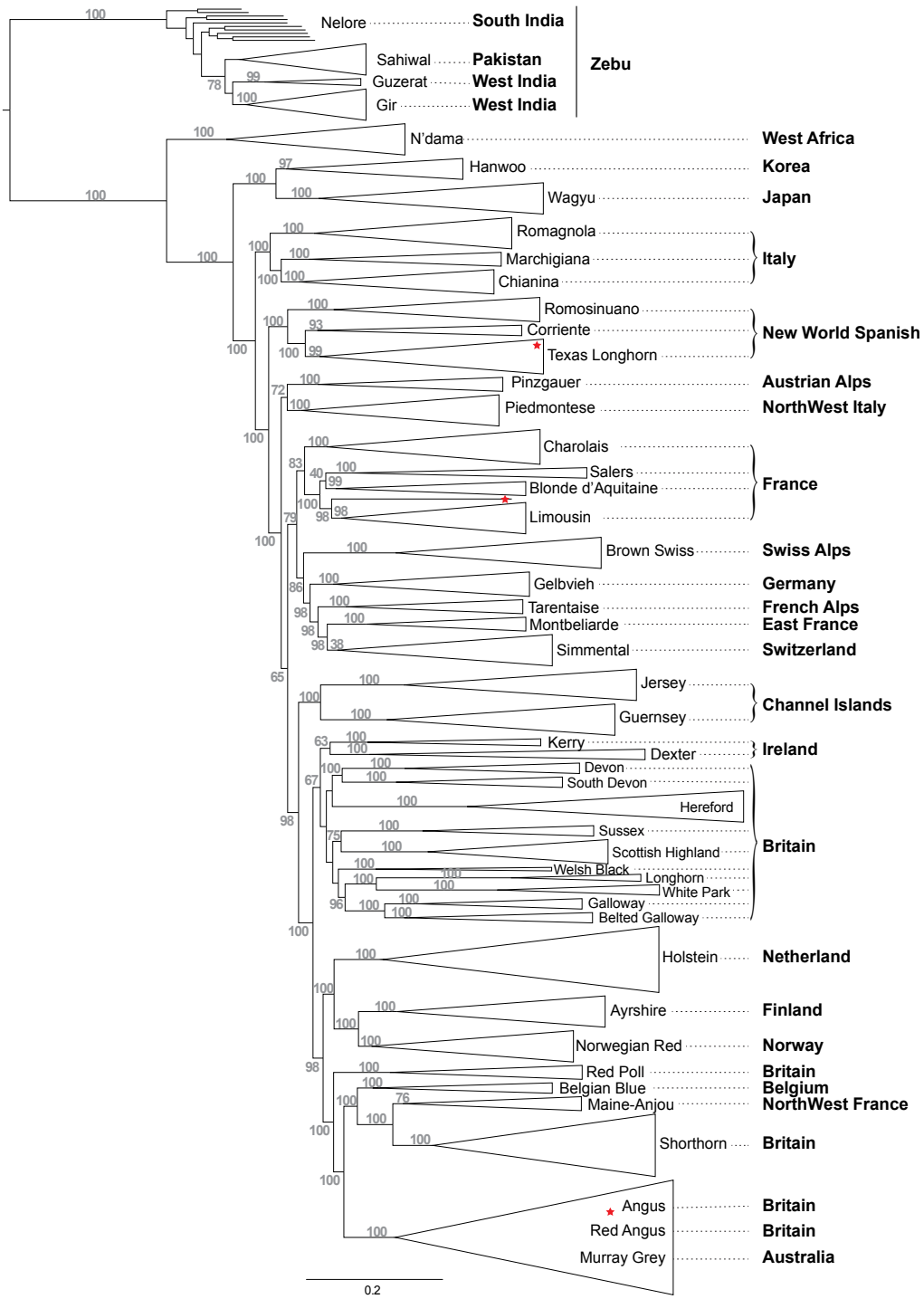


Figure 1. Reanalyzed phylogenetic tree of 47 cattle breeds
BovineSNP50 BeadChip genotyping data for 44 taurine cattle breeds and three indicine cattle breeds (outgroups) from the previous Decker *et al*. (2009) study were reanalysed using the Multi-state option of RaxML to include SNPs called as heterozygotes. Grey numbers represent the bootstrap support values, red stars the paraphyletic groups, and the bold text on the right is the geographical origin of the cattle breed. This reanalysis produced an improved match between phylogeny and biogeographic history to what was seen in the original study.

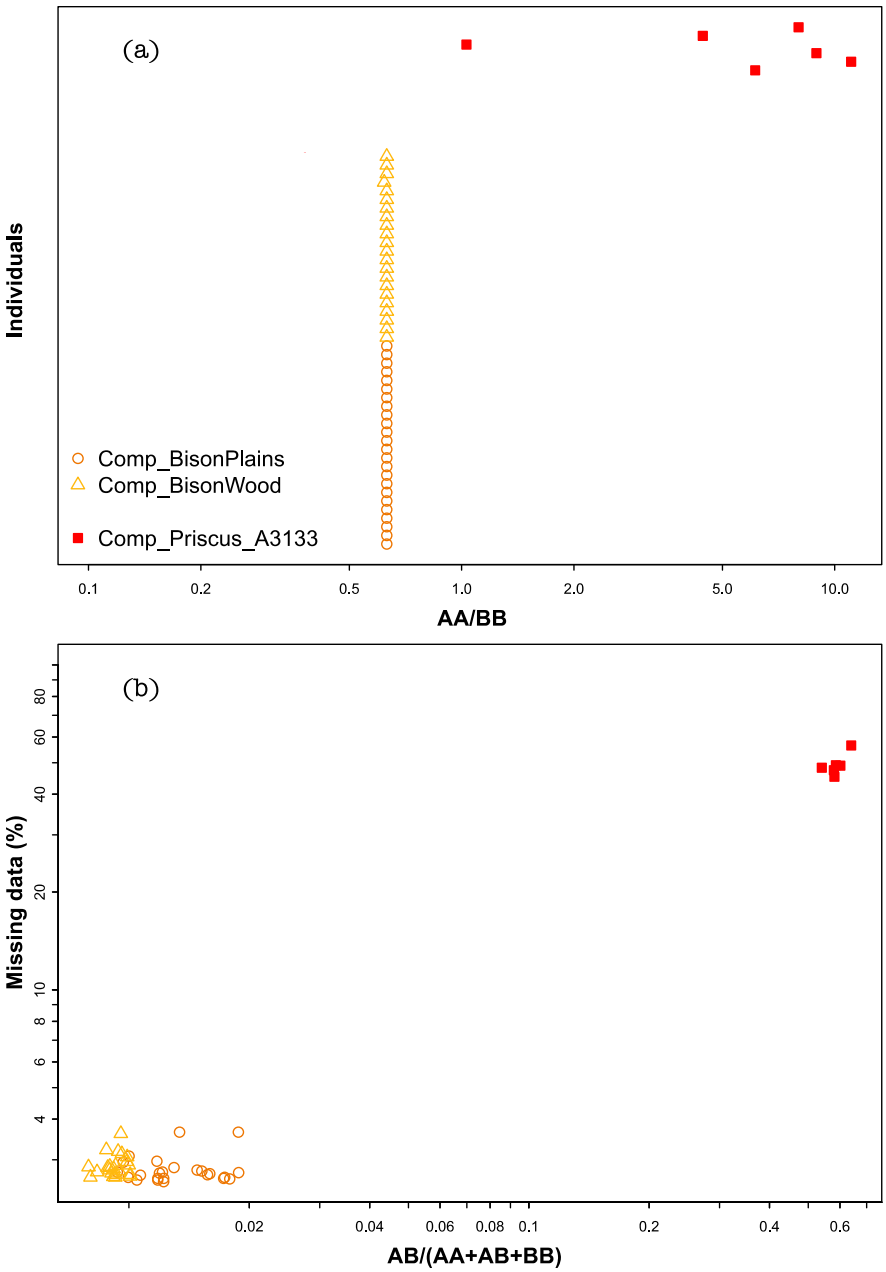


Figure 2. SNP character composition plots
Graph (a) represents the ratio of the two types of homozygote SNP calls in the modern and ancient bison BovineSNP50 BeadChip genotyping data. Graph (b) shows the relationship between the proportions of missing data and the heterozygotes SNP calls in the same bison genotyping data. The ancient bison produced a high homozygote ratio, a high proportion of heterozygotes calls, and a large fraction of missing data indicating poor quality genotyping data.

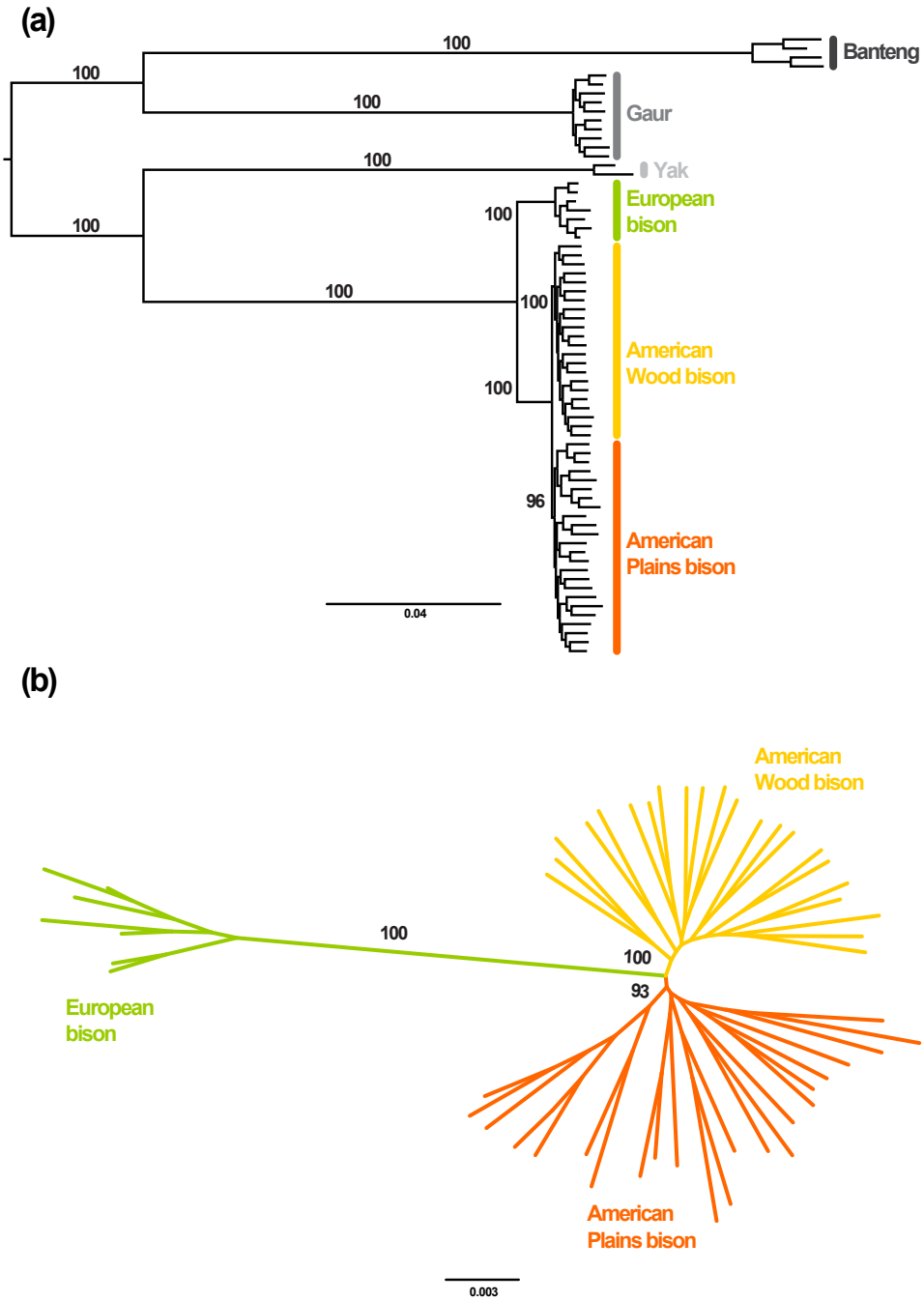


Figure 3. **Reanalysis of European and American bison genotyping data**
Bison BovineSNP50 genotyping data from Decker *et al*. (2009) and unpublished data from European bison (kindly provided by Dr. Ottmar Distl at the University of Veterinary Medicine Hannover) was reanalyzed using the multi-state option of RaxML to include SNPs called as heterozygotes. Black numbers represent branch support (based on 500 bootstrap replicates). 3a - Phylogeny including three outgroup species: the Yak (*Bos grunniens*), the Gaur (*Bos gaurus*) and the Banteng (*Bos javanicus*). European bison were retrieved as sister taxa to both American bison clades, and there is strong support for reciprocal monophyly between American plains and woods bison. 3b- Phylogenetic tree including only European and American bison, showing a clear genetic divergence between plains and woods bison clades. Scale bars indicate substitutions per site.

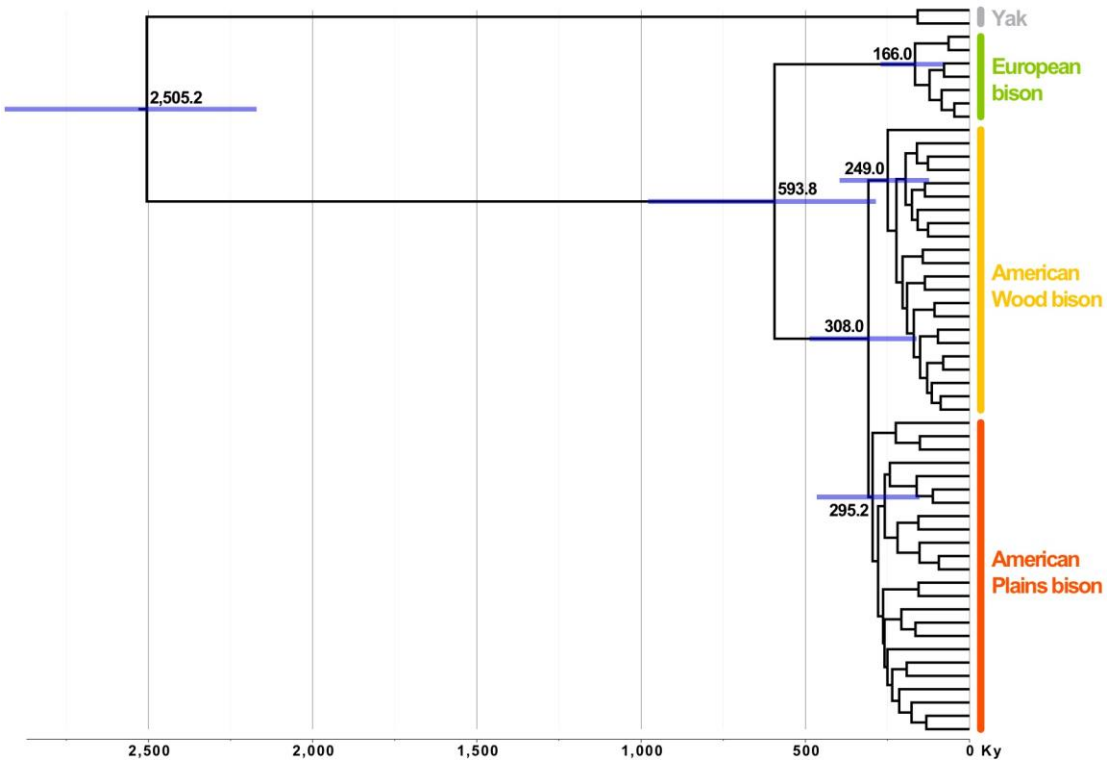


Figure 4. The maximum clade credibility tree estimated using Bayesian analysis of 55 bison and Yak BovineSNP50 genotyping data

Estimated divergence times of the main clades are noted at the node and the blue bars represent confidence intervals (95% Highest Posterior Density).

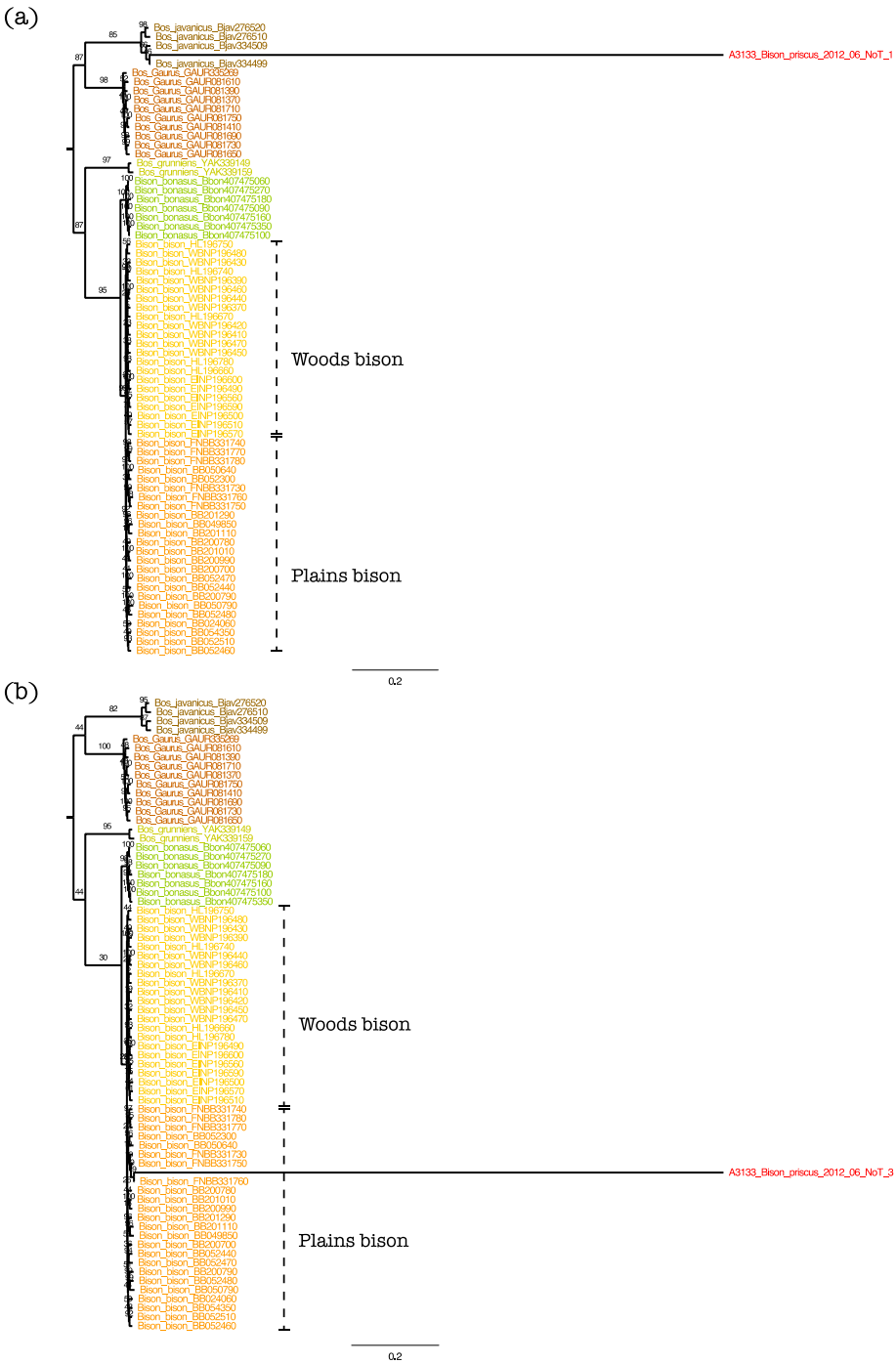


Figure 5. Random phylogenetic placement of steppe bison Bovine SNP50 replicates

To test the reproducibility of aDNA genotyping, replicate libraries made from the same steppe bison aDNA were assayed with the BovineSNP50 BeadChip. The analysis of the SNP character calls indicates that the steppe bison replicates produced poor quality data (Figure 2). The poor quality of the steppe bison data is also shown above through the random phylogenetic placement of the different replicates. Scale bars indicate substitutions per site.

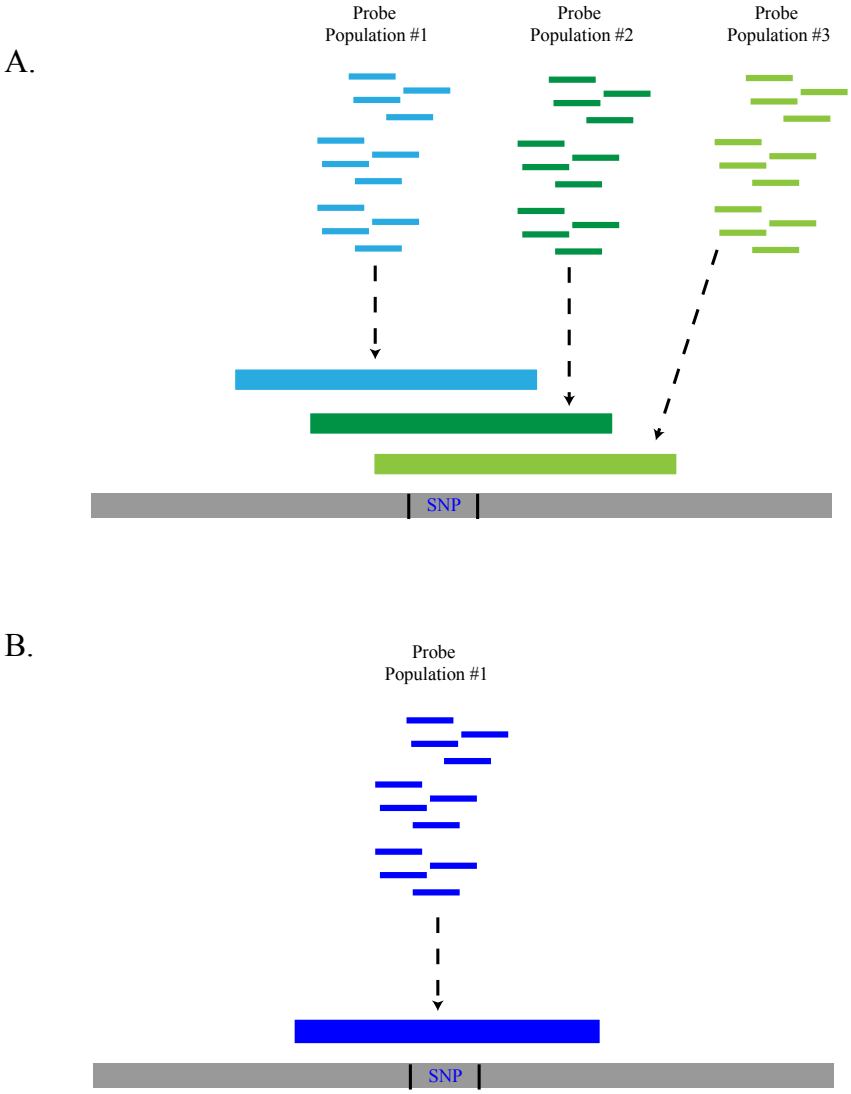


Figure S1. Schematic of probe tiling

In commercially produced hybridization capture methodologies, the term 'probe' refers to a population of identical oligonucleotides that are complimentary to a target in a sequencing library. Consequently, the term 'probes' refers to multiple populations of identical oligonucleotides and each population is complimentary to a different target in a sequencing library. SNP = single nucleotide polymorphism

A. Tiled probe design: In a typical aDNA hybridization capture experiment a locus is enriched with tiled (overlapping) probes that are offset by several nucleotides.

B. Single probe design: In the current study, informative SNPs were targeted for enrichment using a single probe (i.e. a single population of identical oligonucleotides).

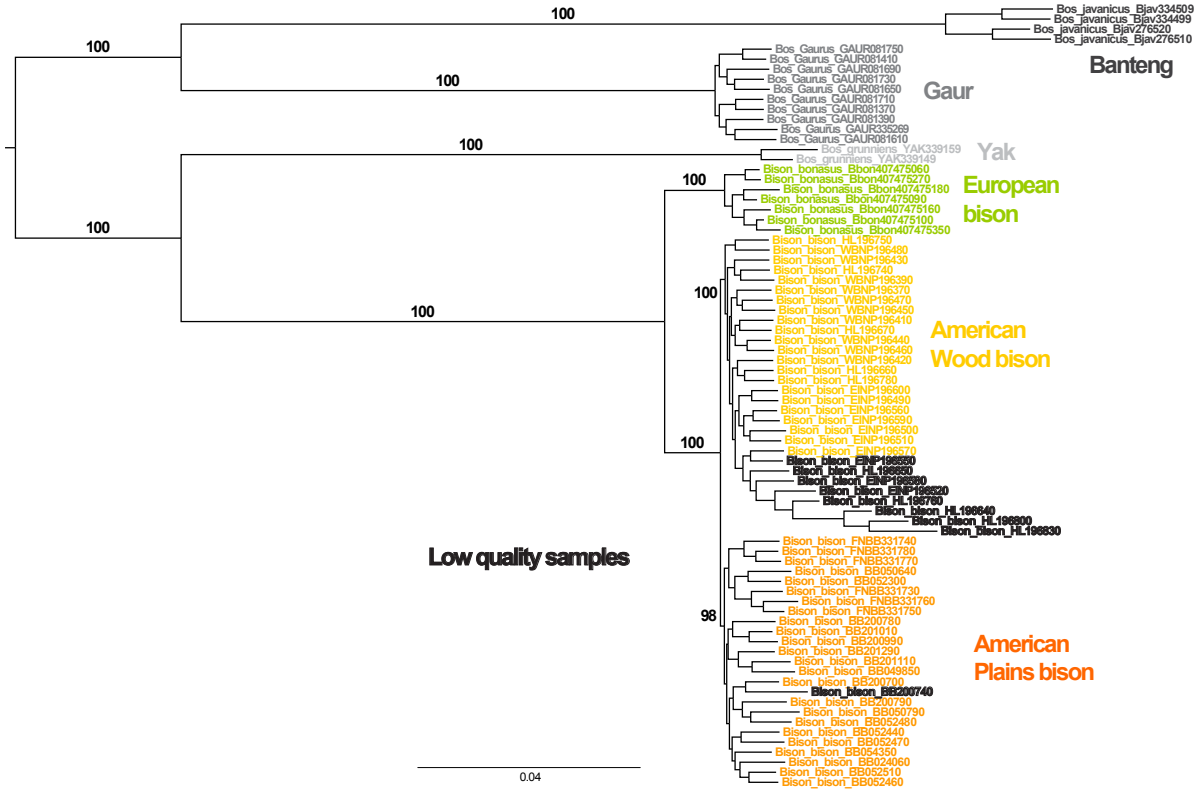


Figure S2. ML phylogenetic tree showing the position of the low quality modern bison samples

A phylogenetic tree showing inclusion of the eight poor quality genotypes does not change the reciprocal monophyly of the plains and woods bison. Scale bars indicate substitutions per site.

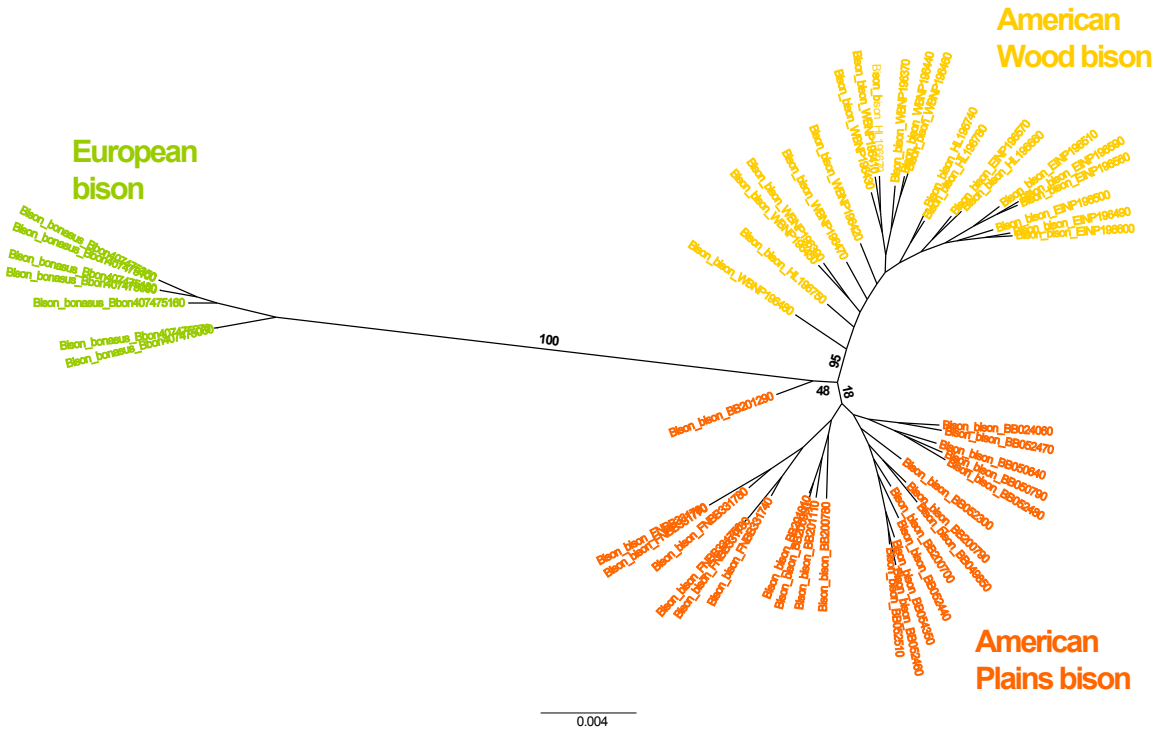


Figure S3. ML phylogenetic tree of European and American bison calculated without heterozygote characters

Although the tree is very similar to the one calculated with heterozygotes (Figure 3b), reciprocal monophyly between plains and woods American bison is contradicted by the placement of a single individual BB201290. Scale bars indicate substitutions per site.

Table 1. Steppe bison samples

ACAD Specimen #	Bone Type	*Age before present in radiocarbon years	Location Collected
A3133	astragalus	26,360 ± 220	Yukon Territory, Canada
A860	metacarpal	29,040 ± 340	Alaska, USA
A885	humerus	12,465 ± 75	Alaska, USA

*Carbon dating was performed at Oxford Radiocarbon Accelerator Unit (Oxford, United Kingdom). Ages are uncalibrated using a ^{14}C half-life of 5,568 years.

Table 2. Read depth coverage of SNPs targeted for enrichment by hybridization capture

Shotgun			
	Number of SNPs for each bison		
Read Depth Coverage	A885	A860	A3133
≥1x	995	6,414	16,309
≥ 5x	0	5	74
≥10x	0	0	1
≥ 20x	0	0	0

1st Enrichment			
	Number of SNPs for each bison		
Read Depth Coverage	A885	A860	A3133
≥1x	5,286	11,812	18,286
≥ 5x	107	2,991	5,430
≥10x	4	667	3,601
≥ 20x	0	9	2,001

2nd Enrichment			
	Number of SNPs for each bison		
Read Depth Coverage	A885	A860	A3133
≥1x	5,180	10,294	16,763
≥ 5x	429	3,775	6,767
≥10x	58	1,377	5,382
≥ 20x	2	144	4,032

Three sequencing libraries constructed from steppe bison (ACAD sample numbers: A885, A860, and A3133) were taken through two sequential round of hybridization capture for 39,294 SNPs found on the BovineSNP50 genotyping microarray. Each SNP was targeted for enrichment with a single probe and library from each enrichment stage was sequenced on an Illumina HiSeq sequencer. Sequencing data was processed and mapped to the *Bos taurus* (UMD 3.1) reference genome to determine the read depth coverage for each SNP. Despite two rounds of hybridization capture only a fraction of the SNPs targeted produced sufficient read depth coverage (≥ 5x) for SNP calling. Several factors may be contributing to the poor performance of the single probe approach used in this study. First the single probe design may not provide a sufficient probe concentration for effective enrichment of the SNP targets. Second, in comparison to overlapping probes, a single probe may be less effective for enriching the small aDNA molecules typical of an ancient extract.

Table S1. Primers and Oligonucleotides

Name	5’ to 3’
IS1_adapter.P5	A*C*A*C*TC TTTCCCTACACGACGCTCTTCCG*A*T*C*T
IS2_adapter.P7	G*T*G*A*CTGGAGTTCAGACGTGTGCTCTTCCG*A*T*C*T
IS3_adapter.P5+P7	A*G*A*T*CGGAA*G*A*G*C
IS7_short_amp.P5	ACACTCTTTCCCTACACGAC
IS8_short_amp.P7	GTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC1	CAAGCAGAAGACGGCATACGAGATcctgcgaGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC2	CAAGCAGAAGACGGCATACGAGATtgcagagGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC3	CAAGCAGAAGACGGCATACGAGATacctaggGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC4	CAAGCAGAAGACGGCATACGAGATttgatccGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC5	CAAGCAGAAGACGGCATACGAGATatcttgcGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC6	CAAGCAGAAGACGGCATACGAGATtctccatGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC7	CAAGCAGAAGACGGCATACGAGATcatcgagGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC8	CAAGCAGAAGACGGCATACGAGATttcgagcGTGACTGGAGTTCAGACGTGT
GAII_Indexing_BC9	CAAGCAGAAGACGGCATACGAGATagttggtGTGACTGGAGTTCAGACGTGT
IS4	AATGATACGGCGACCACCGAGATCT ACACTCTTTCCCTACACGACGCTCTT
B_bison_TLR8_40_F	AGCTAAGGTCAAAGGCTACAGG
B_bison_TLR8_128_R	TGACAGAAGCGTCTTTGGTG
P5_short_RNAblock	ACACUCUUUCCCUACACGAC
P7_short_RNAblock	GUGACUGGAGUUCAGACGUGU

* = Phosphorothioate bond
Lower Case Nucleotide – Barcodes

Conclusion

1.0 Overview

During the course of this thesis several technical advancements for aDNA research were described. First, in Chapters II and V novel hybridization capture protocols were described that generated probes from long-range PCR amplicons and successfully enriched loci from bison aDNA sequencing libraries. Second, again in Chapter II, in comparison to several other library preparation methods, treatment of aDNA with the enzyme cocktail USER to remove deaminated cytosines was shown to be the most effective in optimizing the production of endogenous data. Third, in Chapter III, RPA was demonstrated to be a less biased alternative to PCR in the hybridization capture of genomic targets from aDNA sequencing libraries. Lastly, Chapter VI described an analytical approach to use the three state genotyping data produced by the Illumina SNP BeadChip platform in classical phylogenetic analysis of bovids. However, other techniques investigated were found to be inconclusive. Identification of modified bases in raw aDNA with SMRT sequencing proved to be unsuccessful and genotyping of steppe bison aDNA with the Illumina BovineSNP50 BeadChip generated inconsistent results. Additionally, a hybridization capture approached that targeted informative loci with a single probe proved to be inefficient in enriching targets from aDNA sequencing libraries.

The continued advances in the capability of HTS and associated technologies will make generating large amounts of genomic data from aDNA a widespread occurrence. As these genome-scale datasets become more common, two types of investigations will become more prominent: population-level studies using genome-

wide variations and studies of molecular adaptation. Many of the methodological challenges investigated in the current thesis, such as maintaining library fidelity and identification of modified bases, will continue to be technical issues as aDNA research moves forward in these post-genomic studies.

2.0 Population-Level Studies

The large and inexpensive data production of HTS enables the study of genome-wide variation at a population-level because each site can be examined at significant coverage levels and over a larger number of individuals. Previously, population-level studies with aDNA have been conducted using mitochondrial data [1-3], however, the mitochondrial genome is a single non-recombining loci that can only reveal a fraction of the evolutionary history of a species [4]. In contrast, genome-wide data produced by high throughput technologies (HTS and genotyping microarrays) allows the comparison of thousands of loci, permitting a broader investigation of the evolutionary history of an organism. Genome-wide data permits studies that are not possible with the mitochondrial genome such as admixture analysis, which examines the interbreeding of different populations [5,6]. Depending on the study design, these genome-wide investigations can be conducted on datasets generated with several potential methodologies, three of which are detailed below.

2.1 Methodologies for Generating Genome-Wide Data

Whole genome sequencing (WGS) is one methodology that is used for generating genome-wide data from aDNA [7,8]. For WGS, a sequencing library made from aDNA is sequenced sufficiently to allow reconstruction of as much of a genome as possible. The complete reconstruction of a genome is currently unfeasible because the

small fragmented reads produced by aDNA cannot be used to define repetitive regions [9,10]. WGS of aDNA is traditionally done by shotgun sequencing, which requires a considerable mass of sequencing library and a large amount of sequencing throughput. However, other options now exist to allow hybridization capture (similar to the method described in Chapters II and V) of a whole genome prior to sequencing [11].

Whole or partial exome sequencing (ES) is another methodology to generate variation data from across eukaryotic genomes. Exons are the protein coding regions of a gene and the exome is all the exons within a genome. In humans, the exome is comprised of approximately 180,000 exons and represents 30 million base pairs or approximately 1% of the genome [12]. For ES, loci of interest are first enriched using hybridization capture and then sequenced with HTS [13,14]. ES does not truly generate genome-wide data because there are sections of some chromosomes that are devoid of genes [15]. However, exons are distributed across the majority of most chromosomes and for this discussion ES data will be considered to be genome-wide.

The last methodology that is likely to be used with aDNA to generate genome-wide data is high-density genotyping (HDG). At this time, it is not clear how HDG will be performed with aDNA. Hybridization capture may be used to enrich for molecules containing the SNPs of interest. If the technology can be adapted to produce accurate data, HDG of aDNA may also be performed with SNP microarrays (Chapter VI).

2.2 Advantages and Disadvantages of Genome-Wide Methodologies

Each of the methodologies for interrogating genome-wide variation has advantages and disadvantages. WGS will generate the largest amount of genomic information but the quantity of sequencing required to produce adequate sequence coverage is both costly and time consuming. ES in comparison, generates a much more manageable dataset and will require fewer resources to produce and analyze than WGS. The reduced cost of ES may allow for more samples to be assayed, increasing the power in a study to identify significant SNPs and/or genes. Exomes are among the most understood regions of the genome [13], but in comparison to non-coding regions, these protein-coding sequences have a reduced SNP density [16]. Of all the methods for generating genome-wide data, HDG is the most cost effective as only the SNP or the sequence in the immediate area of the variation are assayed. If current SNP microarray technologies can be adapted to aDNA then the cost benefits could be substantial, while also allowing the analysis of a larger sample size. HDG has disadvantages including the SNPs for genotyping must be known *a priori,* which is unlikely because recent widespread bottlenecks have caused a considerable loss of diversity in many populations [17,18]. Study design for HDG must also account for any ascertainment bias that might be produced by using SNPs found in particular populations (Chapter VI).

2.3 Examples of Population-Level Studies

The motivation for genome-wide population studies is to use large amounts of information to generate a more detailed analysis of evolution than what is possible with the mitochondrial genome or the limited number of variations assayed with PCR based technologies [5]. Genome-wide data permits the use of tests such as cross

population extended haplotype homozygosity (XP-EHH), which can detect selective sweeps where the selected allele is near or has achieved fixation in a subpopulation whilst remaining polymorphic in the overall population [19]. Applying test like XP-EHH to genome-wide data can help elucidate the evolutionary forces causing selective sweeps. For example, XP-EHH has been used to investigate systemic lupus erythematosus (SLE), which is an autoimmune disease with reported increased prevalence among certain populations [20]. SLE is a disorder in which the immune system produces antibodies against self-proteins, causing damage to a number of tissues including cardiac, neural, and joint. SLE is nine times more prevalent in females and primarily occurs in women during childbearing ages. SLE can be a debilitating illness that can lead to death, so there should be considerable selective pressure against any allele associated with the disease [20]. Yet, contrary to this point, SLE has increased prevalence in African-American, Asian, Hispanic, and Native American populations. It has been hypothesized that other selective pressures may have recently acted on alleles that promote SLE. In genotyping of SLE patients, alleles associated with the autoimmune disease produced XP-EHH scores indicating these loci have undergone recent positive selection and might provide some adaptive benefit [21]. Some of the alleles that produced XP-EHH scores that were indicative of selection are known to be involved with resistance to pathogens such as the *Plasmodium* species, which are the causative organisms of malaria.

The high incidence of SLE in populations that originated from regions where malaria is endemic may stem from the protective effect of these loci to *Plasmodium* infection [21]. The use of XP-EHH helped describe an evolutionary scenario that would explain the increased prevalence of alleles associated with SLE in certain populations. XP-

EHH can be applied in a similar manner to aDNA to identify alleles that have undergone recent selection and help develop evolutionary theories to explain the increased incidence of these loci.

Other statistical methodologies, such as principle component analysis (PCA), can also be applied to aDNA genome-wide studies to identify and characterize population structure [22]. PCA reduces the dimensionality of the data whilst maintaining the majority of the variability present in a dataset. PCA allows samples to be plotted, permitting similarities and differences between samples to be visualized. PCA plotting also permits a determination of whether samples exhibit population structure [23], which is the presence of different allele frequencies in the groups that comprise a population. Population structure is caused by a number of factors such as physical isolation, unequal distribution of resources, and migration [24]. PCA of genotype data can help elucidate the factors that shape population structure. For example, in a human genome-wide genotyping study of Indian populations, PCA grouped the Siddi people with African populations. This observation is in accordance with the fact that Siddi are descendant from Bantu peoples brought from Southeast Africa to the Indian subcontinent as slaves by Arab merchants [25].

These are just two of the many analysis that can be applied to genome-wide data. As the use of HTS and genotyping microarrays continues to expand in both the ancient and modern research communities the number of statistical tests that can be applied to genome-wide data will grow even larger. These analyses will be relevant to evolutionary studies of plants and animals such as those performed in this thesis. For

example, PCA analysis of genomic data could help identify the *Bison X* population that may have been isolated in a refugium during MIS 3 (Chapter V) [26].

2.4 Sample Size in Modern DNA and aDNA Population-Level Studies

Currently for modern DNA, genome-wide genotyping studies are being conducted with large sample sizes, especially for humans. A human admixture study has been conducted on 934 individuals using a HDG microarray that interrogates > 629,000 SNPs [27]. In a study of autism spectrum disorders, one research consortium has committed to whole ES of individuals from 2,000 families with a history of the condition [28]. The 1000 Genomes Project and the 1000 Bull Genomes Project aim to sequence at least a thousand genomes from humans and domesticated cattle [29,30]. The vast amount of data produced in modern DNA studies will be a vital reference for aDNA research. In comparison to modern DNA, sample size in aDNA studies will be small with recent investigations sequencing the genomes from 9 European individual ≥ 7,500 years old [31] and the exomes of three Neanderthals [14].

The use of such small aDNA sample sizes will be problematic as informative alleles in populations are those variations that are common enough to be shared by some individuals whilst remaining absent in many other individuals [32]. In some aDNA studies, informative alleles may be lost from analysis because these variations are absent in the few samples examined.

3.0 aDNA Molecular Adaptation Studies

Previously, the majority of aDNA research was concerned with the demographics and evolutionary history of past organisms. The increased utilization of HTS will shift

some of the effort of the aDNA community towards studies of molecular adaptation. Molecular adaptations are changes in gene product activities that contribute to the adaptive phenotype or fitness of an organism [33]. WGS and ES studies will focus effort on molecular adaptation investigations because these datasets contain a large number of candidate variations that may influence gene product activity. Once identified, the functional consequence of a variation can be tested with different molecular biology techniques and some of these techniques are outlined in the discussion below. The molecular techniques included in this discussion pertain exclusively to mammals but similar methodologies are available for other taxonomic groups such as birds [34,35] and plants [36,37].

3.1 Molecular Adaptation – Gene Expression

In a cell, the activity of a gene product can be altered by changing the steady state concentration of the molecule through gene expression.

Regulation of gene expression is very complex with many layers of control. Because of this intricacy there are many different variations within a genome that can influence the expression of a particular gene. In this discussion, only a few of the mechanisms that drive gene expression will be examined.

3.2 Molecular Adaptation – Gene Copy Number Variation

Gene expression can be influenced by gene copy number variation (CNV), which is a structural alteration of the genome where the number of copies of a gene or section of a chromosome is varied [38]. CNV is thought to play a critical role in the evolution of genome complexity and adaptation, as the process provides the raw genetic material needed to generate new functional activity [39]. With gene duplication, each gene

copy is free to mutate and produce a protein or RNA with new activity [39,40]. Gene duplication can have a deleterious effect and in humans the phenomenon is associated with many diseases including psoriasis and lupus glomerulonephritis [41], however gene duplication can be beneficial as having additional gene product from identical genes may produce a phenotype with an improved fitness. In beneficial gene duplication, selective pressure may preserve the same activity in all of the gene copies [40], which can be exemplified by the dietary consequence of farming [42]. Throughout human evolution there has been significant dietary change caused by the development of technologies such as stone tools and agriculture, with starch becoming a prominent nutritional component for many populations [42]. Genomic studies have demonstrated that humans have a considerable variation in CNV for the salivary amylase gene (AMYI), whose product is an enzyme involved with starch metabolism [42,43]. It has also been shown that in humans, high copy numbers for AMYI are positively correlated with increased levels of RNA and protein for the gene [42]. In comparison to humans, chimpanzees (*Pan troglodytes*), which have a low starch diet, possess on average a third fewer copies of the AMYI gene [42]. The high carbohydrate diet found in many human populations appears to have placed selective pressure to preserve the starch metabolism activity for all AMYI copies.

Identification of CNV in aDNA ES and WGS data should be possible using a statistical approach. For modern DNA, analytical tools have been developed to assay CNVs in ES and WGS data using parameters such as read depth coverage [44,45] and it should be possible to adapt these algorithms for aDNA datasets. If genotyping microarrays can be adapted for aDNA, there are analytical approaches that utilize a variety of data including fluorescence intensity and the distance between SNPs to

distinguish CNVs [46]. Once a CNV is identified in aDNA, the functionality of the variation could be examined through several lines of investigation. Comparative studies of different populations could be performed to examine the effect of environmental factors on CNV. For example, the CNV for the AMYI gene in pre-farming hunter gather populations could be compared to the number of copies for the gene in post-farming groups. Functional testing of CNV on phenotype could also be performed by using the cre-lox recombination system to generate transgenic mice with varying number of gene copies [47,48].

3.3 Molecular Adaptation – Mutations in *Cis*-Acting Elements

Molecular adaptation studies of aDNA will also examine genetic variations that affect regulatory mechanisms that control the steady-state concentrations of gene products. In eukaryotic organisms, gene expression is regulated through a complex network of *cis*-acting elements and *trans*-acting factors that act in tandem to control transcription. *Cis*-acting elements are DNA sequences located on the same chromosomal strand as the gene under regulation and can be located near or distant to the transcription start site (TSS). The *cis*-acting elements near the TSS are the core promoter and the proximal *cis*-acting elements. The core promoter is the sequence 40 base pairs upstream and downstream of the TSS whilst the proximal *cis*-acting elements are found in the region between the promoter and approximately1000 base pairs upstream of the TSS. [49,50]. Taken together the core promoter and the proximal *cis*-acting elements are considered to be the promoter region of the gene. Distal *cis*-acting elements are regulatory sequences that can be up to 1 megabase (Mb) upstream from the TSS [51]. *Trans*-acting factors are a complex network of proteins that bind to the *cis*-acting elements to regulate transcription. *Trans*-acting factors include the proteins

that comprise the transcription pre-initiation complex that bind to the core promoter and the transcription factors that associate with the proximal and distal *cis*-acting elements.

Mutations in any of the *cis*-acting elements can change the binding affinities of *trans*-acting factors and thus alter the phenotype of an organism by changing gene expression [52,53]. An example of a mutation in distal *cis*-acting regulatory elements influencing gene expression can be found in humans with the lactase persistence trait. In many human populations, the ability to digest lactose, a sugar found in milk, quickly declines after weaning because of reduced expression of the lactase-phlorizin hydrolase (LPH) gene [54]. In certain European and African populations, expression of LPH persist in adults allowing the digestion of lactose, which provides an adaptive advantage in that milk from domesticated animals remains a food source. LPH persistence is a Mendelian dominant trait and appears to be under the regulation of several SNPs in distal *cis*-acting elements with enhancer activity that are approximately 14 kb upstream from the TSS of the gene. The SNPs that maintain LPH persistence are not the same for all populations [55]. In Europe, the C/T-13910 SNP has a 100% association with LPH persistence in Finnish populations and an 86% to 98% association in other European groups [54,56-59]. An *in vitro* reporter gene assay has also demonstrated that LPH persistence is likely regulated by the C/T-13910 SNP [54,60].

Of the high throughput data types only WGS will contain a substantial amount of information on *cis*-acting elements. Identification of *cis*-acting elements in aDNA data will be aided by the numerous analytical tools that have been developed to study

these sequences in modern genomes including predictive software [61,62] and databases [63,64]. Functional testing of *cis*-acting element variation found in aDNA will be possible with techniques such as the luciferase reporter assay, which makes use of the luciferase enzyme from the firefly *Photinus pyralis* to act as a transcription reporter. In the assay, the *cis*-acting elements that are under investigation are ligated into a vector upstream of the luciferase gene and the vector is then used to transfect mammalian cells. After a certain length of time, the cells are lysed and the activity of luciferase enzyme is measured with the addition of substrate and cofactors. The fluorescent level produced by the luciferase enzyme is approximately proportional to the steady-state mRNA level. Luciferase reporter assay is inexpensive, sensitive, and has a large dynamic range [65].

3.4 Molecular Adaptation – Epigenetic Modifications

Molecular adaptation studies will also examine the epigenetic modification pattern of aDNA. As noted previously, these modifications can regulate gene expression whilst leaving the underlying sequence of DNA unchanged. Epigenetic modifications are biochemical changes to DNA and chromatin associated histones that play a critical role in cell differentiation and these alterations can change throughout the genome as cells transition from a pluripotent to a final differentiated state [66,67]. Environmental stress is known to alter epigenetic modifications, causing altered gene expression and phenotype [68,69]. Changes in epigenetic modification in response to environmental pressure is considered an important process of evolution [67]. The most common epigenetic modification of DNA is methylation of cytosine in a CpG (C—phosphate—G) pair. Methylation can occur in any portion of a gene and the activity of the modification appears to be location dependent [67]. Methylation in the

promoter region is associated with reducing gene expression [70], whilst gene body methylation regulates several aspects of gene expression including splice variants [71] and transcriptional elongation [72]. Lastly, methylation at the 5' end of genes is associated with transcriptional silencing [67,73]. Promoter methylation appears to have evolved in vertebrates, whilst methylation of the gene body was likely present in the last common ancestor of plants and animals [67,74].

Histones are a family of highly alkaline proteins found in eukaryotes that package nuclear DNA into structures called nucleosomes, which are formed by winding ≈ 147 base pairs of DNA around eight histones. Nucleosomes are the basic building blocks of chromatin and chromatin, in turn, is organized into two states: loosely packed euchromatin and tightly packed heterochromatin. Euchromatin is active in gene transcription with the loose packaging of DNA allowing the access of *trans*-acting proteins to *cis*-acting elements. In contrast, heterochromatin is associated with gene silencing, the tight packaging preventing the binding of *trans*-acting proteins to *cis*-acting elements. Chromatin packaging is influenced by many factors including the activity of histone chaperone proteins, ATP-dependent chromatin remodeling enzymes, and epigenetic post-translational modification of histones. Histones can undergo numerous types of post-translational modification and a few of these alterations are known to participate in chromatin remodeling (e.g., methylation and acetylation). However, the function of many of these histone modifications is currently unknown [75].

The mechanisms through which epigenetic modifications act to regulate gene expression are poorly understood, but it has become clear that DNA and histone

epigenetic modifications act in an interactive manner. Methylation of cytosine affects the winding of DNA around histones altering the accessibility of *trans*-acting proteins to *cis*-acting elements [76]. Methylated DNA acts as a binding site for methyl CpG binding proteins, which in turn recruit proteins with histone methyltransferase [77] and histone deacetylase [78] activities. These recruited enzymes then alter the local histone modification pattern causing the chromatin structure to take on a gene silencing conformation [79].

Currently it is possible to obtain data on the epigenetic modifications present in an aDNA sample. As noted earlier in this thesis, there are presently several methodologies to generate cytosine methylation patterns in aDNA HTS data (Chapter IV). Furthermore, information on histone modification can be assayed indirectly from aDNA shotgun HTS data [80]. The winding of DNA around histones appears to protect the nucleic acid from degradation *post-mortem*. DNA molecules that were wound around histones are therefore found in greater abundances in an aDNA sample and consequently produce greater read depth in shotgun HTS data [80]. Loci that produce deeper read depth are therefore from chromatin regions were the histones were modified for tight packaging.

It is not clear how long information on epigenetic modification is retained in aDNA, but evidence of these alterations have been detected in HTS data of a 100 kyr old polar bear. At least under cold preservation, aDNA retains sufficient signal to track genome-wide epigenetic alterations over a considerable timeframe [80]. Currently, there is no practical method to assay the functional differences between different patterns of epigenetic modification observed in aDNA. However, comparative

studies can be performed to identify regions of the genome where modification patterns have undergone change. For example, the DNA methylation and nucleosome packaging patterns could be generated in a time series from bison aDNA. Data from the individuals in this time series could be compared and correlated to climate events to observe how environmental factors have influence the patterns of epigenetic modification. These changing patterns should help identify the genes involved with the bison adapting to the current environmental conditions. Such an epigenetic time series would provide important information on the evolution of the *Bison X* species described in this thesis. The effect of a population bottleneck on epigenetic modifications could be performed by comparing *Bison X* individuals from prior, during, and after the contraction of the species into a refugium (Chapter V).

3.5 Molecular Adaptation – Gene Product Activity

Molecular adaptation studies of aDNA will also examine genetic substitutions that change the activity of a gene product by altering the primary structure of the molecule. It has been well established that non-synonymous substitutions in DNA can alter the activity of a protein by changing the primary amino acid sequence. However, it has become apparent that there are classes of RNAs whose biological activity can be altered through a substitution in the underlying DNA. Approximately 1% of the human genome is protein coding [81], yet it has been estimated that 95% of the genome is transcribed [82]. The function of most of the non-coding RNA is unknown [81,83], but the RNAs with known function are primarily regulators of gene expression and include long non-coding RNA (lncRNA), micro RNA (miRNA), piwi-interacting RNA (piRNA), small nuclear RNAs (snoRNA), small interfering RNA (siRNA), and enhancer RNA (eRNA) [81,84]. Since the majority of the genome is

transcribed into non-coding RNAs most of the genetic variability will be transferred to RNAs and not proteins [82]. Mutations in these biologically active non-coding RNAs have been demonstrated to influence phenotype. For instance, miRNA have a seed sequence that binds to the 3' UTR of mRNA and suppresses translation by facilitating the decay of the transcript [85]. In humans, mutations in the seed region of miRNA-96 are causative agents of progressive hearing loss [86].

Molecular adaptation studies of changes to the primary structure of gene products can be performed with aDNA WGS and ES data. Software has been developed for modern DNA to aid in the analysis of genetic variations that cause alterations to the primary structure of gene products and these algorithms can be applied to aDNA data [87,88]. Once identified, the altered gene product can be reconstructed and compared to a control in a functional assay. The feasibility of reconstruction and functional testing of proteins has previously been demonstrated in a study of mammoth hemoglobin [33]. PCR and Sanger sequencing was used to determine the sequence of the two subunits of hemoglobin from a ≈ 43,000 year-old Siberian mammoth and several non-synonymous substitutions were observed in the β/δ subunit. Functional comparisons were performed by first producing recombinant Asian elephant hemoglobin in *Escherichia coli* using an expression plasmid. Site directed mutagenesis was then used to introduce mammoth-specific substitutions into the Asian hemoglobin vector and the mutated vector was subsequently used to produce recombinant mammoth hemoglobin. In functional testing, the mammoth hemoglobin was found to have a reduced affinity for oxygen at lower temperatures, which is thought to be an evolutionary adaptation to the cold high-latitude environments that the mammoths inhabited [33]. There are a number of technologies that can be used to

test the functionality of proteins and RNAs reconstructed from aDNA. As noted, recombinant proteins can be constructed and studied *in vitro*. Transgenic mice can also be made to produce RNAs and proteins encoded by aDNA sequences using bacterial artificial chromosome or cre-lox recombination systems [89-91].

4.0 Relevance of Thesis to Population-Level and Molecular Adaptation Studies

Many of the technical aspects of the current thesis are relevant for future aDNA population-level and molecular adaptation studies. Molecular techniques will need to be established to optimize data from endogenous aDNA and maintain the fidelity of sequencing libraries. Furthermore, new analytical tools will need to be developed and tested as possible algorithms for processing aDNA data.

4.1 Hybridization Capture

Reducing sequencing requirements through hybridization capture will become a common approach in genome-wide aDNA studies. ES inherently requires hybridization capture and the enrichment procedure will likely become widespread in WGS. Modification of in-house hybridization capture systems such as the method described in this thesis (Chapters II and V) to enrich these large and complex targets will further reduce cost and allow for additional samples to be interrogated.

4.2 Library Fidelity

Library fidelity will continue to be a concern for researchers as loss of genetic variability will be detrimental to aDNA studies. Allelic dropout could reduce the number of informative loci detected in aDNA sequence data and produce inaccuracies in population-level studies. Furthermore, determining the prevalence of mutations that

impart an adaptive advantage in a population will be dependent upon minimizing the loss of genetic diversity during library amplification. Isothermal techniques such as those investigated in this thesis may prove to be a better method than PCR for preserving library fidelity and generating more accurate aDNA data (Chapter III).

Identification of CNVs by read depth coverage will also be dependent on library fidelity. Biased amplification of GC rich regions within genes could cause analysis algorithms to misidentify these loci as having an increased copy number. Amplification methods such as RPA that do not introduce GC bias may generate a more accurate read depth coverage than PCR and improve the identification of true CNVs (Chapter III).

4.3 Identification of Nucleotide Damage

Research on aDNA will continue to be dependent on the accurate identification of SNPs and damaged nucleotides will be a concern, particularly altered residues that act as miscoding lesions. Although, deamination of cytosine has been demonstrated to be a major miscoding lesion, other types of damage that cause misincorporation are likely present in aDNA. The SMRT sequencing data from this thesis (Chapter IV) suggests that there are a variety of modified bases in aDNA of unknown categories. As noted in Chapter IV, gas chromatography/mass spectrometry has previously identified various modified bases in aDNA including 8-OH-Gua [92]. 8-OH-Gua is an analogue of guanine that is formed through oxidative damage [93] and can act as a miscoding lesion [94,95]. 8-OH-Gua does not consistently cause a misincorporation and can insert an adenine or cytosine at ratios of 1:4 to 1:200 depending on the DNA polymerase used for replication [94]. In many types of sequencing data, the

miscoding caused by 8-OH-Gua could be interpreted as heterogeneity. Other forms of damage that cause variable misincorporation may also be present in aDNA. Eventually, in depth investigation of unamplified aDNA using some type of SS-HTS will need to be conducted to generate profiles of the miscoding lesions that can be present in aDNA (Chapter IV). These profiles can then be used to develop repair protocols and analysis algorithms to improve aDNA sequencing data [96]. In the meantime, deep coverage must be used to minimize the misidentification of nucleotides in HTS data.

4.4 Identification of Epigenetic Modifications

Detailed information on the epigenetic modification patterns of DNA and histones of past organisms will greatly aid the study of evolution. Data on these modifications will help researchers identify genes that are contributing to the process of adaptation. Current technology allows for the coarse detection of cytosine methylation but cannot resolve the types of methylation present. For a full understanding of the types of epigenetic modifications present in ancient material, aDNA will have to be examined with technologies such as SS-HTS, which have a greater capacity to identify altered residues than other currently used methodologies (Chapter IV). Analysis of nucleosome packaging in aDNA is dependent on the identification of loci that produce deeper read depth coverage in HTS data and the accuracy of this data type will be dependent on unbiased library amplification (Chapter III).

4.5 Analytical Tools

In this thesis we describe analytical tools to interpret the three state data produced by the Bovid SNP50 BeadChip for classical phylogenetic analysis (Chapter VI). At this

time, it is not clear if these tools will ever be applied to aDNA studies because it may not be possible to generate consistent data using the BovineSNP50 BeadChip to genotype aDNA. However, as discussed in Chapter VI there are molecular methods that may make it possible to accurately assay aDNA with the BeadChip platform and if these techniques can be perfected then the analytical tools developed to interpret the three state data produced by BeadChips will become relevant for aDNA studies.

5.0 Conclusion

Two phylogenetic studies of bison evolution were performed during this thesis. In these studies we identified a new species of extinct bison (*Bison X*) and provided evidence that American plains and woods bison are at least separate sub-species. A possible next step in studying these bison is to generate WGS data for these animals and use these large datasets to perform evolutionary and molecular adaptation studies in parallel. Genotyping data could be analyzed to generate high-resolution phylogenies of these bison and other bovids. WGS data of *Bison X* could be compared to similar data from modern American bison to identify functional differences between the two genomes. Gene products from the two bison species could then be compared to identify the protein or RNA activities that contribute to the specific adaptation of each species. A functional comparison of American woods and plains bison proteins or RNAs would also be an important study of molecular adaptation. These bison are closely related and identifying the functional differences between these animals may provide insight on the forces driving large mammal speciation.

In conclusion, as outlined above, the large amount of information in data from high throughput technologies will be driving aDNA research to a more holistic approach in the study of evolution. Phylogenetic and molecular adaptation studies will both be conducted on the same data. The accuracy of the molecular and analytical techniques used in this holistic approach will be crucial, as these methods will determine the types and quality of the information that can be extracted from the sequence data. Consequently, the development of new techniques to improve the data generated with aDNA is an ongoing affair and this thesis is part of this larger effort to improve ancient methods. Indeed, certain methods described in this thesis offer discernible improvements to protocols currently in use.

References

1. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, et al. (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nature Communications 4: 1-11.
2. Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, et al. (2004) Rise and fall of the Beringian steppe bison. Science 306: 1561-1565.
3. Brandt G, Haak W, Adler CJ, Roth C, Szecsenyi-Nagy A, et al. (2013) Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. Science 342: 257-261.
4. Godinho R, Crespo EG, Ferrand N (2008) The limits of mtDNA phylogeography: complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. Molecular Ecology 17: 4670-4683.
5. Shapiro B, Hofreiter M (2014) A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA. Science 343.
6. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A Draft Sequence of the Neandertal Genome. Science 328: 710-722.
7. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, et al. (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499: 74-78.
8. Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, et al. (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.
9. Jobling M, Hollox E, Hurles M, Kivsild T, Tyler-Smith C (2014) Human Evolutionary Genetics, Second Edition. New York: Garland Science.
10. Haubold B, Wiehe T (2006) How repetitive are genomes? BMC Bioinformatics 7: 541.
11. Carpenter ML, Buenrostro JD, Valdiosera C, Schroeder H, Allentoft ME, et al. (2013) Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries. American Journal of Human Genetics 93: 852-864.
12. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.
13. Teer JK, Mullikin JC (2010) Exome sequencing: the sweet spot before whole genomes. Human Molecular Genetics 19: R145-R151.
14. Castellano S, Parra G, Sanchez-Quinto FA, Racimo F, Kuhlwilm M, et al. (2014) Patterns of coding variation in the complete exomes of three Neandertals. Proceedings of the National Academy of Sciences of the United States of America.
15. Cooper DN, Ball EV, Mort M (2010) Chromosomal distribution of disease genes in the human genome. Genetic Testing and Molecular Biomarkers 14: 441-446.
16. Zhao Z, Fu YX, Hewett-Emmett D, Boerwinkle E (2003) Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. Gene 312: 207-213.
17. Hofreiter M, Barnes I (2010) Diversity lost: are all Holarctic large mammal species just relict populations? BMC Biology 8: 3.

18. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, et al. (2000) The Genetic Legacy of Paleolithic Homo sapiens sapiens in Extant Europeans: A Y Chromosome Perspective. Science 290: 1155-1159.
19. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.
20. Fernandez M, Alarcon GS, Calvo-Alen J, Andrade R, McGwin G, Jr., et al. (2007) A multiethnic, multicenter cohort of patients with systemic lupus erythematosus (SLE) as a model for the study of ethnic disparities in SLE. Arthritis and rheumatism 57: 576-584.
21. Ramos PS, Shaftman SR, Ward RC, Langefeld CD (2014) Genes associated with SLE are targets of recent positive selection. Autoimmune Diseases 2014: 203435.
22. Patterson N, Price AL, Reich D (2006) Population Structure and Eigenanalysis. PLoS Genetics 2: e190.
23. Ringner M (2008) What is principal component analysis? Nature biotechnology 26: 303-304.
24. Nater A, Arora N, Greminger MP, van Schaik CP, Singleton I, et al. (2013) Marked Population Structure and Recent Migration in the Critically Endangered Sumatran Orangutan (Pongo abelii). Journal of Heredity 104: 2-13.
25. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. Nature 461: 489-494.
26. Tollefsrud M, Kissling R, Gugerli F, Johnsen O, Skroppa T, et al. (2008) Genetic consequences of glacial survival and postglacial colonization in Norway spruce: combined analysis of mitochondrial DNA and fossil pollen. Molecular Ecology 17: 4134-4150.
27. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, et al. (2012) Ancient Admixture in Human History. Genetics 192: 1065-1093.
28. Buxbaum JD, Daly MJ, Devlin B, Lehner T, Roeder K, et al. (2012) The autism sequencing consortium: large-scale, high-throughput sequencing in autism spectrum disorders. Neuron 76: 1052-1056.
29. Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, et al. (2013) Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genetics 9: e1004023.
30. Georges M (2014) Towards sequence-based genomic selection of cattle. Nature Genetics 46: 807-809.
31. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, et al. (2013) Ancient human genomes suggest three ancestral populations for present-day Europeans. bioRxiv.
32. Hale ML, Burg TM, Steeves TE (2012) Sampling for Microsatellite-Based Population Genetic Studies: 25 to 30 Individuals per Population Is Enough to Accurately Estimate Allele Frequencies. PLoS ONE 7: e45170.
33. Campbell KL, Roberts JE, Watson LN, Stetefeld J, Sloan AM, et al. (2010) Substitutions in woolly mammoth hemoglobin confer biochemical properties adaptive for cold tolerance. Nature Genetics 42: 536-540.
34. Kang SW, Gazzillo LC, You S, Wong EA, El Halawani ME (2004) Turkey prolactin gene regulation by VIP through 35-bp cis-acting element in the proximal promoter. General and Comparative Endocrinology 138: 157-165.

35. Tyack SG, Jenkins KA, O'Neil TE, Wise TG, Morris KR, et al. (2013) A new method for producing transgenic birds via direct in vivo transfection of primordial germ cells. Transgenic Research 22: 1257-1264.
36. Qin Y, Tian Y, Han L, Yang X (2013) Constitutive expression of a salinity-induced wheat WRKY transcription factor enhances salinity and ionic stress tolerance in transgenic Arabidopsis thaliana. Biochemical and Biophysical Research Communications 441: 476-481.
37. Matthews BF, Saunders JA, Gebhardt JS, Lin J-J, Koehler SM (1995) Reporter Genes and Transient Assays for Plants Plant Cell Electroporation and Electrofusion Protocols. pp. 147-162.
38. Haraksingh RR, Snyder MP (2013) Impacts of Variation in the Human Genome on Gene Regulation. Journal of Molecular Biology 425: 3970-3977.
39. Ohno S (1970) Evolution by gene duplication. New York: Springer-Verlag.
40. Zhang J (2003) Evolution by gene duplication: an update. Trends in Ecology & Evolution (Personal edition) 18: 292-298.
41. Alkan C, Kidd J, Marques-Bonet T, Aksay G, Antonacci F, et al. (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics 41: 1061 - 1067.
42. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nature Genetics 39: 1256-1260.
43. Groot PC, Mager WH, Frants RR, Meisler MH, Samuelson LC (1989) The human amylase-encoding genes amy2 and amy3 are identical to AMY2A and AMY2B. Gene 85: 567-568.
44. Mccallum KJ, Wang J-P (2013) Quantifying copy number variations using a hidden Markov model with inhomogeneous emission distributions. Biostatistics 14: 600-611.
45. Plagnol V, Curtis J, Epstein M, Mok K, Stebbings E, et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28: 2747 - 2754.
46. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. (2007) PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research 17: 1665-1674.
47. Ermakova O, Salimova E, Piszczek L, Gross C (2012) Construction and phenotypic analysis of mice carrying a duplication of the major histocompatibility class I (MHC-I) locus. Mammalian Genome 23: 443-453.
48. Herault Y, Duchon A, Marechal D, Raveau M, Pereira PL, et al. (2010) Controlled somatic and germline copy number variation in the mouse model. Current Genomics 11: 470-480.
49. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics 7: 29-59.
50. Atkinson TJ, Halfon MS (2014) Regulation of gene expression in the genomic context. Computational and Structural Biotechnology Journal 9: e201401001.
51. Symmons O, Spitz F (2013) From remote enhancers to gene regulation: charting the genome's regulatory landscapes. Philosophical Transactions of the Royal Society B: Biological Sciences 368: 20120358.
52. Kwasnieski JC, Mogno I, Myers CA, Corbo JC, Cohen BA (2012) Complex effects of nucleotide variants in a mammalian cis-regulatory element.

Proceedings of the National Academy of Sciences of the United States of America 109: 19498-19503.
53. Yokoyama KD, Thorne JL, Wray GA (2011) Coordinated Genome-Wide Modifications within Proximal Promoter Cis-regulatory Elements during Vertebrate Evolution. Genome Biology and Evolution 3: 66-74.
54. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39: 31-40.
55. Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow DM (2009) Lactose digestion and the evolutionary genetics of lactase persistence. Human Genetics 124: 579-591.
56. Enattah NS, Trudeau A, Pimenoff V, Maiuri L, Auricchio S, et al. (2007) Evidence of Still-Ongoing Convergence Evolution of the Lactase Persistence T-13910 Alleles in Humans. The American Journal of Human Genetics 81: 615-625.
57. Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, et al. (2003) The Causal Element for the Lactase Persistence/ non-persistence Polymorphism is Located in a 1 Mb Region of Linkage Disequilibrium in Europeans. Annals of Human Genetics 67: 298-311.
58. Hogenauer C, Hammer HF, Mellitzer K, Renner W, Krejs GJ, et al. (2005) Evaluation of a new DNA test compared with the lactose hydrogen breath test for the diagnosis of lactase non-persistence. European Journal of Gastroenterology & Hepatology 17: 371-376.
59. Ridefelt P, Hakansson LD (2005) Lactose intolerance: lactose tolerance test versus genotyping. Scandinavian Journal of Gastroenterology 40: 822-826.
60. Lewinsky RH, Jensen TGK, Møller J, Stensballe A, Olsen J, et al. (2005) T−13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Human Molecular Genetics 14: 3945-3953.
61. Lee TY, Chang WC, Hsu JB, Chang TH, Shien DM (2012) GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group. BMC Genomics 13 Suppl 1: S3.
62. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E (2012) MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics 28: 487-494.
63. Zhao F, Xuan Z, Liu L, Zhang MQ (2005) TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Nucleic Acids Research 33: D103-107.
64. Xuan Z, Zhao F, Wang J, Chen G, Zhang MQ (2005) Genome-wide promoter extraction and analysis in human, mouse, and rat. Genome Biology 6: R72.
65. Smale ST (2010) Luciferase Assay. Cold Spring Harbor Protocols 2010: pdb.prot5421.
66. Hochedlinger K, Plath K (2009) Epigenetic reprogramming and induced pluripotency. Development 136: 509-523.
67. Duncan EJ, Gluckman PD, Dearden PK (2014) Epigenetics, plasticity, and evolution: How do we link epigenetic change to phenotype? Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 322: 208-220.

68. Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, et al. (2012) DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. Proceedings of the National Academy of Sciences of the United States of America 109: 4968-4973.
69. Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, et al. (2012) Reversible switching between epigenetic states in honeybee behavioral subcastes. Nature Neuroscience 15: 1371-1373.
70. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature Reviews Genetics 13: 484-492.
71. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, et al. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479: 74-79.
72. Lorincz MC, Dickerson DR, Schmitt M, Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nature Structural & Molecular Biology 11: 1068-1075.
73. Brenet F, Moh M, Funk P, Feierstein E, Viale AJ, et al. (2011) DNA Methylation of the First Exon Is Tightly Linked to Transcriptional Silencing. PLoS ONE 6: e14524.
74. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, et al. (2010) Conservation and divergence of methylation patterning in plants and animals. Proceedings of the National Academy of Sciences of the United States of America 107: 8689-8694.
75. Rothbart SB, Strahl BD (2014) Interpreting the language of histone and DNA modifications. Biochimica et Biophysica Acta.
76. Lee JY, Lee T-H (2011) Effects of DNA Methylation on the Structure of Nucleosomes. Journal of the American Chemical Society 134: 173-175.
77. Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, et al. (2003) The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. The Journal of Biological Chemistry 278: 4035-4040.
78. Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, et al. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393: 386-389.
79. Joulie M, Miotto B, Defossez PA (2010) Mammalian methyl-binding proteins: what might they do? Bioessays 32: 1025-1032.
80. Pedersen JS, Valen E, Velazquez AM, Parker BJ, Rasmussen M, et al. (2014) Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Research 24: 454-466.
81. Kornienko AE, Guenzl PM, Barlow DP, Pauler FM (2013) Gene regulation by the act of long non-coding RNA transcription. BMC Biology 11: 59.
82. Halvorsen M, Martin JS, Broadaway S, Laederach A (2010) Disease-Associated Mutations That Alter the RNA Structural Ensemble. PLoS Genetics 6: e1001074.
83. Pauli A, Rinn JL, Schier AF (2011) Non-coding RNAs as regulators of embryogenesis. Nature Reviews Genetics 12: 136-149.
84. Mousavi K, Zare H, Koulnis M, Sartorelli V (2014) The emerging roles of eRNAs in transcriptional regulatory networks. RNA Biology 11: 106-110.
85. Ameres SL, Zamore PD (2013) Diversifying microRNA sequence and function. Nature Reviews Molecular Cell Biology 14: 475-488.

86. Mencia A, Modamio-Hoybjor S, Redshaw N, Morin M, Mayo-Merino F, et al. (2009) Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nature Genetics 41: 609-613.
87. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research 38: e164.
88. Zorc M, Jevsinek Skok D, Godnic I, Calin GA, Horvat S, et al. (2012) Catalog of MicroRNA Seed Polymorphisms in Vertebrates. PLoS ONE 7: e30737.
89. Belkaya S, van Oers NSC (2014) Transgenic Expression of MicroRNA-181d Augments the Stress-Sensitivity of CD4+CD8+ Thymocytes. PLoS ONE 9: e85274.
90. Baker A, Cotten M (1997) Delivery of bacterial artificial chromosomes into mammalian cells with psoralen-inactivated adenovirus carrier. Nucleic Acids Research 25: 1950-1956.
91. Zhou Y, Grinchuk O, Tomarev SI (2008) Transgenic Mice Expressing the Tyr437His Mutant of Human Myocilin Protein Develop Glaucoma. Investigative Ophthalmology & Visual Science 49: 1932-1939.
92. Hoss M, Jaruga P, Zastawny TH, Dizdaroglu M, Pääbo S (1996) DNA Damage and DNA Sequence Retrieval from Ancient Tissues. Nucleic Acids Research 24: 1304-1307.
93. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA (1992) 8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G----T and A----C substitutions. Journal of Biological Chemistry 267: 166-172.
94. Shibutani S, Takeshita M, Grollman AP (1991) Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. Nature 349: 431-434.
95. McCulloch SD, Kokoska RJ, Garg P, Burgers PM, Kunkel TA (2009) The efficiency and fidelity of 8-oxo-guanine bypass by DNA polymerases δ and η. Nucleic Acids Research 37: 2830-2840.
96. Dabney J, Meyer M, Pääbo S (2013) Ancient DNA Damage. Cold Spring Harbor Perspectives in Biology 5: 1-6.

nature COMMUNICATIONS

ARTICLE

Received 20 Sep 2012 | Accepted 27 Feb 2013 | Published 23 Apr 2013

DOI: 10.1038/ncomms2656

Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans

Paul Brotherton[1,2,*], Wolfgang Haak[1,*], Jennifer Templeton[1], Guido Brandt[3], Julien Soubrier[1], Christina Jane Adler[1,†], Stephen M. Richards[1], Clio Der Sarkissian[1,†], Robert Ganslmeier[4], Susanne Friederich[4], Veit Dresely[4], Mannis van Oven[5], Rosalie Kenyon[6], Mark B. Van der Hoek[6], Jonas Korlach[7], Khai Luong[7], Simon Y.W. Ho[8], Lluis Quintana-Murci[9], Doron M. Behar[10], Harald Meller[4], Kurt W. Alt[3], Alan Cooper[1] & The Genographic Consortium[‡]

Haplogroup H dominates present-day Western European mitochondrial DNA variability (>40%), yet was less common (~19%) among Early Neolithic farmers (~5450 BC) and virtually absent in Mesolithic hunter-gatherers. Here we investigate this major component of the maternal population history of modern Europeans and sequence 39 complete haplogroup H mitochondrial genomes from ancient human remains. We then compare this 'real-time' genetic data with cultural changes taking place between the Early Neolithic (~5450 BC) and Bronze Age (~2200 BC) in Central Europe. Our results reveal that the current diversity and distribution of haplogroup H were largely established by the Mid Neolithic (~4000 BC), but with substantial genetic contributions from subsequent pan-European cultures such as the Bell Beakers expanding out of Iberia in the Late Neolithic (~2800 BC). Dated haplogroup H genomes allow us to reconstruct the recent evolutionary history of haplogroup H and reveal a mutation rate 45% higher than current estimates for human mitochondria.

[1] The Australian Centre for Ancient DNA, School of Earth and Environmental Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia. [2] Archaeogenetics Research Group, School of Applied Sciences, University of Huddersfield, Huddersfield HD1 3DH, UK. [3] Institute of Anthropology, Colonel-Kleinmann Weg 2, Johannes Gutenberg University Mainz, D-55128 Mainz, Germany. [4] State Office for Heritage Management and Archaeology Saxony-Anhalt/State Museum for Prehistory Halle, Richard-Wagner-Straße 9, D-06114 Halle/Saale, Germany. [5] Department of Forensic Molecular Biology, Erasmus MC, University Medical Centre Rotterdam, 3000 CA Rotterdam, The Netherlands. [6] SA Pathology, SA Health, Adelaide, South Australia 5000, Australia. [7] Pacific Biosciences, Menlo Park, California 94025, USA. [8] School of Biological Sciences, The University of Sydney, Sydney, New South Wales 2006, Australia. [9] Institut Pasteur, Unit of Evolutionary Genetics, 75015 Paris, France. [10] Rambam Medical Centre, 31096 Haifa, Israel. * These authors contributed equally to this work. † Present address: Institute of Dental Research, Westmead Centre for Oral Health, The University of Sydney, Sydney, New South Wales 2145, Australia (C.J.A.); Centre for Geogenetics, Natural History Museum of Denmark, 1350 Copenhagen, Denmark (C.D.S.). Correspondence and requests for materials should be addressed to P.B. (email: p.m.brotherton@hud.ac.uk) or to W.H. (email: wolfgang.haak@adelaide.edu.au).
‡A full list of authors for the Genographic Consortium and their affiliations appears at the end of the paper.

A key unanswered issue in human prehistory is the extent to which cultural change identifiable in the archaeological record can be ascribed to the movements of people, as opposed to the movements of just their ideas and artefacts. The Central European archaeological record identifies a succession of profound cultural and economic changes between the last hunter-gatherers of the Mesolithic and the first farmers of the Early Neolithic (ENE), through to the socially stratified chiefdoms of the Early Bronze Age[1–3]. The exact nature and genetic context of the transformative changes that took place over these four millennia remain unclear[4,5], although current genetic patterns of mitochondrial DNA (mtDNA) haplogroup (hg) distribution suggest a complex series of events in European prehistory[4–9] and hint at multiple inputs from outside Central Europe[4,10,11].

Phylogeographic studies suggest that mt hg H arrived in Europe from the Near East before the Last Glacial Maximum (22,000 BP), and survived in glacial refugia in Southwest Europe before undergoing a post-glacial re-expansion[4,12]. Haplogroup H now accounts for over 40% of mtDNA variation in anatomically modern humans across much of Western Eurasia, with declining frequencies south and east to ~10–30% in the Near East and Caucasus[10]. However, it remains uncertain when and how H became the dominant European hg. Traditional approaches (including ancient DNA studies) have been unable to resolve either the phylogeny or phylogeographic distribution of H sub-haplogroups (sub-hgs)[6], however, they have generally relied on sequencing only 300–400 bp of the mt D-loop or control region[10,13]. A number of studies based on complete 16.6 kb human mt genomes have revealed a complex evolutionary history for hg H (for example, refs 12,14–18, with phylogenetic analyses recognizing 87 H sub-hgs[19]). These complete mt genomes revealed that 71% of hg H polymorphic diversity is located outside the D-loop, in the coding region[20] and, as a result, this diversity has not yet been exploited at the population genetics level.

To investigate the relationship between the European genetic and archaeological records, we sequenced whole hg H mt genomes from skeletal remains directly assigned to distinct Central European archaeological cultures. Owing to its excellently preserved human skeletal remains, forming a continuous record across a series of archaeological cultures since Palaeolithic times, the Mittelelbe-Saale region of Saxony-Anhalt (Germany) provided a unique opportunity to address this issue. We analysed a time transect spanning the >3,500 years of the Central European Neolithic period (Table 1, Supplementary Table S1), from the first farmers of the ENE linear pottery culture (LBK, 5450–4775 BC), through the subsequent Rössen (4625–4250 BC), Schöningen (4100–3950 BC), Baalberge (3950–3400 BC) and Salzmünde (3400–3025 BC) cultures. These were followed by two of the first pan-European Late Neolithic (LNE) cultural complexes, the Corded Ware (CWC, 2800–2050 BC) and Bell Beaker (BBC, 2500–2050 BC) cultures, before the emergence of the Early Bronze Age with the Unetice culture (2200–1575 BC). We chose to focus on hg H because of its recent dramatic rise in frequency to become the dominant hg in Europe, because of its presence in all Neolithic cultures in the Mittelelbe-Saale region, and the potential it provided to explore detailed genetic structure on a sub-hg level. Overall, our results suggest that the broad foundations of the Central European mtDNA pool, here approximated via hg H, were formed during the Neolithic rather than the post-glacial period.

Table 1 | Summary of genotyping data against the Reconstructed Sapiens Reference Sequence (RSRS).

Culture/age	Individual	Hg*	Hg H sequence variants compared with RSRS
LBK (5450–4775 BC)	HAL36	H23	**C10211T**
	HAL11	H	T16093C, G16129A!
	HAL32	H26	**T11152C**
	HAL39	H1e	**G3010A**, **G5460A**
	DEB9	H88	**A8596G**
	DEB21	H1j	**G3010A**, **T4733C**
	KAR6a	H1bz	G1719A, **G3010A**, C14380T
	KAR11b	H	T152C!
	KAR16a	H46b	**C2772T**, A11893G
Rössen (4625–4475/4250 BC)	OSH2	H89	A6932G, C8068T, T12696C
	OSH3	H1	**G3010A**
	OSH1	H16	**T152C!**, **C10394T**
	OSH7	H5b	**C456T, G5471A, T16304C**, C16519T
Schöningen (4100–3950 BC)	SALZ18a	H10i	C13503T, **T14470a**, T**16093C**
	SALZ21b	H1e7	T1766C, **G3010A**, **G5460A**
Baalberge (3950–3400 BC)	ESP30	H1e1a5	**G3010A**, **G5460A**, (C5960T), **A8512G**, G8865A, **C14902T**
	HQU4	H7d5	**A4793G**, **C15409T**, G16388A
Salzmünde (3400–3100/3025 BC)	SALZ57a	H3	T152C!, **T6776C**
	SALZ77a	H3	**T6776C**
Corded Ware (2800–2200/2050 BC)	ESP15	H6a1a	**T239C**, **G3915A**, **A4727G**, **G9380A**, **T11253C**, **T16362C**, **A16482G**, C16519T
	BZH6	H1_TBD	**G3010A**, A8149G, A9377G, T9467C, A13671G, T14319C, **T16189C!**
Bell Beaker	BZH4	H1e7	**G3010A**, **G5460A**, A15220G, A15401G, A16293G
(2500–2200/2050 BC)	ROT6	H5a3	**C456T**, **G513A**, **T4336C**, **G15884A**, **T16304C**, C16519T
	ALB1	H3b	**A2581G**, **T6776C**
	ROT1	H3ao2	C4577T, **T6776C**, **C16256T**
	ROT2	H5a3	**C456T**, **G513A**, **T4336C**, **G15884A**, **T16304C**, C16519T
	QUEXII1	H4a1	**C3992T**, **A4024G**, **T5004C**, **G9123A**, **C14365T**, **A14582G**, C16519T
	QUEXII2	H4a1	**C3992T**, **A4024G**, **T5004C**, **G9123A, C14365T, A14582G**, C16519T
	QLB26a	H1	**G3010A**
	QUEXII3	H13a1a2c	**C2259T**, **A4745G**, G9025A, **A13542G**, **C13680T**, **C14872T**, C16519T
	QLB28b	H1	**G3010A**
Unetice (2200–1575 BC)	BZH1	H11a	**T195C!, T961g, T8448C, (G13759A), A16293G, T16311C!**, C16519T
	BZH8	H2a1a3	**G951A**, **G1438A**, **G4769A**, **C6173T**, **T13095C**, A16240t, **C16354T**, C16519T
	BZH14	H82a	T195C!, **A16220G**
	EUL41a	H4a1a1a5	**A73G!**, **C3992T**, **A4024G**, **T5004C**, **G8269A**, **G9123A**, **A10044G**, C13545T, **C14365T**, **A14582G**, C16519T
	EUL57B	H3	T152C!, **T6776C**
	QUEVIII4	H7h	**A4793G**, **G16213A**
Nuragic Bronze Age (1624 BC)	—	H1aw1	**G3010A**, **A8701G!**, C15912T
Iron Age (500 BC)	—	H90	C5435T, T8911C, T10237C, T15109C

Abbreviation: SNP, single-nucleotide polymorphism.
Sub-haplogroup defining diagnostic SNPs are shown in bold and 'private'/as-yet-unknown sequence variants in regular print.
*Haplogroup H designations based on the http://www.phylotree.org mtDNA tree Build 14 (5 April 2012)[19,20].

Results

Sequence and network analyses. From a collection of over 400 European prehistoric human archaeological remains we selected 37 Mittelelbe-Saale individuals, as well as two samples from Italy (Supplementary Table S1), previously assigned to hg H by simplex and multiplex PCR[7]. Work was independently replicated for two samples per individual (Supplementary Methods). We

designed and optimised a hybridisation-based DNA-capture system to sequence complete mt genomes on the Affymetrix MitoChip v2.0 (ref. 21) (Supplementary Methods, Supplementary Fig. S1,S2, and Supplementary Tables S2-S3) via immortalised libraries prepared from the highly damaged and degraded endogenous DNA recovered from archaeological remains[22,23]. Six of the 39 target-enriched libraries were also analysed via a single-molecule, real-time (SMRT[24]) Pacific Biosciences *RS* sequencing platform (Supplementary Table S4, Supplementary Dataset). In addition, 35/391 (9%) of all SNPs identified via the MitoChip were independently confirmed by direct PCR and Sanger sequencing (Supplementary Methods, Supplementary Tables S5 and S6). Mt genomes from all 39 individuals were unambiguously assignable to individual sub-hgs of hg H[20], confirming that a single human was typed in each case (Table 1). The mt hypervariable region I sequences matched those previously determined for each individual. The ancient hg H mt genomes were highly diverse, with 34 distinct haplotypes attributed to 20 major sub-hgs (gene diversity H = 0.997 +/− 0.0071; nucleotide diversity 0.000421 +/− 0.000225), including three novel lineages (provisionally named H88–H90).

Phylogenetic network analysis of these ancient mt genomes reveals evidence of dynamic changes in the composition of H sub-hgs over the ~3,500-year time transect (Fig. 1). Importantly, sequences from older samples (and cultures) tend to represent basal lineages, only one to three mutations away from the ancestral root of hg H, while younger samples (after ~4000 BC) largely comprise more derived haplotypes appearing on longer

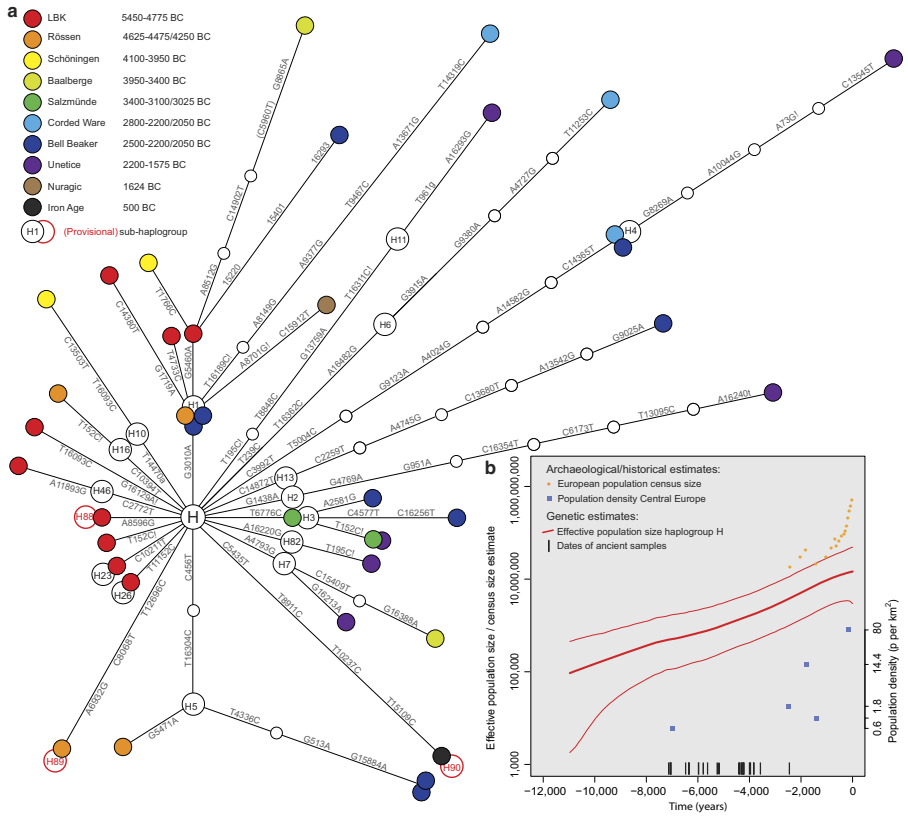


Figure 1 | Mitochondrial haplogroup H sequence evolution. (**a**) Phylogenetic network of 39 prehistoric mitochondrial genomes sorted into two temporal groupings: Early Neolithic (left) and Mid-to-Late Neolithic (right). Node colours represent archaeological cultures. (**b**) A Bayesian skyride plot of 200 representative present-day and 39 ancient hg H mt genomes (the thick red line denotes the posterior median, thinner flanking lines denote the 95% credibility interval; note the logarithmic scale of the *y* axis). Prehistoric samples (18 radiocarbon and 21 mean archaeological dates) served as internal calibration points (black bars). For comparison, census size estimates for the European population are shown as orange dots. Population density estimates from the archaeological record for key periods in Central Europe are plotted as blue squares in chronological order: LBK, Iron Age, Roman period, Merovingian and Pre-industrial modern times (*y* axis on the right)[28].

branches. This temporal relationship provides further support for the authenticity of the ancient mt genomes.

Network analysis (Fig. 1) reveals pronounced differences in the composition of sub-hgs between the ENE cultures (LBK, Rössen, Schöningen), and those of the Mid Neolithic (MNE)/LNE to Early Bronze Age (Baalberge, Salzmünde, Corded Ware, Bell Beaker, Unetice). ENE (and in particular LBK) mt genomes are either rare today (H16, H23 and H26), extinct or have not yet been observed in present-day populations (H46b, H88 and H89). In sharp contrast, most of the later H sub-hgs are more common in present-day European populations (for example, hg H3, H4, H6, H7, H11 and H13)[12,14–16]. Of the 39 haplotypes detected, only three (within the common, basal, sub-hg H1) were shared between ENE and MNE/LNE cultures. As the observed gene diversity is high, we might expect the number of shared haplotypes within and between cultures to be low. However, as the MNE/LNE haplotypes are on different sub-hg branches from the ENE haplotypes, these patterns combined show minimal local genetic continuity over this time period (Table 1).

Genetic distances. To further examine these apparent temporal shifts in sub-hg distribution, we tested whether hg H individuals represent different meta-populations by pooling them into different cultural and/or temporal groups of ENE versus LNE (Table 2, Supplementary Table S7). When pooled in four groups (ENE, MNE, LNE and Bronze Age), pairwise population comparisons via F_{ST} values based on sequence data showed that genetic distances increased with time over the duration of the Neolithic, reaching a significant value ($F_{ST} = 0.08722$; $P = 0.00386 +/- 0.0006$) between the ENE and the early Bronze Age (Table 2). This suggests a transformation of hg H diversity during the Neolithic period. This effect was less apparent (non-significant F_{ST} values) when samples from various sites were pooled in larger temporal groups (Table 2). However, non-parametric multivariate analysis of variance (NP-MANOVA, $P = 0.0072$) also confirmed a significant difference between pooled groups of ENE and LNE individuals when comparisons were based on the presence or absence of sub-hgs (Table 2).

Genetic affinities. To examine potential geographic origins for Neolithic cultures (Supplementary Table S1) and to assess their contribution to present-day Central European mtDNA diversity, we used principal component analysis (PCA) to investigate genetic affinities between three ancient culturally/temporally pooled groups (LBK, MNE and BBC) and 37 present-day Western Eurasian populations (Supplementary Table S8). PCA of the frequencies of the 15 most common H sub-hgs showed that the present-day populations form three significantly supported geographic clusters (a grouping which was also supported using NP-MANOVA, $P < 0.0001$; Table 2: (i) Iberia in the west; (ii) the Caucasus, the Near East and Anatolia; and (iii) Central and Eastern Europe from the Urals to France (Fig. 2a). This particular number of clusters was also the best supported in a model-based test on sub-hg H frequencies followed by Ward clustering (Fig. 2c,d). We also used Procrustes analysis to quantify the relationship between hg H substructure and the geographic locations of both the present-day Western Eurasian and the Mittelelbe-Saale ancient populations. For this analysis, we superimposed the PCA coordinates on the geographic map of the present-day and ancient sampling locations. We found a striking resemblance between the genetic and geographic maps, with a highly significant Procrustes similarity score ($t_0 = 0.733$) obtained for the comparison ($P < 10^{-6}$; 100,000 permutations). The analysis supported a clustering of the transformed genetic data from present-day populations into the three major groups described above (Fig. 2b). In contrast, Procrustes analysis clearly showed that the genetic data for LBK and BBC samples were not related to their geographic location. Although all three ancient groups were sampled from the same Central European location only the MNE group genetically resembles present-day populations from this region.

The combined set of analyses (PCA, Procrustes and Ward clustering) revealed that Mittelelbe-Saale's earliest farmers (LBK; $n = 9$) cluster with present-day Caucasus, Near Eastern and Anatolian populations, as previously noted[7]. In contrast, individuals from the successor series of regional post-LBK (and MNE) Rössen, Schöningen, Baalberge and Salzmünde cultures (ca. 4625–3025 BC, MNE; $n = 10$) cluster with present-day Central European populations (Fig. 2). Mitochondrial genomes from BBC individuals in Mittelelbe-Saale (BBC; $n = 7$) display close genetic affinities to present-day Iberian populations (Fig. 2). The component loadings of the PCA biplot indicate that this is largely based on high frequencies of sub-hgs H1 and H3, which are thought to have spread from a glacial Iberian refugium[13] and which have also been reported from ancient Neolithic sites from France and Spain[8,25]. Other LNE samples add further to the genetic complexity. Individuals from the CWC (2800–2200 BC), which has archaeological associations towards North-Eastern Europe, produced two distinct mt genomes (H1_TBD and H6a1a), which have not been found in their contemporaneous Bell Beaker neighbours, nor in preceding Central European cultures. Similarly, data from the subsequent Early Bronze Age

Table 2 | Population pairwise and linearised Slatkin's F_{ST}s and NP-MANOVA tests.

(a) NP-MANOVA four time periods ($P = 0.0696$)

	Early Neolithic	Middle Neolithic	Late Neolithic	Bronze Age
Early Neolithic (13)	0	0.1262	0.024	0.0574
Middle Neolithic (6)	**0.7572**	0	0.8575	0.7782
Late Neolithic (9)	**0.144**	**1**	0	0.742
Bronze Age (6)	**0.3444**	**1**	**1**	0

F_{ST} four time periods F_{ST}

	Early Neolithic	Middle Neolithic	Late Neolithic	Bronze Age
Early Neolithic (13)	0	0	*0.03179*	*0.09555*
Middle Neolithic (6)	0.01135	0	0	*0.02299*
Late Neolithic (9)	0.02247	− 0.01165	0	*0.01148*
Bronze Age (6)	**0.08722**	0.03081	− 0.02250	0

(b) NP-MANOVA LBK, BBC and pooled intermediate Neolithic (MNE) as used in PCA ($P = 0.2355$)

	LBK	MNE	BBC
LBK (9)	0	0.2084	0.0916
MNE (10)	**0.6252**	0	0.8025
BBC (7)	**0.2748**	**1**	0

F_{ST} LBK, BBC, and pooled intermediate Neolithic (MNE) as used in PCA

	LBK	MNE	BBC
LBK (9)	0	0	*0.03369*
MNE (10)	− 0.02587	0	0
BBC (7)	0.03260	− 0.00704	0

(c) NP-MANOVA two time periods ($P = 0.0072$)

	Early Neolithic	Late Neolithic
Early Neolithic (13)	0	0.0109
Late Neolithic (16)	**0.0109**	0

F_{ST} two time periods

	Early Neolithic	Late Neolithic
Early Neolithic (13)	0	*0.01459*
Late Neolithic (16)	0.01438	0

(d) NP-MANOVA Cultures grouped with geographic regions as in Fig. 2 *($P < 0.0001$)*

	Iberia	Near East	Mainland Europe
Iberia	0	0	0.0001
Near East	**0**	0	0.0004
Mainland Europe	**0.0003**	**0.0012**	0

Abbreviations: BBC, Bell Beaker culture; MNE, Mid Neolithic; NP-MANOVA, non-parametric multivariate analysis of variance; PCA, principal component analysis.
Neolithic samples pooled in different time periods: (a) four time periods; (b) three time periods; (c) two time periods, and (d) from cultures grouped with geographic regions. For NP-MANOVA, Bonferroni corrected values are given in bold print and areas shaded grey indicate significant values ($P < 0.05$). Slatkin F_{ST}s are italicised (upper diagonal) and significant pairwise distances are given in bold print (lower diagonal).

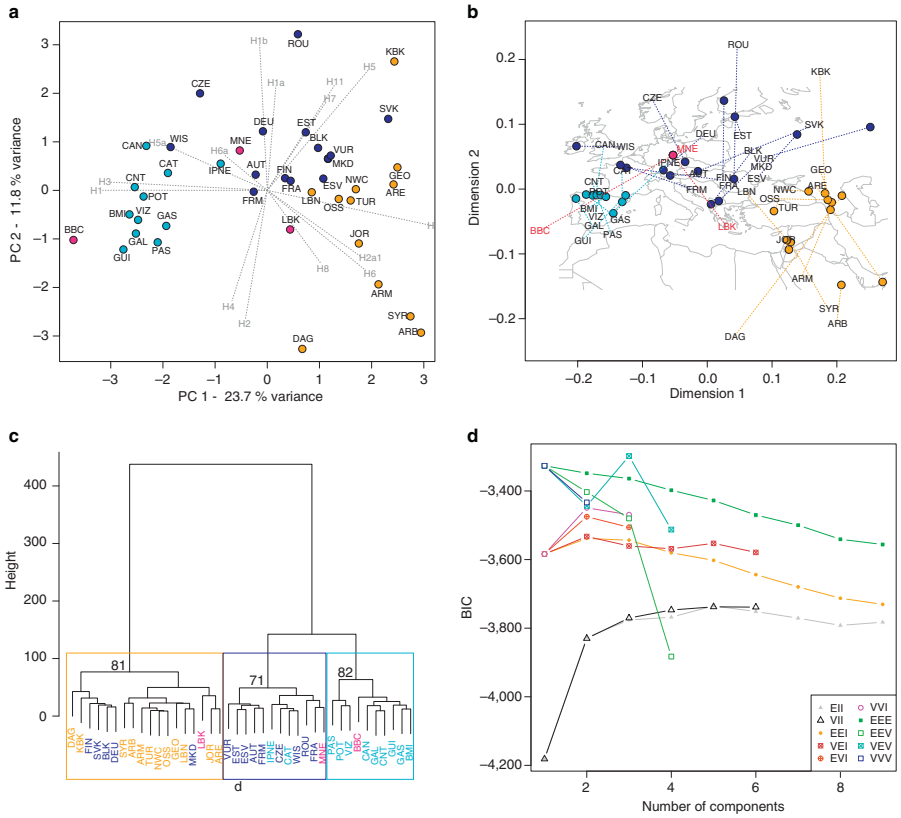


Figure 2 | Population affinities of select Neolithic cultures. (**a**) PCA biplot based on the frequencies of 15 hg H sub-haplogroups (component loadings) from 37 present-day Western Eurasian and three ancient populations (light blue: Western Europe; dark blue: Central and Eastern Europe; orange; Near East, Caucasus and Anatolia; and pink: ancient samples). Populations are abbreviated as follows: GAL, Galicia; CNT, Cantabria; CAT, Catalonia; GAS, Galicia/Asturia; CAN, Cantabria2; POT, Potes; PAS, Pasiegos; VIZ, Vizcaya; GUI, Guipuzcoa; BMI, Basques; IPNE, Iberian Peninsula Northeast; TUR, Turkey; ARM, Armenia; GEO, Georgia; NWC, Northwest Caucasus; DAG, Dagestan; OSS, Ossetia; SYR, Syria; LBN, Lebanon; JOR, Jordan; ARB, Arabian Peninsula; ARE, Arabian Peninsula2; KBK, Karachay-Balkaria; MKD, Macedonia; VUR, Volga-Ural region; FIN, Finland; EST, Estonia; ESV, Eastern Slavs; SVK, Slovakia; FRA, France; BLK, Balkans; DEU, Germany; AUT, Austria, ROU, Romania; FRM, France Normandy; WIS, Western Isles; CZE, Czech Republic; LBK, Linear pottery culture; BBC, Bell Beaker culture; MNE, Middle Neolithic. (**b**) Procrustes analyses of geographic coordinates and PCA scores of the same data set (similarity score $t_0 = 0.733$, $P < 10^{-6}$, 100,000 permutations). (**c**) Ward clustering dendrogram of the three ancient groups and present-day populations (colour code as above and p values in % of approximately unbiased bootstrapping for the following three main clusters). (**d**) Results of the model-based test to identify the number of clusters by the model with the highest support (highest Bayes Information criterion (BIC); VEV = multivariate mixture model (ellipsoidal, equal shape)).

Unetice culture revealed haplotypes with genetic affinities to both the East (sub-hg H2a, H7 and H11) and the West (sub-hg H3 and H4), based on frequency distributions of these sub-hgs in present-day populations[13]. We also included two individuals from outside Central Europe (Sardinia and South Tyrol) and from different time periods (Nuragic Bronze Age and Iron Age, respectively) to further investigate genetic diversity within hg H and to test the power of resolution of complete mt genomes. Both individuals had mt genomes that are not found in samples from the Mittelelbe-Saale region. The Iron Age sample from South Tyrol produced another new sub-hg (provisional H90) and the Bronze Age individual from Sardinia a new H1 haplotype (H1aw1).

Reconstructing the demographic history of mtDNA hg H. It has previously proved difficult to use present-day data alone to determine when hg H became the predominant hg in Europe, as archaeogenetic and palaeodemographic reconstructions have very large uncertainties[4,26]. However, as our 39-dated ancient mt genome sequences provide precise temporal calibration points,

we performed a Bayesian skyride analysis with 200 random present-day mt genome sequences to reconstruct the lineage history of hg H through time (with the caveat of assuming a continuous and panmictic population). The resulting skyride plot (Fig. 1b) is the first real-time estimation of the European hg H population size (and consequently its contribution to Europe's effective population size and demographic history) with a broad temporal coverage over ~3,500 years of the Neolithic period in Central Europe (5500–2000 BC). Hg H shows a consistent and strong exponential growth over the entire course of the Neolithic. The estimated population size tracks the European census size[27] and population density estimates from archaeological sites[28] in the Late Holocene, but also provides detailed estimates for prehistoric times for which data points remain very scarce (Fig. 1b).

Another major advantage of the temporal calibration points provided by ancient hg H mt genomes is that the data allow a relatively precise estimate of the evolutionary substitution rate for human mtDNA. The temporal dependency of evolutionary rates predicts that rate estimates measured over short timespans will be considerably higher than those using deep fossil calibrations, such as the human/chimpanzee split at ~6 million years[29]. The rate calibrated by the Neolithic and Bronze Age sequences is 2.4×10^{-8} substitutions per site per year ($1.7–3.2 \times 10^{-8}$; 95% high posterior density) for the entire mt genome, which is 1.45 × (44.5%) higher than current estimates based on the traditional human/chimp split (for example, 1.66×10^{-8} for the entire mt genome[30] and 1.26×10^{-8} for the coding region[31]). Consequently, the calibrated 'Neolithic' rate infers a considerably younger coalescence date for hg H (10.9–19.1 kya) than those previously reported (19.2–21.4 kya for HVSI[10], 15.7–22.5 kya for the mt coding region[31] or 14.7–22.6 kya when corrected for purifying selection[30]).

Discussion

Despite recent successes in sequencing portions of nuclear genomes from Meso- and Neolithic samples[11,32,33], mtDNA remains the most widely studied and best described marker in population genetics. Although its interpretation is limited to the matrilineal genetic history[4,13,19], this can be an important socio-cultural and demographic signal additional to that gained from autosomal loci[34,35]. Our results clearly demonstrate that high-resolution full mt genome-typing, combined with the ability to analyse large numbers of individuals from multiple cultural layers, can provide highly resolved temporal views that are not yet practical with nuclear DNA studies.

The phylogenetic network analysis of our chronological hg H mt genome data set (Fig. 1a) provides the first detailed real-time view of mutations in human mtDNA. It has enabled the direct observation of the mutation rate over thousands of years and revealed a distinct temporal distribution pattern of hg H diversity. Although a temporal pattern could be expected in an expanding population with stable/increasing hg H frequencies (Fig. 1b, Supplementary Fig. S3), ENE and MNE/LNE/Bronze Age samples clearly show a mutually exclusive sub-hg distribution with the exception of sub-hg H1, which is the most common and basal sub-hg within H[14–16]. Under an assumption of genetic continuity, we would expect MNE/LNE and Bronze Age individuals to be on the same sub-hg branches as ENE individuals. Instead, ENE mt genomes are generally either rare today[19] or have not yet been observed in present-day populations, possibly owing to subsequent extinction of these lineages. In contrast, most MNE/LNE and Bronze Age sub-hgs are still common today. This suggests that individuals from the ENE made a marginal contribution to LNE and present-day hg H diversity. Although the relatively small sample numbers from each time period limit detailed analyses of the causes of the distribution shifts, we interpret this phylogenetic pattern as a genetic discontinuity between Early and subsequent Neolithic cultures in Europe, potentially mirroring genetic structure in Neolithic European populations. Genetic drift could also have played a role in generating discrepant hg distributions over time and space. However, if drift was the sole cause we would expect a random distribution across all sub-hgs rather than a clear distinction between ENE and MNE/LNE/Bronze Age mt genomes.

Our genetic distance data also indicate minimal local genetic continuity between the ENE and the MNE/LNE in Central Europe (Fig. 1; Table 1), again suggesting that ENE lineages were largely superseded during the MNE/LNE (~4100–2200 BC) in a previously unrecognised major genetic transition. This pronounced genetic changeover between ENE and MNE/LNE cultures is comparable to other known major genetic transition, thus far revealed by ancient DNA and coalescent simulations (between indigenous European hunter-gatherers and incoming early farmers from the Near East during the initial Meso-Neolithic transition from ~7500 BC in Central Europe)[6,7]. When compared with hg H diversity in present-day Central Europe[14,15,18,36], the network in Fig. 1 suggests that much of the present-day diversity can be attributed to the incorporation of new lineages in the MNE/LNE and emerging Bronze Age (from 2200 BC). The LNE in particular is known to have been a period of profound cultural and economic change[37], with newly emerging pan-European cultures such as the Bell Beaker phenomenon in Western Europe and the Corded Ware culture in north-eastern Europe. It therefore seems likely that these pan-European cultures were associated with the introduction of lineages from outside Central Europe. Fortunately, the ranges of both these groups overlapped in the Mittelelbe-Saale sample area (Supplementary Methods), allowing this possibility to be further investigated.

Our data on genetic affinities (PCA, Procrustes and Ward clustering) revealed that Mittelelbe-Saale's earliest farmers (LBK; $n = 9$) cluster with present-day Caucasus, Near Eastern, and Anatolian populations. These findings are consistent with a highly detailed archaeological record tracing the temporal and spatial spread of agriculture into Central Europe; beginning initially in Anatolia and the Near East, where farming originated ~12,000 years ago[7].

Our observation that individuals from the successor series of regional post-LBK and MNE cultures (Rössen, Schöningen, Baalberge and Salzmünde) cluster with present-day Central European populations could be explained by a loss of lineages from the ENE LBK period during a short phase of population decline in the centuries after 5000 BC (as proposed in some archaeological models)[38]. However, our results suggest that mtDNA H sub-hg diversity established during the MNE is still present in Central European populations today. This is consistent with independent archaeological evidence of a phase of more localised cultural development during the MNE period, potentially involving influences from contemporaneous MNE cultures outside Mittelelbe-Saale, which (perhaps in concert with LBK population decline) could have resulted in a replacement of most ENE H sub-hgs. Together, the genetic and archaeological evidence highlight the complexities of both the formative and consolidation phases in Central Europe.

From around 2800 BC, the LNE Bell Beaker culture emerged from the Iberian Peninsula to form one of the first pan-European archaeological complexes. This cultural phenomenon is recognised by a distinctive package of rich grave goods including the eponymous bell-shaped ceramic beakers. The genetic affinities between Central Europe's Bell Beakers and present-day Iberian

populations (Fig. 2) is striking and throws fresh light on long-disputed archaeological models[3]. We suggest these data indicate a considerable genetic influx from the West during the LNE. These far-Western genetic affinities of Mittelelbe-Saale's Bell Beaker folk may also have intriguing linguistic implications, as the archaeologically-identified eastward movement of the Bell Beaker culture has recently been linked to the initial spread of the Celtic language family across Western Europe[39]. This hypothesis suggests that early members of the Celtic language family (for example, Tartessian)[40] initially developed from Indo-European precursors in Iberia and subsequently spread throughout the Atlantic Zone; before a period of rapid mobility, reflected by the Beaker phenomenon, carried Celtic languages across much of Western Europe. This idea not only challenges traditional views of a linguistic spread of Celtic westwards from Central Europe during the Iron Age, but also implies that Indo-European languages arrived in Western Europe substantially earlier, presumably with the arrival of farming from the Near East[41].

Other LNE population movements appear to have added further genetic complexity, as exemplified by the CWC (2800–2200 BC), which preceded the Bell Beaker culture in Mittelelbe-Saale and has archaeological associations with North-Eastern Europe. A genetic affinity to eastern populations is consistent with two distinct CWC mt genomes (H1_TBD and H6a1a) not identified in either their contemporaneous Bell Beaker neighbours or in preceding Central European cultures. The subsequent Early Bronze Age Unetice culture, associated with emerging metallurgy and increasingly stratified societies[37,42], marks a consolidation of social and cultural systems in Mittelelbe-Saale that were established during the LNE by the two pan-European Bell Beaker and CWCs. The Unetice culture appears contemporaneously with the last Neolithic horizon (~2200 BC) in areas where elements of both the Bell Beaker and CWCs are present, sometimes overlapping at the same sites. It is therefore not surprising that individuals ascribed to the newly emerging Unetice culture carry mt genomes with both Western (sub-hgs H3 and H4) and Eastern (sub-hgs H2a, H7 and H11) associations.

The demographic reconstruction, which is based on direct calibration points, has major implications for understanding post-glacial human history in Europe. Our new estimate is incompatible with traditional views that the majority of present-day hg H lineages were carried into Central, Northern and Eastern Europe via a post-glacial human population expansion before the Holocene (12 kya)[13]. Our data complement a recent study, based on present-day mt genomes, which describes a pronounced population increase at ~7000 BC (interpreted as a Neolithic expansion into Europe), but followed by a slow population growth until the present day[26]. By including ancient DNA data from across the critical time points in question, our skyride plot corrects for missing temporal data and suggests substantial growth of hg H from the beginning of the Neolithic and continuing throughout the entire Neolithic period. This emphasizes the role of farming practices and cultural developments in the demographic expansions inferred in subsequent time periods, which have not yet been explored genetically.

Although an expansion of hg H could in principle be compatible with a post-glacial resettling of Northern and Central Europe from southwestern refugia[12,16] (as indicated by our population skyride and PCA plots), we instead propose that the rise of hg H to become the predominant mtDNA branch in Europe was mediated by subsequent demographic events during the Neolithic, as shown by a general increase in hg H frequency and strong population growth during this period (Fig. 1b). Support for this position comes from data suggesting that hg H was virtually absent among Central and Northern European hunter-gatherers[6,43] and formed only 19% in LBK individuals, most likely introduced from Southeast Europe and/or the Near East[7]. In our updated data set from Mittelelbe-Saale, hg H appears to have been established by the LBK period and increased in frequency after 4000 BC (Supplementary Fig. S3). Interestingly, MNE/LNE cultures with cultural associations to the North and Northeast, such as the Bernburg and CWCs, show reduced hg H frequencies and hg H only moved northwards into southern Scandinavia during the Neolithisation of Northern Europe around the Middle Neolithic, as exemplified by individuals from the Funnel Beaker Culture[11,43]. However, hg H appears to have been generally more frequent in prehistoric Western Europe: at 20% from a Middle Neolithic (3030–2890 calBC) site in France[9]; at ~25% from Iberian (Epi-)Cardial Neolithic samples[8,25]; at 36% from a Neolithic site in Catalonia[44]; and at 44% from Neolithic sites from the Basque Country and Navarre[45]. Importantly, a recent study on Iberian hunter-gatherers revealed the presence of hg H there in Mesolithic times[45]. In Mittelelbe-Saale, the Bell Beaker samples signpost a significant increase in hg H frequency (the 95% confidence intervals do not overlap with earlier LBK and Schöningen Neolithic cultures; Supplementary Fig. S3). In conclusion, the Western European Neolithic and the widespread pan-European Bell Beaker phenomenon appear to be important factors in driving the spread of H sub-hgs throughout large parts of Western Europe. In particular, high proportions of sub-hgs H1 and H3 seem to have made substantial contributions to the hg H diversity that exists in Western and Central Europe today[16]. Having reached significant levels, and assuming a generally higher rate of population growth in southern and western Europe in post-Neolithic times[27], these Neolithic processes appear to have been the major factor in hg H becoming the predominant European mtDNA hg.

Overall, our results suggest that the broad foundations of the Central European mtDNA pool, here approximated via hg H, were formed during the Neolithic rather than the post-glacial period. ENE hg H mt lineages brought in from the Near East by Central Europe's first farmers do not appear to have contributed significantly to present-day Central Europe's hg H diversity, instead being largely superseded during the MNE and LNE (with the process starting around 4000 BC), after which there appears to have been substantial genetic continuity to the present-day in Central Europe. These developments have been revealed by comparative full mt genome sequencing and would have remained obscure using standard HVS I data.

In conclusion, demographic changes across the MNE, followed by the widespread Bell Beaker cultural phenomenon, are likely to have been the key factors in the expansion of hg H across Western Europe and the eventual rise of hg H to become the predominant mtDNA hg. However, LNE Corded Ware and Early Bronze Age data suggest a complex series of additional genetic contributions, which require further investigation.

Methods

Ancient DNA analyses. DNA was extracted from two independent samples for each individual (Supplementary Methods). HVS I was amplified using a minimum of four short overlapping primer pairs, following established protocols and authentication criteria as described previously[7,46]. Multiplex SNP typing of 22 hg informative SNPs (GenoCoRe22) was carried out using a SNaPshot-based protocol as described previously[7].

Ancient DNA Library preparation. Ancient DNA extract polishing, phosphorylation, adaptor ligation and polymerase 'fill-in' reactions were used sequentially to create fully double-stranded adaptor-tagged aDNA libraries (Fig. 3). Following every step, DNA was purified using MinElute spin columns (Qiagen) as per the

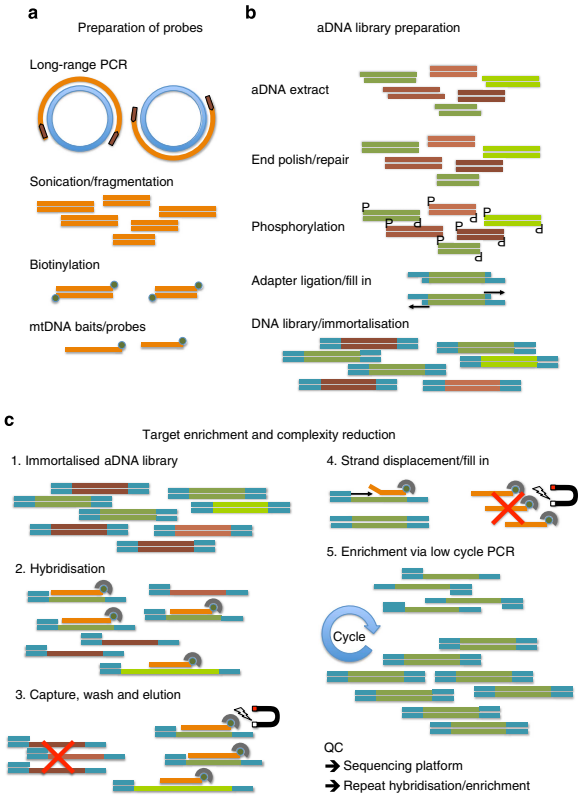


Figure 3 | Schematic representation of experimental steps. (**a**) Probe DNA was prepared by amplifying a complete mitochondrial genome in two overlapping fragments by long-range PCR, followed by DNA fragmentation and biotinylation to form mtDNA 'baits' for targeted hybridisation. (**b**) Ancient DNA was enzymatically blunt-ended and phosphorylated, ligated to custom library adaptors, followed by polymerase 'fill-in' to create 'immortalised' double-stranded DNA libraries. (**c**) Hybridisation-based DNA-capture using biotinylated probe bound to Streptavidin magnetic beads; following stringency washes, captured library constructs enriched in mtDNA sequences are eluted from the beads/probe via a novel polymerase strand-displacement reaction followed by PCR library reamplification. These steps can be carried out iteratively to maximise mtDNA content in enriched libraries (see Supplementary Methods for full details).

manufacturer's instructions. PCR amplification reactions were then performed to create 'primary' DNA libraries, ready for DNA-capture hybridisation steps, and amplification products were sized and quantified (Supplementary Methods).

Hybridisation-based enrichment of human mtDNA. The basic conceptual design for the hybridisation of tracer DNA (aDNA library) to biotinylated driver DNA sequences (human mt probe) was previously described[47] and the overall scheme is outlined in Fig. 3. The two library-specific PCR primers were included as part of the hybridisation mix as blocking oligonucleotides to minimise unwanted hybridisation between the adaptor-tagged flanking regions of otherwise unrelated single-stranded library DNA molecules[48]. A key innovation of this methodology was the use of a DNA polymerase with strand-displacing activity after post-hybridisation stringency washes. This allowed primer extension from the bound library (blocking) primers to disrupt the double-stranded region of stable hybridisation between human mt probe DNA sequences and single-stranded library DNA molecules that had inserts with complementary sequences. These mtDNA-enriched library DNA molecules captured in the hybridisation step could thereby be cleanly separated from biotinylated probe molecules, which remained bound to magnetic Streptavidin beads. PCR reamplification reactions from the mt-enriched library DNA molecules comprised the 'first enrichment' DNA libraries. In general, we used three cycles of hybridisation/enrichment/reamplification to produce DNA libraries highly enriched for short endogenous mtDNA sequence fragments ready for genotyping (Supplementary Methods).

Affymetrix Mitochip v2.0 array typing and Pacific Biosciences SMRT sequencing. MtDNA-enriched libraries underwent biotin labelling using terminal deoxynucleotidyl transferase (TdT) as per the Affymetrix GeneChip Whole-Transcript Sense Target Labelling Assay Manual (P/N 701880, rev. 4). Biotin-labelled DNA libraries were hybridised to Affymetrix GeneChip Human mt Resequencing 2.0 Arrays for 17 h at 49 °C. Arrays were washed, stained and scanned as per the GeneChip CustomSeq Resequencing Array Protocol (P/N 701231, rev. 5). Affymetrix GeneChip Command Console software (v3.2) was used to generate CEL files, which were then analysed using GeneChip Sequence Analysis Software (GSEQ v4.1, Affymetrix) and validated using the software Geneious[49] (Supplementary Fig. S1,S2, Supplementary Tables S2,S3). Six of the mt-enriched

DNA libraries were also converted to SMRTbell template libraries for sequencing on a Pacific Biosciences *RS* platform (Supplementary Methods).

Network analyses. A median joining network of all ancient hg H mt genomes (Fig. 1a) was constructed manually using the most up-to-date version of the mt phylogenetic tree (PhyloTree.org, mtDNA tree Build 14) as a scaffold on which to place the observed hg H lineages[19,20]. This version included a revised version of the hg H sub-tree comprising 1203 sequences in total. As per convention, insertions at np 309.1C(C), 315.1C, 523-524d (aka 522-523d), 16182C, 16183C, 16193.1C(C) and mutation 16519 were not considered for phylogenetic reconstruction[20].

Procrustes-based PCA and Ward Clustering. PCA was used to describe and visualise the maternal genetic relationships among the Neolithic cultures investigated, as well as to 37 present-day European and Near Eastern populations (Fig. 2a). PCA was performed on the frequency of H sub-hgs taken from the literature (Supplementary Table S8). To minimise statistical noise caused by rare sub-hgs and to allow for data compatibility across published studies, we considered only the following 15 most common H sub-hgs in Europe and the Near East: H*, H1, H1a, H1b, H2, H2a1, H3, H4, H5, H5a, H6, H6a, H7, H8 and H11. PCAs were performed and visualised in R version 2.11.1 (ref. 50) using a customised script based on the function prcomp.

Ancient hg H individuals were pooled into three different groups based on the numbers of samples available: two for 'pan-European' archaeological phenomena/cultures alongside hypothesised geographic origins (LBK, $n=9$ and BBC, $n=7$); and a temporally transitional group pooling regional (mostly MNE) cultures (MNE, $n=10$). Small sample sets such as the Corded Ware ($n=2$) and later Bronze Age Unetice ($n=5$) were excluded. To test whether the clustering pattern observed in the PCA was significantly supported, we performed a number of statistical tests including Ward clustering, Procrustes analysis and NP-MANOVA (as described below). First, we performed model-based cluster tests to identify the number of clusters via the model with the best support (highest Bayes Information criterion) followed by Ward hierarchical clustering of sub-hg H frequencies using the packages mclust, pvclust (for bootstrap values) and hclust in R, respectively. Procrustes analysis was also performed in R using the package vegan based on PCA scores and geographic coordinates (Supplementary Table S8) and the function protest to calculate the similarity score (100,000 permutations).

Summary statistics. Population pairwise F_{ST}, Slatkin's linearised F_{ST} and haplotype diversity were calculated in Arlequin version 3.5 (ref. 51). We used jMODELTEST 0.1.1[52] in order to find the best fitting evolutionary model and, if required, to estimate a discrete γ shape parameter for our 39 non-partitioned mt genomes. Based on the resulting scores for each model (AIC and Bayes Information criterion), we subsequently used the Tamura and Nei model and a γ value of 0.049 for our calculations of population distances in Arlequin. The ancient hg H individuals were pooled into different temporal/cultural groups in order to calculate genetic diversity indices and to test for genetic differentiation (Table 2, Supplementary Table S7).

Multivariate analysis of variance. We performed a NP-MANOVA to test whether the temporal grouping of ancient individuals according to archaeological time periods are statistically supported. The NP-MANOVA was performed on a Raup-Crick distance matrix, which was produced from the presence/absence of the 15 hg H sub-hgs used in the PCA. Calculations were performed in PAST version 2.09 with 10,000 permutations per test and *post hoc* Bonferroni correction to account for multiple comparisons and small sample sizes[53]. We also tested whether the clustering pattern between the ancient and present-day populations observed in the PCA was significantly supported.

Bayesian skyride analyses and mutation rate calculation. The data set comprised 37 newly sequenced, non-related, ancient mt genomes, five sets of randomly chosen, distinct, present-day hg H mt genomes from Phylotree (http://www.phylotree.org, mtDNA tree Build 12 (20th July 2011)) and 420 newly available hg H sequences[17]. The Sequences were manually aligned to the revised Cambridge Reference Sequence (rCRS: AC_000021)[54] using the program SeaView[55]. The alignment was partitioned into four subsets, representing the D-loop, the protein-coding regions (1st + 2nd codon positions and 3rd codon position) and a concatenation of tRNA and RNA genes. Insertions at nps 309.1C(C), 315.1C, 523-524d (aka 522-523d), 16182C, 16183C, 16193.1C(C) were not considered for phylogenetic reconstruction and position 16519 was removed from the D-loop subset[20]. The best substitution models were selected using ModelGenerator 0.85 (ref. 56). by comparison of Bayesian Information Criterion scores: HKY + G for D-loop, TN + G for protein-coding regions and HKY for RNA genes. Considering the short evolutionary timescale being studied (intra-hg), models including a proportion of invariant sites were excluded. A Bayesian skyride analysis[57] was performed using the phylogenetic software BEAST 1.6.1 (ref. 58). and calibrated using radiocarbon dates from 18 of the ancient individuals and mean archaeological dates for the remaining individuals. This allowed us to achieve a broad temporal coverage for ~3500 years of the Neolithic period in Central Europe (5500–2000 BC) and to generate the most precise demographic reconstruction of hg H. Results were replicated using independent sets of 100 (1 ×), 200 (3 ×), and 300 (1 ×) mt genomes. A strict molecular clock was used, allowing for a distinct rate in each subset of the alignment. Additional analysis using an uncorrelated log normal relaxed clock to account for potential rate variations could not reject the strict clock assumption. Convergence was checked by sampling from two independent Markov chains. Each MCMC analysis was run for 100,000,000 steps and samples from the two chains were combined, after discarding the first 10% of samples as burn-in. All parameters showed sufficient sampling, indicated by effective sample sizes above 200. Tracer 1.5 was used to produce the skyride plot (Fig. 1b)[59].

We carried out a 'date randomisation test', to test whether the signal from the radiocarbon dates associated with the ancient sequences was sufficient to calibrate the hg H phylogeny[60]. This test randomises all dates associated with the sequences (including present-day ones) and replicates of the phylogenetic analysis as described above. If the structure and spread of the ancient sequences in the tree were sufficient to calibrate the analysis, the inferred mean rate of the randomised analysis should be significantly different from the rate calculated using the correct association date/sequence. In other words, the 95% HPD of the randomised analysis should not overlap with the mean rate estimated without randomisation. The comparison of estimated rates from the main analysis and from 10 replicates with randomised dates presented in Supplementary Fig. S4 confirms the presence of sufficient signal to calibrate the tree provided by dates from the 37 ancient samples.

References

1. Whittle, A. W. R. & Cummings, V. *Going over: The Mesolithic-Neolithic Transition in North-West Europe* 632 (Oxford University Press, Oxford, 2007).
2. Sherratt, A. Plough and pastoralism: aspects of the secondary products revolution. *Patterns of the Past: Studies in honour of David Clarke*. In: Hodder, I., Isaac, G. & Hammond, N. (eds). 261–305 (Cambridge University Press, Cambridge, 1981).
3. Bogucki, P. I. & Crabtree, P. J. *Ancient Europe 8000 B.C.-A.D. 1000: Encyclopedia of the Barbarian World,* 1221p (Charles Scribner's Sons, 2004).
4. Soares, P. *et al.* The archaeogenetics of Europe. *Curr. Biol.* **20,** R174–R183 (2010).
5. Pinhasi, R., Thomas, M. G., Hofreiter, M., Currat, M. & Burger, J. The genetic history of Europeans. *Trends genet.* **28,** 496–505 (2012).
6. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326,** 137–140 (2009).
7. Haak, W. *et al.* Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. *PLoS Biol.* **8,** e1000536 (2010).
8. Gamba, C. *et al.* Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. *Mol. Ecol.* **21,** 45–56 (2012).
9. Lacan, M. *et al.* Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proc. Natl Acad. Sci. USA* **108,** 9788–9791 (2011).
10. Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67,** 1251–1276 (2000).
11. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336,** 466–469 (2012).
12. Pereira, L. *et al.* High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res.* **15,** 19–24 (2005).
13. Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H. J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* **22,** 339–345 (2006).
14. Roostalu, U. *et al.* Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: The near eastern and Caucasian perspective. *Mol. Biol. Evol.* **24,** 436–448 (2007).
15. Loogväli, E. L. *et al.* Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol. Biol. Evol.* **21,** 2012–2021 (2004).
16. Achilli, A. *et al.* The molecular dissection of mtDNA haplogroup H confirs that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. *Am. J. Hum. Genet.* **75,** 910–918 (2004).
17. Behar, D. M. *et al.* The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since Pre-Neolithic times. *Am. J. Hum. Genet.* **90,** 486–493 (2012).
18. Alvarez-Iglesias, V. *et al.* New population and phylogenetic features of the internal variation within mitochondrial DNA macro-haplogroup R0. *PLoS ONE* **4,** e5112 (2009).
19. Behar, D. M. *et al.* A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).
20. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).
21. Hartmann, A. *et al.* Validation of microarray-based resequencing of 93 worldwide mitochondrial genomes. *Hum. Mutat.* **30,** 115–122 (2009).
22. Maricic, T., Whitten, M. & Paabo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5,** e14004 (2010).

23. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
24. Korlach, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol.* **472,** 431–455 (2010).
25. Lacan, M. *et al.* Ancient DNA suggests the leading role played by men in the Neolithic dissemination. *Proc. Natl Acad. Sci. USA* **108,** 18255–18259 (2011).
26. Fu, Q., Rudan, P., Pääbo, S. & Krause, J. Complete mitochondrial genomes reveal Neolithic Expansion into Europe. *PLoS ONE* **7,** e32473 (2012).
27. Livi-Bacci, M. *A Consice History of World Population* 279 (Blackwell Publishing, Malden, Oxford, Carlton, 2007).
28. Zimmermann, A., Hilpert, J. & Wendt, K. P. Estimations of population density for selected periods between the Neolithic and AD 1800. *Hum. Biol.* **81,** 357–380 (2009).
29. Ho, S. Y., Shapiro, B., Phillips, M. J., Cooper, A. & Drummond, A. J. Evidence for time dependency of molecular rate estimates. *Syst. Biol.* **56,** 515–522 (2007).
30. Soares, P. *et al.* Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am. J. Hum. Genet.* **84,** 740–759 (2009).
31. Mishmar, D. *et al.* Natural selection shaped regional mtDNA variation in humans. *Proc. Natl Acad. Sci. USA* **100,** 171–176 (2003).
32. Keller, A. *et al.* New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat. Commun.* **3,** 698 (2012).
33. Sanchez-Quinto, F. *et al.* Genomic affinities of two 7,000-year-old iberian hunter-gatherers. *Curr. Biol.* **22,** 1494–1499 (2012).
34. Behar, D. M. *et al.* The Basque paradigm: genetic evidence of a maternal continuity in the Franco-Cantabrian region since pre-Neolithic times. *Am. J. Hum. Genet.* **90,** 486–493 (2012).
35. Gunnarsdottir, E. D. *et al.* Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat. Commun.* **2,** 228 (2011).
36. Brandstätter, A. *et al.* Dissection of mitochondrial superhaplogroup H using coding region SNPs. *Electrophoresis* **27,** 2541–2550 (2006).
37. Heyd, V. Families, prestige goods, warriors & complex societies: Beaker groups of the 3rd millennium cal BC along the upper & middle Danube. *Proc. Prehist. Soc.* **73,** 327–379 (2007).
38. Shennan, S. & Edinborough, K. Prehistoric population history: from the late glacial to the late neolithic in central and northern Europe. *J. Archaeol. Sci.* **34,** 1339–1345 (2007).
39. Cunliffe, B. & Koch, J. T. (eds) *Celtic from the West: Alternative Perspectives from Archaeology, Genetics, Language and Literature,* 384 (Oxbow Books, Oxford, 2010).
40. Koch, J. T. *Tartessian. Celtic in the South-west at the Dawn of History* (Aberystwyth, 2009).
41. Renfrew, C. *Archaeology and Language: The Puzzle of Indo-European Origins.* XIV, 346 S. Ill., Kt(Cape, London, 1988).
42. Nowak, M. Transformations in East-Central Europe from 6000 to 3000 BC: local vs. foreign patterns. *Documenta Praehistorica* **XXXIII,** 143–158 (2006).
43. Malmström, H. *et al.* Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr. Biol.* **19,** 1758–1762 (2009).
44. Sampietro, M. L. *et al.* Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. *Proc. Biol. Sci./Royal Soc.* **274,** 2161–2167 (2007).
45. Hervella, M. *et al.* Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PLoS ONE* **7,** e34417 (2012).
46. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310,** 1016–1018 (2005).
47. Patel, M. & Sive, H. PCR-based subtractive cDNA cloning. *Curr. Protocols Mol. Biol.*. Chapter 25, Unit **25B,** 2 (2001).
48. Tao, S. C., Gao, H. F., Cao, F., Ma, X. M. & Cheng, J. Blocking oligo--a novel approach for improving chip-based DNA hybridization efficiency. *Mol. Cell. Probes* **17,** 197–202 (2003).
49. Drummond, A. J. *et al.* Geneious v5.4, Available from http://www.geneious.com/ (2011).
50. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2010).
51. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resources* **10,** 564–567 (2010).
52. Posada, D. jModelTest: phylogenetic model averaging. *Mol. Biol. Evol.* **25,** 1253–1256 (2008).
53. Hammer, O., Harper, D. A. T. & Ryan, P. D. PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4** (2001).
54. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet* **23,** 147 (1999).
55. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27,** 221–224 (2010).
56. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McLnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6,** 29 (2006).
57. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25,** 1459–1471 (2008).
58. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
59. Rambaut, A. & Drummond, A. Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer (2007).
60. Ho, S. Y. *et al.* Bayesian estimation of substitution rates from ancient DNA sequences with low information content. *Syst. Biol.* **60,** 366–375 (2011).

Acknowledgements

We are indebted to Matt Kaplan and Ryan Spriggs at Arizona Research Laboratories, Division of Biotechnology, University of Arizona Genetics Core Facility, http://uagc.arl.arizona.edu/, Tyson Clark, Michael Brown, Kristi Spittle and Matthew Boitano (Pacific Biosciences) for sequencing work, Jeremy Timmis for help with DNA sonication protocols, and Robin Skeates, and Hubert Steiner for additional samples and contextual information. We thank the Australian Research Council (grant LP0882622), the Deutsche Forschungsgemeinschaft (Al 287/7-1 and Me 3245/1-1) and National Geographic's Genographic Project for funding. M.v.O. was supported in part by the Netherlands Forensic Institute (NFI) and a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

Author contributions

P.B., W.H. and A.C. conceived and designed the project. P.B. designed and developed the DNA extraction, DNA library construction and hybridisation-based DNA-capture protocols (with assistance from J.T.). P.B., J.T. and W.H. generated and analysed the data. S.M.R., C.D., R.K. and M.B.v.d.H. contributed experimental steps and C.J.A., J.S., S.Y.W.H., J.K. and K.L. contributed analytical steps. G.B., R.G., S.F., V.D., M.v.O., L.Q., D.M.B., H.M. and K.W.A. provided ancient samples, contextual information, radiocarbon dating and access to critical population data. P.B., W.H. and A.C. wrote the manuscript with input from C.J.A., J.S., S.Y.W.H., S.M.R., J.K. and members of the Genographic Consortium. All authors discussed the paper and gave comments.

Additional information

Accession codes: The complete consensus mt genome sequences have been deposited to NCBI GenBank under accession numbers KC553980 to KC554018.

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

Competing financial interests: The authors claim no competing financial interests associated with this paper.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat. Commun.* 4:1764 doi: 10.1038/ncomms2656 (2013).

Syama Adhikarla[11], Arun Kumar Ganesh Prasad[11], Ramasamy Pitchappan[11], Arun Varatharajan Santhakumari[11], Elena Balanovska[12], Oleg Balanovsky[12], Jaume Bertranpetit[13], David Comas[13], Begoña Martínez-Cruz[13], Marta Melé[13], Andrew C. Clarke[14], Elizabeth A. Matisoo-Smith[14], Matthew C. Dulik[15], Jill B. Gaieski[15],

Amanda C. Owings[15], Theodore G. Schurr[15], Miguel G. Vilar[15], Angela Hobbs[16], Himla Soodyall[16], Asif Javed[17], Laxmi Parida[17], Daniel E. Platt[17], Ajay K. Royyuru[17], Li Jin[18], Shilin Li[18], Matthew E. Kaplan[19], Nirav C. Merchant[19], R. John Mitchell[20], Colin Renfrew[21], Daniela R. Lacerda[22], Fabrício R. Santos[22], David F. Soria Hernanz[23], R. Spencer Wells[23], Pandikumar Swamikrishnan[24], Chris Tyler-Smith[25], Pedro Paulo Vieira[26] & Janet S. Ziegle[27]

[11] The Genographic Laboratory, School of Biological Sciences, Madurai Kamaraj University, Madurai 625 021, Tamil Nadu, India. [12] Research Centre for Medical Genetics, Russian Academy of Medical Sciences, 115478 Moscow, Russia. [13] Institut de Biologia Evolutiva (CSIC-UPF), Departament de Ciences de la Salut i de la Vida, Universitat Pompeu Fabra, 08003 Barcelona, Spain. [14] Department of Anatomy, University of Otago, Dunedin 9054, New Zealand. [15] Department of Anthropology, University of Pennsylvania, Philadelphia, Pennsylvania, 19104-6398, USA. [16] National Health Laboratory Service, Sandringham 2131, Johannesburg, South Africa. [17] IBM, Yorktown Heights, New York 10598, USA. [18] School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, People's Republic of China. [19] Arizona Research Laboratories, University of Arizona, Tucson, Arizona 85721, USA. [20] Department of Genetics, School of Molecular Sciences, La Trobe University, Melbourne, Victoria 3086, Australia. [21] McDonald Institute for Archaeological Research, University of Cambridge, Cambridge, CB2 3ER, UK. [22] Departamento de Biologia Geral, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, CEP 31270-901, Brazil. [23] National Geographic Society, Washington, District of Columbia 20036-4688, USA. [24] IBM, Somers, New York 10589, USA. [25] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [26] Universidade Federal do Rio de Janeiro, Rio de Janeiro, CEP 21941-901, Brazil. [27] Applied Biosystems, Foster City, California 94494, USA