

A Multiomic Approach to  
Paleogenetic Investigation of Ancient  
North American Bison

Joseph Boyle

Masters of Science Thesis

Presented to the Department of Biology

Lakehead University

Supervisor: Dr. Carney Matheson

# Table of Contents

<b>Declaration</b> .....	<b>III</b>
<b>Acknowledgements</b> .....	<b>IV</b>
<b>List of Figures</b> .....	<b>V</b>
<b>List of Tables</b> .....	<b>VI</b>
<b>Abstract</b> .....	<b>1</b>
<b>1. Introduction</b> .....	<b>2</b>
1.1 Objective .....	2
1.2 Theoretical Background.....	3
1.2.1 Structure and Nature of DNA .....	3
1.2.2 Ancient DNA .....	8
1.2.3 Structure of Protein .....	10
1.2.4 Proteomics.....	13
1.2.5 Multiomics .....	16
1.2.6 Multiple Simultaneous Extractions.....	17
1.2.7 Bison in North America .....	17
1.3 Methodological Background.....	19
1.3.1 Separation Chemistry.....	19
1.3.2 Quantitation.....	21
1.3.3 Nucleic Acid Methods .....	23
1.3.4 Proteomic Methods .....	32
1.3.5 Statistical Scoring .....	41
1.3.6 Inhibition and Damage.....	43
<b>2. Methods</b> .....	<b>49</b>
2.1 Sample Background and Preparation.....	49
2.1.1 Sample Background .....	49
2.1.2 Sample Preparation .....	51
2.1.3 Bone Milling .....	51
2.1.4 Bone Demineralization .....	52
2.1.5 Quantitation.....	53
2.2 Nucleic Acid Methods .....	53
2.2.1 Extraction .....	53
2.2.2 Purification.....	54
2.2.3 Amplification .....	55
2.2.4 Detection .....	57
2.2.5 Sequencing.....	57
2.2.6 Sequence Analysis .....	58
2.2.7 Phylogeny .....	58
2.3 Proteomic Methods .....	59
2.3.1 Alkylation, Denaturation, and Digestion .....	59
2.3.2 Purification.....	59
2.3.3 Liquid Chromatography - Tandem Mass Spectrometry .....	60
2.3.4 Protein Analysis .....	60

<b>3. Results</b> .....	<b>62</b>
3.1 Nucleic Acids .....	63
3.2 Phylogeny .....	67
3.3 Proteomics.....	68
3.4 Chloroform Separation.....	74
<b>4. Discussion</b> .....	<b>75</b>
4.1 Nucleotide Sequencing .....	75
4.2 Phylogeny .....	75
4.3 Protein Sequencing .....	78
4.4 Multiomics .....	79
4.5 Authenticity.....	80
4.6 Human Serum Albumin for Inhibition Relief.....	82
4.7 Chloroform Separation for Protein Purification .....	83
<b>5. Conclusion and Future Directions</b> .....	<b>84</b>
<b>6. Works Cited</b> .....	<b>86</b>
<b>Appendix</b> .....	<b>100</b>

# Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person, except where due acknowledgment has been made in the text.

Joseph Boyle

# Acknowledgements

I would like to acknowledge the help and support of several people, without whom this thesis would not have been possible.

Thank you to Dr. Chris Widga and the Illinois State museum for providing most of the samples and enthusiasm at the beginning of this project.

Thank you to Greg Kepka at the Lakehead University Instrumentation Lab for your assistance and patience with the mass spectrometry, and to Stephen Fratpietro and the staff at Paleo DNA Laboratory for assistance with sequencing.

Thank you to Dr. Matthew Collins for the invaluable insights on protein sequencing and impossibly prompt email replies.

To my committee members, Dr. David Law and Dr. Scott Hamilton, thank you for your patience, support, and advice with this thesis. Having both of you on my committee has helped put me at ease.

Thank you to Dr. Dongya Yang for serving as the external reviewer for this thesis. Your time and effort are greatly appreciated.

To my former fellow students Felicia Joseph and Karen Giffin, who have since graduated and moved on to greater things, thank you for your support and teachings at the beginning of this project. I'm glad you made it out.

To current fellow student Ashley Salamon, so much of this project would not have been possible without your incredibly hard work. Thank you.

Thank you to Ryan Lehto and Neil Esau, for both your advice and resources, and for your positive attitudes and encouragement throughout this process.

Thanks Jake Keller, for your timeless and enduring motivation.

Of course, to my family, without whom my continuing education would simply not be possible.

And especially to Dr. Carney Matheson. The extent of your knowledge is rivalled only by your enthusiasm. Your kindness and dedication to your students is incredibly appreciated. I could not have had a better supervisor.

# List of Figures

<b>Figure 1:</b> The double helix structure of DNA .....	5
<b>Figure 2:</b> The semi-conservative replication of DNA .....	6
<b>Figure 3:</b> The circular shape of mtDNA .....	7
<b>Figure 4:</b> The different levels of collagen molecule arrangement.....	12
<b>Figure 5:</b> Map of archaeological sites where samples were recovered.....	49
<b>Figure 6:</b> Gel image of amplified product from ISM11.....	63
<b>Figure 7:</b> Electropherogram of ISM11, amplicon 9.....	64
<b>Figure 8:</b> Graphical alignment of two amplicons from ISM11 .....	64
<b>Figure 9:</b> Graphical view of the top BLAST search result for ISM11 .....	65
<b>Figure 10:</b> Top ten alignments for a BLAST search of ISM11 .....	65
<b>Figure 11:</b> Alignment view of the top BLAST search match for ISM11 .....	66
<b>Figure 12:</b> Simplified phylogenetic tree incorporating ISM11 with the literature .....	67
<b>Figure 13:</b> GPM search results from ISM11 .....	68
<b>Figure 14:</b> Amplified product from ISM3 showing PCR inhibition.....	72
<b>Figure 15:</b> Amplified product from ISM3 after addition of HSA.....	73
<b>Figure 16:</b> Amplified product of a modern bison extract .....	101
<b>Figure 17:</b> Amino acid alignment for ISM11 .....	102
<b>Figure 18:</b> Graphical view of the top alignments from a BLAST search for ISM11.....	103
<b>Figure 19:</b> Top ten alignment identities from a BLAST search for ISM11.....	103
<b>Figure 20:</b> Alignment view of the top alignment from a BLAST search on ISM11 .....	103
<b>Figure 21:</b> Top 5 Global Proteome Machine search results from sample ISM2 .....	104
<b>Figure 22:</b> Global Proteome Machine search results from sample ISM5.....	104
<b>Figure 23:</b> Global Proteome Machine search results from sample ISM6.....	105
<b>Figure 24:</b> GPM search results for a negative control .....	105
<b>Figure 25:</b> Complete phylogenetic tree with samples from this study and the literature .....	110

# List of Tables

<b>Table 1:</b> Summary of results from all samples .....	<b>49</b>
<b>Table 2:</b> Reagent conditions for a typical PCR.....	<b>55</b>
<b>Table 3:</b> Reagent conditions for PCR with the addition of HSA.....	<b>56</b>
<b>Table 4:</b> Cycling conditions for all polymerase chain reactions.....	<b>56</b>
<b>Table 5:</b> Summary of results from all samples .....	<b>62</b>
<b>Table 6:</b> Summary of lane identities from gel electrophoresis of ISM 11 .....	<b>63</b>
<b>Table 7:</b> Summary of all peptide matches for ISM11 from a GPM search .....	<b>68</b>
<b>Table 8:</b> Summary of all peptide matches for ISM 2 from a GPM search .....	<b>70</b>
<b>Table 9:</b> Summary of lane identities for gel electrophoresis in Figure 14.....	<b>72</b>
<b>Table 10:</b> Summary of lane identities for gel electrophoresis in Figure 15.....	<b>73</b>
<b>Table 11:</b> Spectrophotometric readings for HSA purified with chloroform.....	<b>74</b>
<b>Table 12:</b> All primer sequences used in this study.....	<b>100</b>
<b>Table 13:</b> Summary of lane identities from gel electrophoresis in Figure 16.....	<b>101</b>

# Abstract

Twelve ancient bison bone samples from north-central North America were examined using genetic and proteomic sequencing to determine relationships to other bison populations. Mitochondrial DNA sequences suggest a genetic affinity that most closely matches populations from contemporaneous bison populations located in central North America. Proteomic sequencing by liquid chromatography tandem mass spectrometry could only resolve relationships to broad taxa and could not determine intra-specific relationships. Also, a novel multiple and simultaneous extraction protocol is presented to extract material suitable for both genetic and proteomic analysis from the same bone sample. It was also found that human serum albumin can be used as a replacement for bovine serum albumin as an effective additive to improve DNA amplification. In addition, chloroform alone can be used as an efficient organic solvent for the purification and separation of protein.



# 1. Introduction

## 1.1 Objective

Understanding the nature of ancient bison populations and speciation has been nothing if not dynamic over the last century. Since the earliest zooarchaeological studies, species classification of archaeological samples has been largely based on morphological qualities of recovered bison bones (e.g. size, shape, and robusticity). However, findings from more recent genetic studies have suggested that the relationships and lineages suggested by the phenotype are not reflected in the genotype, and previously identified samples may need reclassification. The genetics-focused projects have included large population sizes, but have left a gap in ancient bison populations from east of what is now the Great Plains. This study will attempt to fill this gap in genetic knowledge of North American bison through mitochondrial DNA (mtDNA) sequencing and subsequent phylogenetic analysis and attempt to further elucidate the nature of ancient bison lineages.

This question of bison species classification in zooarchaeology will also be used as a portal to develop new methodologies in ancient biomolecular science. This study will incorporate both Deoxyribonucleic Acid (DNA) and protein sequencing and will present a new streamlined simultaneous extraction of both biomolecules. Additionally, new efficiencies in genetic and proteomic methods will be explored and quantifiably evaluated, namely the use of human serum albumin as a PCR additives and the use of chloroform for protein purification.

# 1.2 Theoretical Background

## 1.2.1 Structure and Nature of DNA

### Chemical Structure

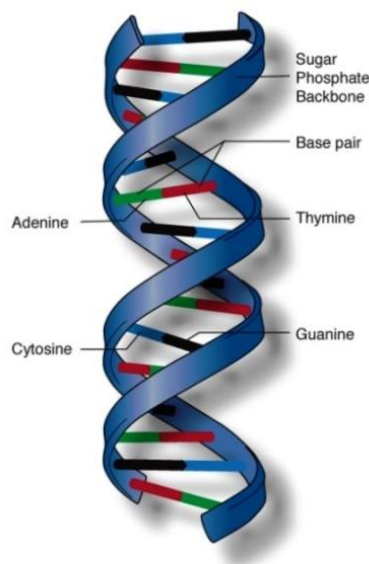
Deoxyribonucleic Acid is a critical molecule to life and has become a vital component in biological analysis. The whole story of DNA is complex. Despite the strong focus this molecule has received, there are still many questions about its nature and function; there are countless species and populations whose genetics are unstudied and the actual role of most of the genome is unknown.

Although the functional aspects of DNA are complicated, the structural form is relatively simple: four repeating nucleotide bases – adenine (A), thymine (T), guanine (G), and cytosine (C) - are arranged to code the information that DNA holds (Figure 1). These bases are attached to an alternating chain of pentose sugar and phosphate that acts as a backbone to hold the molecule together. The DNA molecule is naturally double stranded thanks to the complementary pairing of the nucleotide bases - adenine to thymine and guanine to cytosine - that creates two identical and opposite strands held together in antiparallel by hydrogen bonds between the base pairs (Hartl and Jones 2005).

The backbone of DNA contains two primary components: a deoxyribose sugar and a phosphate group. Deoxyribose is a single-ringed, five carbon sugar, with two oxygen atoms - one between the 1' and 4' carbon and the other as part of a hydroxyl group attached to the 3' carbon. This 3' hydroxyl group and the 5' carbon act as the binding sites for the second component of the backbone chain: the phosphate group. This group exists between each pentose sugar, as a

phosphodiester bond, connecting the 5' end of one molecule to the 3' end of the other, linked together to polymerize the backbone structure that supports DNA. The nucleotide bases are bound one each to the 1' carbon of the deoxyribose sugar, completing the structure of DNA (Hartl and Jones 2005).

Chemically, the nucleotide bases are made up of heterocyclic (carbon and nitrogen) rings, with additional amine, oxygen, or methyl functional groups. Two of these nucleotides - adenine and guanine - are double-ringed structures, called purines. Thymine and cytosine, on the other hand, are single-ringed molecules called pyrimidines. The mechanism of pairing, hydrogen bonding, also differs between the bases, with one of each purine and pyrimidine containing two sites for hydrogen bonding (A and T) while the other two (C and G) accommodate three hydrogen bonds (Hartl and Jones 2005). It is due to the specificity of these hydrogen bonds that the perfect complementarity of bases arises. The differing number of hydrogen bonds also carries differential bonding strengths and therefore bond lengths, which in turn give rise to the iconic



**Figure 1:** The basic double-helix structure of DNA (Wold 2016)

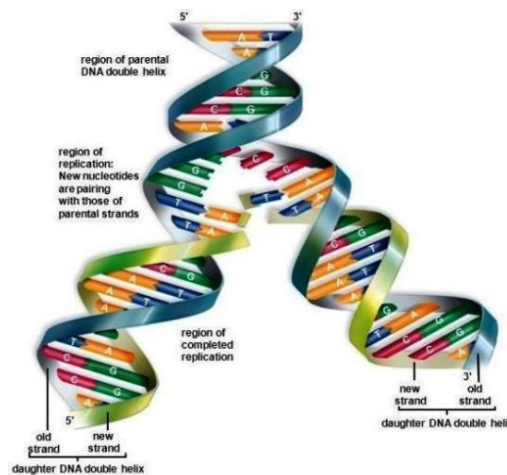
geometric structure of DNA.

The double helix structure of DNA was originally made famous by Watson and Crick in their 1953 manuscript, building on a number of publications. This understanding was critical to

the study of genetics. The double helix structure not only visually represents one of the most important biomolecules, but also provided the foundation for understanding semi-conservative replication - one of the key components of hereditary genetics. This was confirmed by the Meselson-Stahl experiment, empirically showing this fundamental function of biology (Meselson and Stahl 1958).

## **Replication**

One of the defining features of DNA is its perpetually and naturally self-replicating nature. This is done semi-conservatively (Figure 2), meaning that one half of the double helix is used to make a new double strand - from one parent strand, two daughter strands are formed (Hartl and Jones 2005). A cocktail of enzymes, proteins, and other chemicals mediates replication *in vivo*, each contributing to the fragile and common process of creating new strands of DNA. Critically, this replication is the fundamental means of genetic inheritance. Not only does this replication ensure that all new cells in an organism carry identical copies of the



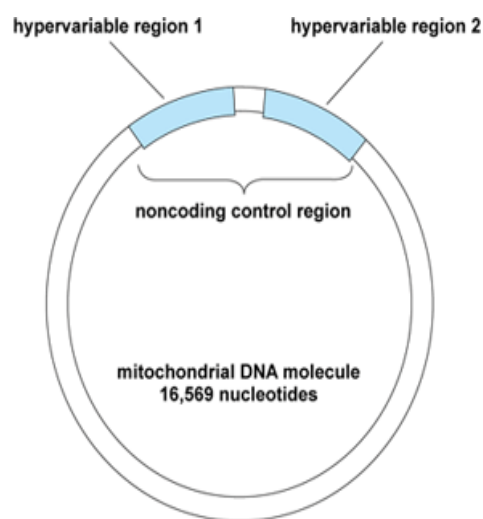
**Figure 2:** The semi-conservative replication of DNA (Wold 2016)

genome, it also leads to any offspring of that organism to inherit exactly half of a parent's nuclear genome. This forms the basis of evolutionary genomics: if all genes are inherited, it is

possible to trace this inheritance to determine relationships and to use changes in the genetic sequence to elucidate distance of individuals, populations, and species.

## Types

Within animal cells there are two types of DNA: nuclear DNA (nDNA) and mtDNA. Nuclear DNA is found in the nucleus of each cell, with cells containing two copies of each nDNA molecule. Mitochondrial DNA is contained, appropriately, in the mitochondria and can exist in many thousands of copies per cell (Montier et al. 2009). It is this high copy number that has led researchers to favour mtDNA when considering ancient samples, as there is a far higher



**Figure 3:** Mitochondrial DNA. Note the circular shape and the highlighted hypervariable regions 1 and 2 (Melton 2003)

likelihood of the mtDNA being preserved in analyzable quantities compared to nDNA.

Within the mitochondrial genome there are several regions which have become research hotspots for ancient DNA (aDNA) projects. One of the most common, and the one that will be used in this project, is called the hypervariable region (HV, control-loop, non-coding region,

displacement loop, or D-loop) (Figure 3). The hypervariable region itself is divided into two distinct sections, named HV1 and HV2 respectively. The HV1 and HV2 have been extensively studied in an aDNA context, as they are much more susceptible to accumulating mutations compared to other sections of the mitochondrial genome (Stoneking 2000). This elevated mutation rate allows a better resolution when studying evolution, as more data points can arise and one can better differentiate between species, populations, and individuals. These qualities make HV1 and HV2 an excellent region to study when the scope of an aDNA project must be constrained by limited resources.

## **1.2.2 Ancient DNA**

Analysis of aDNA form a distinct methodology within the larger field of genetics and genomics. Studies of aDNA began in earnest during the mid-1980's with the first demonstration of preservation of genetic material in ancient remains (Paabo 1984). This was followed in close order by the first sequencing of ancient genetic material, achieved with a 140-year-old quagga, an equine relative (Higuchi et al. 1984). Shortly after this, DNA from ancient mammoth (between 10,000 and 50,000 years old) was characterized (though not sequenced), showing for the first time the potential for ancient genetics to be used in evolutionary studies (Johnson et al. 1985). Studies of aDNA accelerated towards the end of the decade, with the exciting sequencing of mtDNA from a 7000-year-old human brain, by far the earliest sequences recovered at that time (Paabo et al. 1988).

Studies of the aDNA of megafauna further accelerated during the 1990s, with sequences obtained from archaeological cave bear (Hanni et al. 1994), mammoth (Hoss et al. 1994;

Hagelberg et al. 1994), mastodon (Yang et al. 1997) and ground sloth (Hoss et al. 1996; Poinar et al. 1998), among others.

Previous genetic studies of ancient North American bison are few, but have included a large number of individual bison and provided an extensive knowledge base on these megafauna. The first study was a publication of a small section of mtDNA from one *Bison priscus* individual older than 55.6 kya, alongside two osteocalcin protein sequences (Nielsen-Marsh et al. 2002). The first population-level study of ancient bison, published in 2004, explored the extinction of Beringian bison. Using sequence data from both modern and ancient samples, a drastic decrease in genetic variability was observed around 37 kya, which the authors attribute to a climate-driven extinction event (Shapiro et al. 2004). This study also built a large dataset from which further studies would be built upon and expanded.

Another notable ancient bison DNA project was undertaken by Wilson et al. (2008) and built on the earlier 2004 publication. Wilson et al. (2008) expanded the sample area to include archaeological bison from most of western and central North America, studying individuals from Alaska to New Mexico. The Wilson team chose to focus on a wider phylogeny for North American bison and mapped dispersal patterns of bison throughout time. Ultimately, Wilson et al. (2008) concluded that the species name *Bison occidentalis* is not applicable to any southern populations of bison, contrary to its standard treatment in the literature of the time.

Another genetic study was undertaken to examine the nature of the earliest appearance of *Bison latifrons* in North America (Froese et al. 2017). This was conducted on individuals from the Northern Yukon and a newly-discovered deposit in Colorado. This study both characterized the mitochondrial genetics of the earliest known bison in North America, and identified a pattern of rapid divergence following this species' arrival on the continent (Froese 2017).



The distribution of early North American bison was further examined in a recent publication, looking at fossil bison remains from where the Ice Free Corridor would have been located (Heintzman et al. 2016). Using a combination of radiocarbon dating and mtDNA, Heintzman et al. (2016) were able to show that the Ice Free Corridor did not exist during the late Pleistocene, at least not in a way that would have allowed for migration. This shows that North American bison regional populations were physically separated for at least several millennia, providing a mechanism for the genetic divergence noted in the publications above.

None of these publications have examined bison east of Alberta or Colorado, and so the large number of specimens recovered remain genetically uncharacterized. This study will attempt to fill in these gaps and explore the genetic nature of bison in these as-of-yet unstudied areas.

## **1.2.3 Structure of Protein**

If DNA is the instruction manual, proteins can be considered the product. The nucleic acid sequence, transcribed and carried by ribonucleic acid (RNA), is used to build a sequence of amino acids, the building blocks of polypeptide chains, which in turn combine to create proteins. All amino acids have the same generalized structure: an amine group (N-terminus), a carboxyl group (C-terminus), a central carbon atom, and a side chain. The side chain defines one amino acid from another, creating 20 unique molecules. Following the instructions given by DNA, these amino acids arrange and polymerize in any order and number (generally between 100 and 1000) to form peptide chains. These peptides in turn combine to form proteins that go on to perform many functions in the organism (Hartl and Jones 2005).

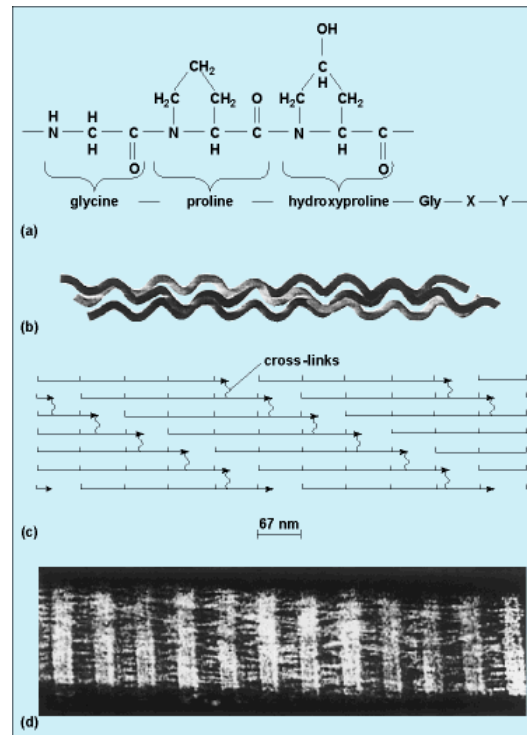
### **Orders of Structure**

The structure of proteins is divided into four levels of organization: primary through quaternary. Primary structure describes the raw ordering of single amino acids, forming the initial polypeptide chain (this sequence of amino acids forms the basis of the proteomic analysis in this project). These peptide chains then undergo geometric transformation due to intra-chain interactions (e.g. hydrogen bonding), forming secondary structures like alpha-helices or beta-pleated sheets in what is described as the secondary structure (Pauling et al. 1951). The tertiary structures of proteins are formed through additional folding within the polypeptide, arising from two dominant factors. Firstly, there are intra-molecular bonds, specifically sulphur-sulphur covalent bonds and further hydrogen bonding that contribute to this folding. Secondly, there are inter-molecular forces, particularly in relation to water molecules. Since polypeptides are such highly complex molecules, they possess different affinities to water along a single peptide chain. Folding of the polymer is therefore driven by this affinity, with a tendency to sequester the hydrophobic components in the interior of the structure (Kauzman 1959; Dill 1990). This, referred to as the hydrophobic principle (Pace et al. 2004), is important in the discussion on separation chemistry. Multiple polypeptides in tertiary structure can then combine to create complex, multi-chained proteins, described as quaternary structures (Bauman 2010). Quaternary structures are effectively the end of the line for protein geometry, though they may be further modified through combination with other molecules (e.g. glycoproteins and metalloproteins) (Bauman 2010).

### **Collagen**

Collagen is composed of three polypeptide chains woven together to form a triple helix (Figure 4) (Rich and Crick 1955; Cown et al. 1955). These polypeptide chains can be highly variable, with 46 distinct chains and giving rise to 29 different types of collagen, known so far,

that provide structure, strength, and flexibility to bones, teeth, blood vessels, skin, and many other aspects of vertebrate anatomy (Buehler 2006; Soderhall et al. 2007; Shoulders and Raines 2009). This variability makes collagen a highly versatile and therefore abundant protein, making up 1/3 of all protein in mammals and definitively the most abundant protein in the world



**Figure 4:** The different levels of collagen molecule arrangement, from amino acid chains (a), to triple-helix arrangements (b), to cross-linked fibrils (c), to a micrograph of a collagen fiber (d) (Brodsky 2014).

(Buehler 2006; Shoulders and Raines 2009). Though collagen does not offer the specificity that DNA analysis can, amino acid sequencing of the peptides that make up collagen can still offer a high level of information that can help discern species, populations, and to a limited extent, individuals (Collins et al. 2010).

Collagen is also an extremely well preserved protein, often surviving in the archaeological and paleontological record. This is due to its length, structure (triple helix),

abundance, and critically its context within the body; collagen is frequently found protected by the hard mineral structure of bones or underneath the durable enamel of teeth, facilitating its survival long past the current limits of aDNA. This survivability and the aforementioned variability make collagen a fantastic target for ancient proteomic research.

## 1.2.4 Proteomics

Broadly, proteomics is the process of deriving a global understanding of proteins and gene expression. Currently, work in the field of proteomics produces thousands of publications annually from a wide variety of disciplines, from pathology to pharmacology to archaeology. These studies examine all aspects of protein construction, function, and interactions (Graves and Haystead 2002).

The traditional methods employed in proteomics were first developed in the 1970s with 2-dimensional gel electrophoresis (2DGE) (O'Farrell 1975). The 2DGE uses the electrical properties to separate and characterize proteins in a polyacrylamide gel, a method that is still in use today, primarily for protein expression profiling (Graves and Haysted 2002). However, 2DGE is limited by its loading capacity and resolution, and so the field moved on to Edman sequencing. First developed in 1949, Edman sequencing caught on strongly with proteomics in the 1980s as an alternative to gel electrophoresis. Based on predictable protein degradation at the N-terminal of amino acids, Edman sequencing proved favourable as it offered better resolution than 2DGE, but had the drawbacks of significant risk of sample loss and again a relatively low specificity (Aebersold et al. 1987). Current proteomics came with the application of mass spectrometry (MS) to protein sequencing. Though MS instrumentation had been around for decades prior, it was not until the development of electrospray ionization (ESI) that allowed

interface with liquid phases and the ability to analyze non-volatile biomolecules (i.e. proteins). Mass spectral analysis of proteins relies on the fragmentation of peptide chains followed by collection of the mass-to-charge ratios of those fragments. Based on this data, amino acid sequences can then be determined, analyzed, and compared, forming the foundation of contemporary proteomics.

The increased sensitivity in protein sequencing brought on by the application of MS opened the door to analysis of ancient material through proteomics. The idea of archaeological protein sequencing was met with enthusiasm from the beginning, as it offered a potentially farther-reaching alternative to the comparatively fragile genetic analysis (Poinar and Stankiewicz 1999). In an archaeological context, amino acid sequences can provide the same type of information as genetic sequencing. Though the function of the protein is highly dependent on the peptide sequence, there still exists room for variation within the peptide sequence in the same protein across different individuals. It is these differences that can be exploited and studied in much the same way as genetics; models of relationship, function, and evolution can be constructed based on the similarities and differences between individuals both ancient and modern. At the time of writing, genetic studies have only reached back as far as a few hundred thousand years, whereas significant amounts of protein have been recovered and studied from animal sub-fossils as old as 1.5 million years (Wadsworth and Buckley 2014). Conveniently, many of these early studies have focused on bovids in relatively temperate European environments, giving strong support to the possibility of recovering useful proteins (particularly collagen, which was targeted in this study) from bison samples (Wadsworth and Buckley 2014).

Though ancient proteins can be studied in a number of ways, most popular methods for exploring ancient proteomics is Zooarchaeological Mass Spectrometry (ZooMS). A much more

recent method than aDNA, the applications of ZooMS are still growing. Not only has this method been used for collagen, as mentioned above, but it has also been employed to identify proteins from blood (such as albumin) and eggshells, among many others, and has obvious wide-ranging utility in archaeology and biology (Cappelini et al. 2011; Stewart et al. 2013). In archaeology, ZooMS is already being used for the identification of fragmented biological remains that may be too small to identify morphologically and may not have enough well-preserved DNA to facilitate genetic identification. Similarly, in biology, ZooMS provides a tool to analyze the relative uniqueness or relatedness of two individuals based on their proteins, giving a useful complement in comprehensive morphological, genetic, and proteomic studies and can be the only option in cases where remains are too small to obtain useful information using other approaches.

Many of the challenges that ancient genetics face are less of a concern with ancient proteomics. Proteins are far more abundant and many are much more durable than DNA, leading to longer survival of sample material. In addition, analysis of proteins does not require sensitive amplification and is not concerned with inhibiting contaminants the same way as DNA. Overall, this makes proteins far more resilient in an archaeological context than nucleic acids.

However, ancient proteomics is not without its drawbacks. As useful as they can be, amino acid sequences cannot provide the same level of resolution as nucleic acids. In theory, the whole genome can reveal the nature of every protein, but a protein can only speak to one small part of the genome. In addition, not all proteins can survive on an ancient time scale, and so ancient proteomic research is limited to only the most resilient or abundant proteins (e.g. collagen).

There has been only one paper published on ancient bison proteomics so far (Hill et al. 2015). This study looked at bone proteins from *B. latifrons* recovered from the same Colorado site as in the Froese et al. (2017). Using bone sampled from the inner skull of one individual, Hill et al. (2015) were able to identify several dozen proteins, primarily types of collagen, using bottom-up tandem MS.

## 1.2.5 Multiomics

Multiomics is a broad term used to describe any study that combines multiple global biomolecular methods (e.g. genomics, proteomics, transcriptomics, etc.) to study a single topic with the goal of reaching a better understanding than one method alone would allow. Multiomic approaches are quickly growing in popularity across the fields of medicine and biology, but the uptake to ancient samples has been slow, with only a small number of publications attempting to examine both genetics and proteins (or other molecules) simultaneously (e.g. Wadsworth et al. 2017).

The studies by Hill et al. (2015) and Froese et al. (2017), were able to obtain both proteomic and genetic information from the same species from the same location. Recent studies on ancient European bovids have compared protein content to DNA preservation in what is the closest to a true multiomic study of ancient material (Wadsworth et al. 2017). Both these studies provide encouraging evidence toward the feasibility of multiomic approaches focussed on ancient material, but stop short of applying a full-scale multiomic approach.

## 1.2.6 Multiple Simultaneous Extractions

Multiomics has great potential to increase the information gathered from individual samples, yet, it has not been widely applied in ancient studies. One possible way to increase the use of multiomic approaches is to improve extraction methods, namely with multiple simultaneous extractions of different biomolecules. This can improve efficiency and save valuable resources, which is of particular interest with ancient materials as sample material may already be restricted.

The current options for multiple extractions are limited. Currently, ready-made reagents are available, such as TRIzol (Life Technologies) or TRI Reagent (Molecular Research Centre) and associated pre-made kits (e.g. AllPrep (Qiagen)), but these can be time-consuming, expensive, and difficult to optimize (Pena-Lopez and Brugaolas 2014). A simpler method using basic reagents would give researchers more incentive to apply multiple simultaneous extractions to a sample, and would facilitate the application of multiomic approaches. This project will work to develop a new procedural workflow while examining the larger questions of ancient bison genetics and its application to zooarchaeology.

## 1.2.7 Bison in North America

This project focuses exclusively on the ancient genomics and proteomics of North American bison. Bison arrived in North America at least 130kya, with the first populations arriving in what is now Alaska and the Yukon from Asia across the Bering land bridge (Froese et al. 2017). These earliest bison belonged to the species *Bison bison*, and quickly spread



throughout the North American continent. At its peak, species of bison could be found literally from coast to coast, with specimens found as far south-east as Florida (Webb 1984).

There have been five recognized species of bison known to have existed in North America: *Bison priscus*, the earliest known bison to arrive, *B. latifrons*, an extreme-longhorn species, the intermediately-sized *B. occidentalis* and *Bison antiquus*, and finally modern *B. bison* (modern bison are furthered divided into subspecies *B. bison bison* and *B. bison athabascae*). Bison have been integrally linked with early human populations and their associated archaeological sites, bison have become a well-studied component of zooarchaeology. How bison have been classified in the archaeological record has been nothing if not complicated. Traditionally, bison species were identified based on morphology, particularly the size and shape of horn cores (most notably horn core breadth - tip to tip) and relative size and robusticity of long bones, tarsals/carpals, crania, etc. (Widga 2014). This has led to a fairly diverse classification of species, in some cases having several ancient species being identified in the same geographic area (Jenks 1937; Shay 1971; Widga 2014).

Species classification received an additional tool with the advent of modern genetic analysis. As noted above, these large-scale projects suggested a simpler model of North American bison relationships, broadly classifying bison into four distinct clades separated by geography (Shapiro et al. 2004; Wilson et al. 2008; Froese et al. 2017). These insights have brought about a reclassification of species identity in some cases (Widga 2014), with *B. occidentalis* designations dropped in favour of the *B. bison* identification.

## 1.3 Methodological Background

### 1.3.1 Separation Chemistry

To understand how nucleic acids and proteins are separated from each other one must first understand the chemical structure of both biomolecules. As discussed above, the backbone of DNA is composed of alternating pentose sugar (deoxyribose) and a phosphate group. While the ribose carries no charge, the phosphate group is highly negative due to the charged oxygen molecules (recall:  $\text{PO}_4^{3-}$ , though the charge will be 1- in the molecule as two oxygen atoms are bound to the deoxyribose sugar), and so the molecule as a whole carries an overall negative charge.

The electric principles of proteins are more complex than nucleic acids. Due to the high variability of the amino acid structures, polypeptides can have varying charges and can even vary along different stretches within a single molecule. This is in turn due to the high variability of the side chains in each amino acid. Some of these side chains - leucine, for example - lack any charge and are effectively non-polar. Others, such as lysine, do contain charged segments, and will act as a polar molecule. *In vivo*, proteins will coordinate the hydrophobic sidechains to the interior of the structure during tertiary and quaternary folding, rendering them soluble in the aqueous biological environments (Kauzmann 1959). However, when that protein is exposed to a non-polar solvent, the combined repulsive forces on the hydrophilic outer layer and attraction to the hydrophobic inner layer will effectively turn that protein inside out, denaturing that protein and rendering it soluble in non-polar solvents (Pace et al. 2004). This interaction is exploited particularly in organic phase separations (discussed below).

## **Organic Phase Separation**

One of the standard methods for nucleic acid purification has been the use of an organic phase separation. Since DNA is soluble in a polar solution (e.g. water) and protein soluble in a non-polar solution (e.g. chloroform), these macromolecules can be separated by exposing the sample to a mixture of both types of solvents and centrifuging to create a distinct phase separation, with the denser organic phase settling below the aqueous phase (Butler 2001). The separate phases can then be removed and analysed. Commonly, this is done using a phenol-chloroform mixture (in a 1:1 ratio, or 25:24:1 with iso-amylalcohol), as this has been shown to produce high yield and purity for nucleic acid purification (Hummel 2003) and variations of other organic solvents (e.g. chloroform:methanol) have been used in similar methods to purify proteins (Abramsky and London 1975; Vellaichamy et al. 2010).

## **Ethanol Precipitation**

Precipitation of nucleic acids by ethanol is one of the more widely used methods of nucleic acid purification, owing to its simplicity and effectiveness (Maniatis et al. 1982). The DNA molecule carries a slight charge, making it soluble in polar solvents like water. With the addition of 100% ethanol, the overall polarity of the solvent is decreased towards the point where DNA will be rendered insoluble (Mullay 1987). This on its own is not enough to precipitate nucleic acids, as water molecules will form a hydration shell around the DNA molecules, effectively neutralizing its polar attributes. To mitigate this, a strong ionic salt is added (such as sodium acetate) to remove the hydration shell and return the polar qualities to DNA (Churprina et al. 1991). Under these conditions, DNA can be readily separated and centrifuged to form a pellet. After removing the supernatant, the pellet can be dried, washed and re-suspended to produce a purified DNA containing solution.

## **Size Exclusion Chromatography**

In addition to the chemical separation methods discussed above, there are also methods based on physical properties. One of the more common methods is size exclusion chromatography (SEC). As the name suggests, SEC separates molecules in a mixture based on size. This is accomplished by using a column with hollow packing, the size of which can be tailored to trap particles of a certain size (Barth et al. 1994). For example, the column packing can be designed to prevent the passage of small molecules (e.g. metal ions) letting only larger molecules (e.g. DNA) pass through unobstructed. This is usually done under either centrifugal or vacuum force. Size exclusion chromatography has been shown to be effective at removing both metal ion and humic acid inhibitors from samples (Matheson et al. 2009; Matheson et al. 2010) and is now used extensively in both nucleic and amino acid sample preparation.

## **1.3.2 Quantitation**

Quantitation of nucleic acid and protein content is a critical step in the analysis process. The primary tool of quantitation used in this study is spectrophotometry. This technique has been used for decades for the determination of nucleic and amino acid content and purity, and remains a commonly used tool today. Quantitation of DNA relies on its preferential absorbance of light at the 260nm wavelength due to the double bonds in the purines and pyrimidines in the nucleotide bases (Ogur and Rosen 1950). The absorbance, determined by the spectrophotometer, is related to the nucleic acid content using Beer's Law. This same principle can be applied to proteins, which absorb distinctly at a wavelength of 280nm due primarily to the aromatic rings present in tyrosine and tryptophan (Stoscheck 1990). These spectrophotometric readings can not only determine quantity, but purity as well, by taking the ratio of the absorbance in a sample at 260nm

and 280nm (Desjardins et al. 2010). Generally, a 260/280 ratio of between 1.7 and 1.8 is regarded as “pure” double stranded DNA, 2.0 as pure RNA, and 0.5 as pure protein (Desjardins et al. 2010). Determination of purity is critical to the analysis process, as impurities will often result in failed amplification of DNA or erroneous proteomic results.

Though usually performed automatically by software, DNA and protein concentration is calculated using the Beer-Lambert equation:

$$A = \epsilon lc$$

where  $A$  is the measured absorbance,  $\epsilon$  is the molar extinction coefficient ( $0.020 (\mu\text{g}/\text{mL})^{-1} \text{cm}^{-1}$  for double stranded DNA at a 260nm wavelength, proteins are more variable, but  $\epsilon=10$  is assumed for unknown samples (Pace 1995)),  $l$  is the path length, and  $c$  is the concentration. This can be re-arranged to solve for concentration, thus:

$$C=A/\epsilon l$$

Purity, through the 260/280 ratio is calculated simply as:

$$\text{Purity}=A_{260}/A_{280}$$

It is important to note that all absorbance values can be corrected by subtracting readings from an initial blank read and by a read at 320nm to account for turbidity.

The advantage of spectrophotometric quantitation is that it is very rapid - usually only taking a few minutes, simple to perform - requiring only a few steps, and requires very little sample - only 2 $\mu$ L per microwell. Though commonly used, spectrophotometry is very susceptible to interference. Substances common to DNA and protein workflows, such as phenol, ethylenediamine-tetraacetic acid (EDTA), and detergents, among other things, can absorb in the 260-280nm range, and would lead to inaccurate readings and determinations of quantity and purity (Ahn et al. 1996).

## 1.3.3 Nucleic Acid Methods

### DNA Amplification

Following many years of gradual progress, a field-defining method for the rapid amplification of DNA was developed. This method, called the polymerase chain reaction (PCR) allowed for the exponential amplification of a small number of DNA molecules, mitigating the limitations of the small number of individual templates present in cells (Mullis et al. 1987). As the name would suggest, PCR relies on an enzyme-mediated chain reaction to double the number of copies each cycle. The products of amplification are then used as templates for the next cycle, creating an exponentially growing number of replicates of the original sample. This efficient amplification allows for a workable amount of DNA to be synthesized from a small sample volume, a key for working with heavily degraded samples as found in aDNA.

#### Thermal Cycling

A typical PCR consists of three main steps: denaturation, annealing, and extension. During denaturation, the reaction temperature is raised to around 95°C to break the hydrogen bonds of DNA and split the double strands into single-stranded templates. Typically, this temperature is held for about 1 minute. Temperature is then reduced to allow annealing of primers and enzymes to the template strand, usually for 30 seconds. This temperature can be variable, and is generally set to a few degrees lower than the melting point of the primers used in the reaction. Achieving the optimal annealing temperature is critical for the specificity of the reaction. If the annealing temperature is too low, primers may not bind to the target sample template exclusively and non-specific product may be produced. If this temperature is too high, primers may not bind to the template at all and no product will be produced. Once the primers

are bound to the target, the DNA polymerase will attach to the double stranded primed region of the template, and the reaction temperature is increased to around 72°C to initiate the enzymatic incorporation of single nucleotide bases, creating a new strand. This core process of denaturing, annealing, and extension is repeated anywhere between 15 and upwards of 40 times per reaction. The amount of DNA increases each cycle, effectively doubling the amount of DNA. Most PCR also includes an initial denaturing phase - usually one minute at the same temperature as the denaturation temperature to initiate the hot start for DNA polymerase. In addition, a final extension is often included, typically around 5 minutes at the cycled extension temperature to ensure all amplified fragments are adenylated (i.e. the addition of an adenine at the 3' end of each amplified product).

## **PCR Reagents**

### **Polymerase**

Amplification procedures can be highly variable, but all contain a set of core reagents that carry out the reaction. The active ingredient in PCR is a polymerase, the enzyme that facilitates the incorporation of nucleotides. In regular PCR, DNA polymerase is used and was originally commercially derived from *E. coli* bacteria for use in the earliest enzymatic extension experiments. While effective, this DNA polymerase was vulnerable to the high temperatures of the thermal cycling, and polymerase had to be re-added in each reaction, making easy enzymatic amplification very time-consuming. A significant leap forward in PCR protocol came with the derivation of DNA polymerase from the *Thermus aquaticus* bacteria, who live in extremely hot underwater heat vents (Saiki et al. 1988). The polymerase from *T. aquaticus* is naturally resilient to high temperatures as a result of their natural habitat in thermal pools, and so now the classic thermal cycling of PCR can be carried out with only an initial addition of the enzyme. There are

many types of polymerase available now, but this discovery of heat-resistant enzymes represented a major step in improving the efficiency of PCR, allowing for full automation of the thermal cycling process.

Relevant to this study was the development of “Hot-Start” polymerase. When added to the reaction, DNA polymerase has a tendency to start working immediately, even before thermal cycling begins. Since this is obviously not under ideal parameters, this early operation can generate non-specific product and inhibit amplification of the target sequence. To combat this action, “Hot Start” techniques were developed to delay the activation of the polymerase and improve results. Initially, this was accomplished by creating a physical barrier of wax between the polymerase and the rest of the reaction. Once the high temperatures of PCR were reached, the wax would melt and enzyme would mix with the reaction as intended. More sophisticated techniques are now used, predominantly the attachment of another enzyme to the polymerase molecule that prevents binding until a certain temperature is reached, denaturing this blocking enzyme and allowing the reaction to continue as normal. Other methods of mediation include antibody and ligand release systems.

### **Magnesium**

Related to the DNA polymerase activity is the addition of magnesium ions ( $Mg^{2+}$ ) to the reaction, commonly in the form of magnesium chloride ( $MgCl$ ) or magnesium sulphate ( $MgSO_4$ ) salts. Magnesium ions are a cofactor with DNA polymerase, catalyzing the synthesis of free nucleotides to the template chain (Erlich 1989). Not only is the presence of  $Mg^{2+}$  ions key to the outcome of PCR, but the concentration is vital due to its interaction with the template DNA, the primers, and their melting properties; too much magnesium and the reaction may carry out non-specifically and amplify non-targeted templates, too little magnesium and the reaction may not



amplify anything at all. For this reason, it is common to optimize a PCR protocol for magnesium concentration ahead of time. Another salt, potassium chloride (KCl), is also often added to similar effect, aiding in primer and template annealing (McPherson and Moller 2000).

### **Nucleotides**

For any replication to occur, the building blocks - nucleotide bases - must be available. These are added in the form of deoxynucleotide triphosphates (dNTPs) to each reaction. Again, hitting a "Goldilocks" zone of concentration of dNTPs is vital to the reaction; too high and the risk of misincorporation and error increases, too low and there may not be enough raw material to effectively amplify a workable amount of product. It is important to add each nucleotide (dATP, dCTP, dGTP, and dTTP) in equal concentrations to avoid affecting reaction fidelity (McPherson and Moller 2000).

### **Buffer**

All PCRs are heavily pH dependent, and so a buffer is included in each reaction. Tris-HCl is one of the more common buffers used, and will hold the pH between 6.8 and 8.3, though other buffers with different pH ranges may be used depending on the needs of the reaction (usually dependent on the type of DNA polymerase being used) (Moretti et al. 1998; Killelea et al. 2014). The acidity or alkalinity of the reaction is significant because, in addition to enzyme behaviour, at a lower pH nucleotides can cause the template to depurinate, reducing yield.

### **Primers**

For any replication to occur, the reaction requires a starting point. Primers initiate the start of PCR amplification. These are pre-made oligonucleotides designed to bind to a specific area along the template DNA, which generate a small section of double stranded DNA to facilitate DNA polymerase binding. Each amplification requires a set of at least two primers: one

forward and one reverse. These primer pairs are designed to bookend a section of DNA to be amplified (amplicons), the design of which can be tailored to the needs of the project. When working with highly degraded material (in ancient or forensic contexts, for example), designing primers to target small amplicons (i.e. under 100bp) is preferred, whereas DNA of better quality can be amplified in larger sections for efficiency (Hummel 2003). There are many other variables to account for when considering primer design. Generally, primers should be between 20-30bp in length, allowing enough variation for specific binding while mitigating risk for secondary structure formation that would arise for longer primers (McPherson and Moller 2000). Primers should also be designed so that the forward and reverse primers are not complementary to themselves, avoiding self-amplification.

### **Template**

Of course, PCR needs template DNA to amplify. This means an amount of as-pure-as-possible DNA, though the quantity can be theoretically as low as a single copy. It is critical to eliminate any source of contamination and inhibition from the sample DNA, either through sample processing, purification steps, or the inclusion of PCR additives.

### **Additives**

The components above provide the necessities for a modern PCR protocol to amplify DNA, but often times they alone are not enough. When troubleshooting, PCR additives can be considered. There are a wide range of options for additives, and should be approached based on the specific causes of problems with amplification; there is no one-size-fits all solution.

One additive relevant to this project is serum albumin. Simple additions of small amounts of this blood protein (i.e. 200-400ng/  $\mu$ L) have been shown to be particularly effective at relieving inhibition caused by humic acids and metal ions (Kreader 1996). The most common

source of this additive in biological research is bovine serum albumin (BSA). A by-product of the cattle industry, BSA is abundantly available and inexpensive, leading it to be the nearly-universal form of this additive used in PCR. However, since this project is dealing with bovids, a derivative of cow's blood can lead to obvious concerns about purity and possible contamination. In previous studies on ancient bison, alternative forms of serum albumin have been used - namely rabbit serum albumin (Shapiro et al. 2004).

There are many more additives that can be included in a reaction, and a more in-depth discussion on inhibition and damage is included below.

## **Sequencing**

Gathering information from DNA would be impossible without the ability to read the sample's nucleic acid sequence. There are several ways that this can be accomplished.

### **Chain-Termination Sequencing**

The earliest practical sequencing methods were built on the idea of chain-termination, first used in the "Plus/Minus" sequencing method (Sanger and Coulson 1975). These chain-termination methods were further developed by Fred Sanger and colleagues a few years later (Sanger et al. 1977), for which Sanger won his second Nobel prize. Chain-termination sequencing, or Sanger sequencing, relies on the incorporation of dideoxynucleotide bases (ddNTPs) during replication that will block further chain extension. A small number of ddNTPs are mixed with regular dNTPs. This results in the ddNTPs being incorporated randomly during extension, creating one chain of unique length for each nucleotide in the original template. Gel electrophoresis is used to resolve these strands, producing a series of bands, each differing by a single nucleotide. Originally, these were radioactively-labelled and were read following gel electrophoresis after exposure and development of X-ray film. The early radio-labelling methods

required as many as 8 reactions and lanes in electrophoresis to be able to read the nucleotide sequence, as the same label had to be used in each experiment on each chain-terminating nucleotide.

A significant step forward in Sanger sequencing came with the development of fluorescent dye-labelled ddNTPs (Smith et al. 1986). Using a different colour for each of the four nucleotide bases allowed chain-termination sequencing to be carried out in a single reaction and processed using a single lane in gel electrophoresis. In short order, this led to the use of capillary electrophoresis (CE) and laser induced fluorescence detection for full automation of the sequencing process (Cohen et al. 1988; Cohen et al. 1990; Ruiz-Martinez et al. 1993). Sanger sequencing with CE has become one of the most popular sequencing methods of the last couple of decades because of its ease of automation and applications to high-throughput scales, and is the sequencing method used in this study (França et al. 2002). Sanger sequencing is limited in length to reading sequences of up to around several hundred base pairs, but since aDNA is almost always found in segments smaller than this, it is a good match for an aDNA project.

### **Other Methods**

There are several other sequencing methods other than chain-termination. Another method developed at about the same time as the first Sanger sequences was the Maxam and Gilbert's chemical sequencing method (1977). This method relied on predictable chemical modification and subsequent chain cleavage of nucleic acid bases. Unlike Sanger sequencing, which is a sequence by synthesis method, Maxam-Gilbert sequencing is direct and does not require a replication step. This reduces the risk of sequencing errors from base misincorporation and eliminates a step in the analysis process. However, Maxam-Gilbert sequencing requires large

amounts of DNA, radioactive labels, and complex chemical reactions that are less amenable to automation, leading it to lose favour over chain-termination as Sanger sequencing improved.

Pyrosequencing was developed around 1990, gaining favour as it offered a real-time option for direct DNA sequencing. This method relies on the release of inorganic phosphates with enzymatic conversion producing photons upon dNTP incorporation during polymerase-mediated DNA replication (Nyman and Lundin 1985; Ronaghi et al. 1996). Pyrosequencing is still used today, though it can only sequence small stretches of DNA (no longer than 100bp) and is limited by non-linear responses to stretches of a single repeated nucleotide base (França et al. 2002).

In addition to these methods, there exists a whole host of proprietary techniques used by different biotechnology companies to sequence DNA, including single-molecule real-time sequencing, Illumina sequencing, and many others.

## **Nucleic Acid Bioinformatics**

To analyze DNA sequences, modern genomics relies heavily on bioinformatics. This project uses simple bioinformatics analysis, specifically three main tools for analysis: a sequence editor, a DNA sequence database search tool, and phylogeny software, used in serial to obtain meaning from the nucleic acid sequences.

### **BioEdit**

Once Sanger dye-terminated sequencing is completed, the raw data files must be examined, edited, and combined to create a usable DNA sequence. Sequence editing software allows the researcher to visually examine the chromatographic peaks produced during Sanger sequencing and make changes to the produced nucleic acid sequence as necessary. This visual inspection is crucial, as there are frequently errors in the produced sequence that can only be

identified and corrected manually. This data will often contain sections at the beginning and end of each sequence that contains errors because of primer and DNA polymerase behaviour and enzymatic drop off and may need to be discarded all together. Essentially, sequence editing requires a needed cleanup of the data prior to further analysis.

The program used in this study is BioEdit (version 7.0.5). First released in 1997, BioEdit is a free, open source program that allows for chromatogram viewing, sequence editing, and sequence alignment (Hall 1997). In addition to manual sequence editing, BioEdit also offers in-client interfacing with the ClustalW and MUSCLE algorithms to produce sequence alignments. These alignments are vital, as they can show the position of a sample sequence in a larger segment of DNA and reveal the location and nature of any polymorphisms in a sample against a reference sequence.

### **Basic Local Alignment Search Tool**

Once a sequence has been edited and aligned, further information can be gleaned using a search of that sequence against a database. This can elucidate the identity of the genetic sequence, the identity of the organism, and its relationship to other organisms. The tool used in this study is Basic Local Alignment Search Tool (BLAST), from the National Centre for Biotechnology Information (NCBI). Basic Local Alignment Search Tool is a web-based client that allows for the rapid searching of nucleic or amino acid sequences against NCBI's database of genes and proteins. First developed in 1990 (Altschul et al. 1990), BLAST's easy-to-use interface and database size has made it an extremely valuable tool in bioinformatics. Part of BLAST's user-friendly qualities comes from the program's heuristic approach to determining sequence matches; the simplifications and shortcuts used by BLAST help keep search times and server loads manageable (Madden 2013).

The outputs from BLAST searches contain several elements. The most essential of these are the alignments, showing the queried sequence aligned with sequences from the database. Each alignment also comes with multiple scores and evaluations of authenticity, discussed in detail below. These searches can be done for both nucleic acid and amino acid sequences, with both searches presenting similarly formatted results. Basic Local Alignment Search Tool also offers immediate generation of phylogenetic trees and estimations of relationships between the queried search and the database sequences, making the sequence analysis process streamlined and efficient.

### **Molecular Evolutionary Genetics Analysis**

Molecular Evolutionary Genetics Analysis (MEGA) is a free software commonly used to resolve phylogeny following nucleotide sequencing. First developed in the mid-1990s, MEGA has gone through many iterations and improvements and now offers a suitable collection of features for use by the modern geneticist (Tamura et al. 1994). The two primary tools used in this project are the distance calculator and phylogeny constructor. To resolve phylogeny between two or more sequences, the genetic distance between each one must be determined. There are many algorithms available, but they all share the same fundamental principle of computing relative genetic distance based on the number of nucleic acid substitutions. Once distance has been calculated for a dataset, a phylogenetic tree can be constructed to illustrate relationships.

## **1.3.4 Proteomic Methods**

Proteomic analysis was also included in this research. The workflow that this project follows to analyse proteins involves separation through liquid chromatography followed by tandem mass spectrometry (LC-MS/MS). Respectively, each of these three steps will allow the

separation of a mixture of peptides, mass spectral analysis of those peptides, and finally determination of the amino acid sequence of each peptide.

## **High-Performance Liquid Chromatography**

Fundamentally, chromatography is the collection of methods used to physically separate different substances (the mobile phase) as the mixture moves through a medium (the stationary phase). High Performance Liquid Chromatography (HPLC) is a method of separation for substances in a liquid. Developed during the 1960s, HPLC marked a massive leap forward over earlier forms of liquid chromatography, shortening the run time from hours to minutes and shrinking the apparatus size from columns as large as several stories to benchtop machines (Hamilton 1966; Snyder 1968; Karger 1997). The development of HPLC was important for the development of proteomics. Since large biomolecules cannot be readily volatilized, gas chromatography cannot be used to separate them prior to MS. Liquid chromatography setups at the time were impractical, there was no way to easily separate protein mixtures prior to MS. HPLC allowed for this separation to be easily carried out, and ushered in a whole new field of proteomic discovery.

Analysis by HPLC relies on a sample (the analyte) dissolved in a liquid solvent. This can be an aqueous solution, a non-polar solvent, or commonly both used in combination to make a gradient. As a whole, the analyte and solvent are referred to as the mobile phase. Using high pressure from a pump, the mobile phase is forced through a column containing the stationary phase. The composition of the stationary phase is made from is highly variable and can be tailored to exploit the properties of the target analyte and address the needs of the project. No matter the makeup, all columns rely on the differential adsorption of the mobile phase to produce separation; substances with a higher affinity for the stationary phase will adsorb more strongly



and take longer to elute out than those with lower affinity, and so the different fractions of the mobile phase will exit the column at discrete times. After exiting the column, the substances in the mobile phase can be analyzed in a number of ways, such as spectrophotometry or MS (or a combination of many methods).

The columns used for the stationary phase in HPLC come in a very wide range of compositions, with the packing being tailored to the needs of the project. The column used in this project is a C18 Reversed Phase column. Reversed Phase Chromatography (RPC or RP-HPLC) describes any chromatography that has a hydrophobic stationary phase. This style of column packing has become so ubiquitous that the “Reversed Phase” nomenclature is often just dropped and “normal phase” is more commonly adopted as the modifier to signify a hydrophilic column. A reversed phase (recall: hydrophobic stationary phase) is usually paired with an aqueous (polar) mobile phase (or a gradient of polar and non-polar solvents), meaning that non-polar compounds in the mobile phase will have a higher affinity for the stationary phase. In this case, the most-polar substances will elute first with the most non-polar substances eluting later.

## **Mass Spectrometry**

Fundamentally, MS measures the mass-to-charge ratio of ions. These ratios can then be used to infer the composition of the original analyte. Mass spectrometers have been used since 1919 (Aston 1919) and consist of three primary components: an ion source, a mass analyzer, and a detector.

### **Ion Source**

Before being analyzed by MS, the analyte must first be converted into an ionic gas. This process is performed by the mass spectrometer’s ion source. There are many ways this can be performed, but two specifically are used in proteomics: Electrospray Ionization (ESI) and

Matrix-Assisted Laser Desorption Ionization (MALDI) (Han et al. 2008). The method used in this project is ESI. First developed in 1989 (Fenn et al.), ESI creates ions by passing the analyte, dissolved in a solvent, through a narrow nozzle under high voltage, creating a fine mist. The solvent then rapidly evaporates leaving only a stream of ionized analyte. This stream is directed into the mass analyzer. The development of ESI was key to the use of MS in proteomics, since proteins cannot be made volatile, it allowed for the use of a liquid phase prior to analysis.

The other sourcing method that is used in proteomics is MALDI, which co-crystallizes the protein with an organic solvent to create a solid ion source (Karas and Hillenkamp 1988). This solid is bombarded with a laser to explosively desorb and ionize the protein, creating an ion cloud that is fed into the mass analyzer.

### **Mass Analyzer**

Like the ion source, there are several different mass analysis techniques used in MS, including Quadrupole (Q), Ion Trap (these can be either Quadrupole (QIT) or Linear (LIT)), Time-of-Flight (TOF), and Fourier-Transform Ion Cyclotron Resonance (FT-ICR) (Han et al. 2008). Although each method has its own specifications, they all operate under the same principle of using a variable electric field to separate the ions from the ion source by their mass-to-charge ratio.

The instrument used in this project was an ion trap analyzer. Ion traps work by using electrodes aligned on multiple axis to create alternating magnetic fields that contain ions suspended in a vacuum (Jonscher et al. 1997). These magnetic fields can be adjusted to contain only ions of a certain mass-to-charge ratio, which can then be sent to the detector. The adjustment of the magnetic fields can be run across a wide range of charges, allowing for scanning of a wide range of particles.

## **Detector**

The role of the detector in MS is to determine the presence and number of ions at a specific  $m/z$  ratio following sorting by the mass analyzer. Though often less specialized than the ion source or mass analyzer, detectors come in a variety of forms, including induced-current detection, collision detection, and fluorescent detection. One of the most popular types of detectors are electron amplifiers. When a charged particle collides with the electron multiplier, the multiplier will release several secondary electrons. These secondary electrons are then measured and recorded, allowing for the inference of the abundance of the primary charged particle. The instrument used in this project uses an electron multiplier as its detector (Koster 2015).

## **Tandem Mass Spectrometry**

Key to proteomics is Tandem Mass Spectrometry (MS/MS). As the name implies, MS/MS uses two rounds of mass spectrometry arranged in series to achieve a higher resolution of information. First, one round of MS is used to detect individual peptides. Once peptides have been sorted, they then undergo fragmentation to further break them down into amino acids. Fragmentation can be achieved in a number of ways. One method is by Collision Induced Dissociation (CID), where ions from the first round of MS (conducted in a vacuum) are passed through a chamber containing an inert gas. Collisions between the analyte and the gas cause violent fragmentation before being sent to the second round of MS.

Another method is by Electron Transfer Dissociation (ETD). Based on the earlier method of Electron Capture Dissociation (ECD) (Zubarev et al. 1998), electrons are transferred on to the charged ions following the first round of MS (Syka et al. 2004). These charged particles are unstable and will fragment. The products of this fragmentation are directed into the second round

of MS. This fragmentation is often combined with another method, Proton Transfer Reduction (PTR) to reduce the charge on the fragmented peptide ions to further prepare them for MS (Loo et al. 1994). These fragmentation techniques can be used individually or in combination. This project uses CID as its method of fragmentation before the second round of MS.

These fragmentation methods break the peptides at variable lengths along their backbone, and so the second round of MS can detect the  $m/z$  ratios of this mixture of amino acids from the same peptide, and can then be used to reconstruct the original peptide sequence.

### **Mass Spectrometry in Proteomics**

Though MS has been in use for nearly 100 years, it has only been used in proteomics for the past three decades (since the development of ESI as an ion source) and has since become the dominant tool in the field (Han et al. 2008). Previously, other methods such as microarrays and Two-Dimensional Gel Electrophoresis (2DE) had been used, but they do not provide nearly the same resolution as the current MS methods.

There are two general approaches to MS proteomics that both work towards obtaining amino acid sequences: top-down and bottom-up. Top-down proteomics start with whole proteins without any prior digestion (Han et al. 2006; McLafferty et al. 2007). In the top-down approach, intact proteins are directly analysed with MS, relying entirely on the fragmentation within the mass spectrometer to provide resolution. Top-down proteomic studies are advantageous because they ensure the entire protein is being studied; without prior fragmentation, there is no chance for loss of information. The drawback here is that they may sacrifice resolution, as the fragmentation cannot be controlled or fine-tuned in the same way as bottom-up proteomics.

Bottom-up proteomics start with a digestion of the protein prior to mass spectral analysis. This digestion reduces proteins to their peptide components, which are used in MS analysis.

Tandem MS can be used to further fragment these peptides to read the individual amino acids. There are two principle workflows with bottom-up proteomics: a “sort-then-break” and a “break-then-sort” method. The “sort-then-break” method is where a sample is separated into its constituent proteins prior to digestion. The “break-then-sort” method - also called Shotgun proteomics – is where no initial fractioning is done and all available proteins are digested and analyzed with the mass spectrometer, sorting of the peptides and their sequence is done post-analysis via bioinformatics (McDonald et al. 2003; Han et al. 2008). In comparison, bottom-up proteomics has a higher tendency to lose information during the digestion than the top-down approach, but the bottom-up approach can provide better resolution (Han et al. 2008). It is this second approach, shotgun proteomics, that is conducted in this study.

## **Proteomic Reagents**

There are several reagents used here to prepare our samples for MS. Generally, this process functions to isolate, prepare, and optimize collagen for LC-MS/MS analysis.

### **Ethylenediamine-tetraacetid Acid**

The EDTA is used to demineralize the sample, as it will chelate the calcium ions in the bone. The removal of calcium ions is significant to both the protein workflow, cleaning up the protein samples, and the DNA workflow, removing the divalent cations that will inhibit PCR. EDTA functions by binding to divalent metal cations (critically,  $\text{Ca}^{2+}$ ), removing them from the bone matrix. Since calcium makes up a significant portion of animal bone, EDTA does much of the heavy lifting during sample preparation in both proteomic and genetic analysis, and is widely used in both fields (Tuross et al. 1988; Baron et al. 1996).

### **Iodoacetamide**

Iodoacetamide (IAA) is used to alkylate cysteine in our samples. Alkylation of cysteine is used to reduce the size of the database that mass spectra need to be searched against (Sechi and Chait 1998). By forming a covalent bond with the charged sulphur group on cysteine, IAA effectively “tags” cysteine, adding a reliable modification with which to constrain database search parameters and increasing efficiency of the data analysis process.

### **Ammonium Bicarbonate**

Ammonium bicarbonate ((NH<sub>4</sub>)HCO<sub>3</sub>) (ABC) is used here as a buffering agent during the trypsin digestion of protein. The ABC is widely used in protein preparation for MS analysis as its buffering capacity (around pH 8) closely matches the optimal range for trypsin activity (Rosenfeld et al. 1992). The ABC is a volatile salt, making it acceptable for processing in MS (though salts are removed from our samples prior to MS in this case).

### **Dithiothreitol**

Dithiothreitol (DTT) is a thiol-containing molecule with wide applications in biochemistry. Specific to proteomics, DTT is commonly used as a reducing agent, acting in similar fashion to IAA, binding covalently to sulphur to prevent re-formation of sulphide bridges (Getz et al. 1999). This prevents reformation of complex folding and maintains protein suitability for proteomic analysis.

### **Trifluoroacetic Acid**

Two steps in this procedure requires the use of a strong acid. Trifluoroacetic acid (TFA) is used as it is widely available and less dangerous than other similar acids. First, TFA is used to quench the trypsin digestion by inactivating the enzyme, which is achieved by decreasing the pH to around 3 (Kunitz and Northrop 1934). The TFA is also used during the purification process, acting as a strong buffering acid during protein binding to the C18 purification tips.

## **Formic Acid**

Formic acid (FA) is another useful acid in the proteomic preparation process and is widely used in biochemistry involving HPLC and MS. Addition of even small amounts of FA prior to analysis provides better peak resolution and improves specificity. The action of FA is as an ion-pairing agent, aiding the retention of small non-polar molecules (such as the modified beta-Asp residue) that may resolve poorly in reversed-phase HPLC (Gustavsson et al. 2001; Krohkin et al. 2006).

## **Trypsin**

One of the most critical components to the protein preparation process is the enzyme trypsin. *In vivo*, trypsin is found in the small intestine of animals to hydrolyze proteins during digestion (Rawlings and Barnett 1994). In proteomics, trypsin is used to digest proteins, predictably cleaving them to facilitate mass spectrometry, one of the fundamental components of shotgun proteomics. The advantage to trypsin is that it cleaves in highly specific locations, always acting on the C-terminus of lysine and arginine (Brown and Wold 1973). This predictability translates well into database searching for protein and peptide identity, and is now one of the most widely-used tools for protein digestion in proteomics.

## **Acetonitrile**

To ensure that both polar and non-polar molecules successfully elute through HPLC, it is standard to use a mixture of both a polar (i.e. water) and non-polar solvent as the mobile phase. Acetonitrile (ACN) has emerged as one of the preferred non-polar solvents for this purpose, as its simple chemical structure and low UV absorbance minimizes the interference with spectrophotometric analysis following HPLC and its volatility makes it amenable to in-line MS

(Wilson et al. 1981). The solvent ACN is also miscible with water allowing a wide gradient range to be generated.

## **Bioinformatics and Proteomics**

After mass spectra have been produced, they must be analyzed by one of the many available software programs to obtain protein and peptide identity. There are two main categories of programs that are used in proteomic bioinformatics. Some programs will construct protein sequences *de novo*, without any interpretation of protein identity. *De novo* sequencing is best used when working with single proteins, and can be limited when considering complex protein mixtures. Other programmes rely on comparing the mass spectra of the analyte with a database of known sample, then scoring the comparisons across multiple criteria to determine the most likely match (Yates 2004). These programs are much more efficient at determining protein identities in complex mixtures, though they are limited in that they rely on existing databases and it can be challenging, to identify novel proteins, or when the target species is not in the database.

The primary bioinformatic tool used in this project is the Global Proteome Machine (GPM). The GPM is a free open-source server-side tool for protein identification. This program is built on the X! TANDEM architecture and its offspring X! Proteotypic Protein Profiler (X!P3)(Craig and Beavis 2004). These algorithms are used in concert with the Global Proteome Machine Database (GPMDB) to generate peptide identities.

### **1.3.5 Statistical Scoring**

Whenever mathematically generated results are considered, statistical significance is imperative. Both the GPM and BLAST use robust statistical scoring methods that incorporate multiple measurements of confidence when assigning identities.



## **E-Value**

Central to both the GPM, BLAST, and many other bioinformatic tools is the e-value. When comparing against a large database, as the GPM and BLAST do, it is useful to determine the probability of a match happening by chance (i.e. what is the “expectation” of a random match, hence e-value) (Collins et al. 1988). Though based on the same principle, different search tools may employ specialized algorithms for calculating e-values based on the properties and needs of that program (Ericsson and Fenyo 2004). Both the GPM and BLAST report e-values with the search results, reported as a base log value.

## **False Positive Rate**

In statistics, the p-value has become the ubiquitous catch-all for statistical significance of data. This value, however, can only apply to single instances of data, and so a different metric needs to be used (Aggarwal and Yaday 2016). In global searches (such as in the GPM) a p-value cannot be calculated, and so a new value must be used to evaluate confidence in the results. To accomplish this, the GPM reports the false positive rate (FPR) with search results (Ericsson and Fenyo 2004). Essentially, the FPR is the chance that any match within the entire set of results has been matched incorrectly (a false positive). Though not definitive on its own, when combined with an e-value, these two measurements provide statistical robustness for results from the GPM.

## **BLAST Scores**

To corroborate statistical significance alongside the e-value, BLAST also reports a score for each sequence alignment. Expressed as a logarithmic exponent, an alignment score essentially represents the size of a database you would have to search to find an alignment that matches as well or better than the initial search result (Karlín and Altschul 1990). Unlike the e-

value and FPR, a higher score represents higher confidence, and can be used in combination with BLAST's e-value to provide statistical significance to the results.

## 1.3.6 Inhibition and Damage

Inhibition and DNA damage are primary concerns when working with aDNA. Inhibition here is a broad term that can describe anything that prevents the successful amplification and with PCR being such a sensitive procedure, there are numerous sources for inhibition. These sources can be derived from environmental conditions (such as humic acids in soil) (Abbaszadegan et al. 1993), from procedural artefacts (such as EDTA from extraction or ethanol left over from a purification step) (Wilson 1997), and from chemical modifications of the DNA template itself (such as cross-linking with protein) (Wilkins and Smart 1996).

DNA damage presents similar challenges to inhibition, though are borne at a different stage. Following the deposition of the organism, there are numerous processes that act to damage and degrade the template strand, giving rise to failed PCR, sequencing, and to errors even when amplification and sequencing are successful.

Environments typical of the northern North American mid-continent (i.e. densely forested with harsh winters, continental climates reflecting repeated freeze/thaw and wet/dry cycles, podzolic soils, shallow deposition and intense microbial activity, among other things) provide unique and significant challenges when considering DNA. The challenges derive primarily from the environmental chemistry and climate of this area. Chemically, boreal forests are some of the most difficult depositional environments to recover amplifiable DNA as they contain some of the largest variety of substances that can damage DNA and inhibit PCR. The large amplitude of temperatures in this environment only further contributes to this problem. Some of the most

common sources of inhibition and damage specific to the environment from the archaeological context of the samples in this project are discussed below.

### **Humic Acids**

One of the substances that typifies northern soil chemistry are humic acids. Not a single chemical entity itself, humic acids are the collection of substances containing carboxyl and phenolate active groups (McCarthy 2001). These form as a product of biological degradation and are virtually ubiquitous in organic soil. Humic acids can inhibit amplification even at very low concentrations (as low as 10ng per sample) (Tsai and Olsen 1992).

The actual mechanism of humic acid as an inhibitor is currently poorly understood due to the complicated nature of humic complexes (Matheson et al. 2010). One mechanism has been proposed that this action is caused by the binding of humic acids to the template DNA and preventing enzyme activity (Abbaszadegan 1993). Nevertheless, inhibition by humic acids can be resolved, with the current prescription being the use of size-exclusion chromatography and gel filtration to physically separate the humic acids from the sample (Matheson et al. 2010). In addition to the inhibition that poses issues during amplification, humic acids will also interact with DNA to induce oxidative damage (discussed further below), creating further issues during both amplification and sequencing (Cheng et al. 2003).

In the context of proteins, humic acids may actually have preservative qualities, and have been shown to significantly strengthen collagen (Riede et al. 1992). Though not fully studied in an archaeological time scale, this may indicate that humic-rich environments may be amenable to collagen preservation, but not DNA.

## **Tannins**

Like humic acids, tannins are a collective of different naturally occurring molecules that can contribute to failed DNA amplification from archaeological samples. Tannins are polyphenolic compounds that can be derived from both organic (e.g. vegetables) and inorganic (e.g. minerals) sources and have been used by humans to process leather for millennia, from which they derive their name (Annick et al. 2006). Tannins are found abundantly in northern environments, primarily in trees and entering the soil through decomposition. In a PCR, tannins will act as chelators, chelating magnesium ions, vital cofactors to DNA polymerase activity and prevent enzyme-mediated replication of the template (Opel et al. 2010).

Conversely, the preservative qualities of tannins may benefit proteomic analysis. In the leather-making process, tannins will bind to collagen to increase stability in a post-biological context. This action has been correlated with the exceptional physical preservation of human remains recovered from tannin-rich bogs (so-called “bog bodies”) (Stankiewicz et al. 1997). What is detrimental to DNA analysis may be beneficial to proteomics.

## **Water Damage**

One of the most common forms of damage to DNA is hydrolytic damage. Caused by interactions with water molecules, hydrolytic damage is present to some degree in almost every archaeological sample but is particularly abundant in consistently wet environments.

The actual effects of hydrolytic damage can manifest in several ways. Water can cleave the DNA backbone, fragmenting the strand. This cleavage occurs at the phosphate group that joins the ribose sugars in the nucleic acid backbone, and will result in fragmentation of the DNA between 100 and 500bp long (Paabo 1989). Water can also cause cleavage between a nucleotide base and the ribose sugar in the backbone (known as depurination and depyrimidation). This

results in abasic sites that can prevent primer binding or enzymatic extension during PCR (Lindahl 1993). Finally, water can also attack the nucleotide base itself (e.g. deamidation), resulting in a modified base within the template and hindering replication during PCR (Shapiro 1981). This process is damaging in any context, but is catalyzed in acidic environments (such as where the samples in this project have been recovered), adding further complications to obtaining workable DNA.

Hydrolysis can also damage proteins. Hydrolysis of proteins is naturally occurring and is a vital digestive process in organisms. *In vivo*, this process is catalyzed by enzymes and occurs rapidly. In post-mortem samples, this process is significantly slower, but can still have an impact when considering an archaeological time scale. Fragmented proteins are less of a concern as fragmented DNA, since their size, abundance, and survivability relative to DNA helps to mitigate the drawbacks of fragmentation, but it can still pose problems with sample preservation.

Water can also physically damage template DNA, through the formation of ice crystals during freeze/thaw cycles. These ice crystals can physically shear the DNA, causing an increase in fragmentation and a reduction in yield even after only an hour of being frozen (Grecz et al 1980; Ross et al. 1990). These freeze/thaw cycles can be extremely damaging to DNA strands over time, causing many double stranded breaks and limiting the amount of workable DNA in a sample. This process is extremely relevant to archaeological samples from northern environments, as biological remains may be subjected to several freeze/thaw cycles each year, causing extremely degraded sample DNA.

### **Oxidative Damage**

Oxidative damage to DNA is nearly impossible to avoid and can affect any sample to some degree. In an archaeological context, DNA oxidation is caused primarily by hydroxyl and

peroxide radicals and can attack DNA at multiple locations (Paabo et al. 2014). Oxidation can occur at the double bonds found in purines and pyrimidines, creating either modified bases or abasic sites (Lindahl 1993). This in turn can cause failed amplification from either failed primer binding, termination of enzymatic extension, or errors in sequencing from misincorporated nucleotide bases (Mitchell et al. 2005). Oxidative damage can also affect the deoxyribose sugar of DNA, causing ring opening and cleaving the strands, producing similar effects to hydrolytic and freeze/thaw damage on DNA (Lindahl 1993).

### **Metal Ions**

The presence of metal ions from the soil can also negatively impact DNA in a number of ways. Unfortunately, metal ions (iron and copper, for example) make up a significant portion of northern soil chemistry, and present challenges when working with samples taken from archaeological contexts. Metal ions can catalyze the generation of reactive oxygen species that will oxidatively damage DNA, impeding amplification (Henle and Linn 1997). Metal ions are able to form DNA/metal crosslinks that can inhibit enzymatic extension. Metal ions can also cause significant problems during the amplification process itself, acting as competitive inhibitors to magnesium - an essential cofactor for the DNA polymerase - and preventing enzymatic mediated replication. These reactive oxygen species can also attack proteins, degrading them through oxidative damage, causing fragmentation and making them further susceptible to other forms of damage (Wolff and Dean 1986).

### **pH**

The soils of the boreal forests of North America are characteristically low in pH due to the high presence of decomposing organic matter, with soils easily reaching a pH of 4.5 or lower (Bauhusa et al. 1998). This presents a number of issues. Of greatest concern is the effect of

acidity on bone preservation. The mineral component of animal bone, hydroxyapatite, makes up the majority of the bone's mass (~70%) and is usually insoluble in water. However, under even mildly acidic conditions the hydroxyapatite will begin dissolving (Nielsen-Marsh et al. 2007). Over time, this will wear away at the protective mineral structure of bone, leaving only the organic components (i.e. DNA and proteins) susceptible to further damage.

## 2. Methods

### 2.1 Sample Background and Preparation

#### 2.1.1 Sample Background

This study examined samples from six different archaeological sites from the northern midwest United States and Northwestern Ontario. Most samples were provided by mail from the Illinois State Museum. An additional sample from the Kenora bison was already available through the collection at Lakehead University, previously provided by the Manitoba Museum. A sample of bone from a modern bison also in Lakehead's collection was taken as the modern control. A summary of samples, their sites, and their ages can be found in Figure 5 and Table 1.

Each of the American sites includes evidence of human activity alongside the bison remains. Some sites, such as Itasca, feature extensive collections of tools and debitage associated



**Figure 5:** Map of central North America showing locations of sites from which archaeological bison were originally obtained for this study.



with the archaeological bison and other animals, while others, such as the Interstate Park site feature a small number of hammered copper tools. In any case, there exists definitive evidence for contemporaneous human activity at each of these locales.

**Table 1:** Summary of bison sample information and origin.

Sample Name	Locality	Site ID	Type of Bone	Illinois State Museum #	Age ( <sup>14</sup> C YPB)	Key Publication
ISM1	Hill, IA	13ML62	Long Bone Shaft	ISM1	7250±200	Frankforter and Agogino 1959
ISM2	Hill, IA		Long Bone Shaft	ISM2		
ISM3	Hill, IA		Long Bone Shaft	ISM3		
ISM4	Smilden-Rostberg, ND	32GR123	Long Bone Shaft	ISM4	5960±230	Larson and Penny 1991 (unpublished)
ISM5	Interstate Park, WI	47PK36	Long Bone Shaft	ISM5	Archaic	Cooper 1937
ISM6	Interstate Park, WI		Long Bone Shaft	ISM6		
ISM7	Interstate Park, WI		Long Bone Shaft	ISM7		
ISM8	Itasca, MN	21CE1	Petrosal	ISM8	7880±90	Jenks 1937
ISM9	Itasca, MN		Petrosal	ISM9		
ISM10	Simonsen, IA	13CK61	Petrosal	ISM10	8430±520	Frankforter and Agogino 1960
ISM11	Simonsen, IA		Petrosal	ISM11		
Kenora	Kenora, ON		Phalange	N/A	4850±60	McAndrews 1982

Environmentally, many of these sites have experienced dynamic local climates since the time of deposition of these bison. For example, pollen analysis has shown that the horizon from where the Itasca bison was recovered, was typical of an open grassland environment at the time of deposition (Shay 1971). At that location now is a bog, with dynamic water table levels and a

consistently wet environment more typical of contemporary boreal environments. Many of the other sites examined, particularly the Kenora site and the Simonsen site, have the same dynamic wet environment.

### **2.1.2 Sample Preparation**

Samples were first cut, using a band saw, into smaller pieces that would fit into the vials for milling (approximately 2cm<sup>3</sup>). The samples were then cleaned using a Dremel tool. A Dremel pencil-tip bit covered in either grinding stone or diamond powder was used to remove the visible layer (approximately 1mm) from all external surfaces of the bone pieces. Each Dremel tip was used only once, then discarded and replaced. Prior to use, the tips were cleaned with 2% hypochlorite bleach and Kimwipes (Kimberley-Clark). All work areas and bench surfaces were cleaned with 2% bleach at the start of the work day and between the cleaning of each sample. Two pairs of latex gloves were worn at one time throughout the experiment. The outer glove was changed after each sample and both gloves were changed if the gloves were torn. Hands were also washed if a glove broke. Bone samples were stored in separate sealed plastic bags. Finished samples were also wrapped in a Kimwipe to preserve cleanliness.

### **2.1.3 Bone Milling**

The most effective way to extract DNA from archaeological bone is to grind the material into a powder. A common way to do this is to use a mechanical mill, employing a metal rod and oscillating magnets to pulverize the solid bone sample in a vial. Though very effective at creating a fine powder, this aggressive and rapid milling also exposes the sample to high temperatures that can damage the target DNA (Liang et al. 2002). One method to counteract the heat is to use a chilled miller. The mill used in this project cooled the samples prior to and during the milling process, greatly reducing the heat the sample is subjected to and mitigating the potential template

damage due to temperature. Additionally, chilled millers have been shown to produce a finer and more efficient powder than non-chilled models, as the low temperatures creates more brittle bone that is more easily ground (Liang et al. 2002).

The mill used in this experiment uses liquid nitrogen (-210°C) to cool samples prior to and during milling. Prepared bone samples were loaded into a plastic vial with a metal rod. The vial is sealed with metal stoppers at each end and loaded into the millers. Samples are chilled in the liquid nitrogen for 5 minutes before being mechanically milled by the metal rod. The vial is then removed and the powdered sample transferred to a sterile 2mL Eppendorf tube using a scoopula. Modern samples were prepared last to avoid potential contamination through transfer in the instruments.

This process was repeated for each of the samples (excluding the Kenora bison sample, which had already been powdered for a previous experiment at the Paleo-DNA Laboratory). The vials, rod, scoopula, and stoppers were cleaned thoroughly with 2% hypochlorite bleach and Millipore water between each trial.

### **2.1.4 Bone Demineralization**

Approximately 0.200g of each sample were weighed out and placed in sterile 1.5mL tubes. An aliquot of 1.000mL of 0.5M EDTA (pH 8.01) was added to each sample. A demineralisation negative control was also employed. Each tube was then vortexed until the powder was suspended and left to sit at room temperature under gentle agitation.

The EDTA solution was changed every 1-3 days. To do this, the tubes were centrifuged at 25°C and 7500rpm for 15 minutes. As much of the supernatant was then removed with a pipette and transferred to a clean 1.5mL tube, taking care to not disturb the pellet. The supernatant was saved and stored at -4°C for further analysis. Observations of the quality of the

supernatant and pellet were made whenever the supernatant was changed. Samples were then vortexed until the pellet was fully resuspended. Extra agitation of the pellet was done with either a syringe tip or pipette tip if the pellet did not fully resuspend. The tubes were left at room temperature under medium agitation until the next supernatant change. These steps were repeated until the pellets were consistently gelatinized and the supernatant was clear after several days of incubation, a period of approximately 3 weeks and 8 solution changes.

### **2.1.5 Quantitation**

Quantitation in this project used an Epoch spectrophotometer (BioTek), a Take 3 plate (BioTek), and Gen5 software (BioTek, version 1.10.8) running on a desktop using Windows XP. Absorbance at wavelengths of 260nm, 280nm, and 320nm were taken during each read. Prior to loading, the Take 3 plate was cleaned using 75% ethanol and a Kimwipe (Kimberly-Clark) on both the wells and the contact side of the glass plate. The non-contact side of the glass was also cleaned in this same manner as required.

First, 2 $\mu$ L of the resuspension solvent was loaded using a micropipette with appropriate tip and read as a reagent blank. Then, 2 $\mu$ L of each sample were loaded in duplicate using a micropipette with appropriate tip, up to 7 samples were analysed at a time, on the Take 3 plate. The plate was loaded into the spectrophotometer, and readings were taken. Results of the quantification were exported to an Excel (Microsoft) spreadsheet.

## **2.2 Nucleic Acid Methods**

### **2.2.1 Extraction**

The extraction protocol used was carried out in two steps. First, approximately 0.2g of powdered bone sample was measured and added to a 2mL Eppendorf tube. Then, 1.4mL of 0.5M

EDTA, 75 $\mu$ L of 20% Sarkosyl solution, and 75 $\mu$ L of 10mg/mL proteinase K were added to the bone sample. The samples were vortexed overnight and left to incubate at 56°C under gentle agitation overnight.

Following this, the contents of the 2mL tubes were transferred to a sterile 15mL tube. A volume of 2.7mL of 3M Guanadinium thiocyanate (GuSCN) solution and 15 $\mu$ L of silica bead suspension was added to each sample, which were left to incubate overnight at 4°C. The contents of the 2mL tubes were transferred using 3 consecutive washes of 900 $\mu$ L of GuSCN.

100 $\mu$ L of supernatant from the previous demineralization step was also taken as an “extract” and used directly in the purification step below.

## **2.2.2 Purification**

### **Ethanol Precipitation**

Quantitation revealed the potential for DNA isolation directly from the EDTA supernatant from the earlier demineralization of bone (section 2.1.4). To accomplish this, an ethanol precipitation step was performed. First, 3M sodium acetate was added to 750 $\mu$ L of supernatant to a final concentration of 10% v/v in a 1.5mL tube. The tube was then vigorously vortexed for 1 minute. A volume of 2.5 times the purification volume of cold 100% ethanol was added to the tube and the sample was placed on ice for 30 minutes. The tubes were centrifuged at 13,000rpm for 5 minutes to generate a pellet. As much ethanol as possible was removed by pipette with care taken not to disturb the DNA pellet. A volume of 1.000mL of 95% ethanol was added to each sample and vortexed briefly followed by another 5 minutes of centrifugation at 13,000rpm. The ethanol was again removed by pipette and the tube was left to air dry. The pellet was resuspended in 150 $\mu$ L of sterile water and incubated at 37°C for 15 minutes. The samples

were re-quantified to assess the efficacy of the purification, and additional purification steps were taken as required.

### **Size-exclusion Chromatography**

Size-exclusion chromatography was used as an additional purification step, as needed, on samples using commercially available P30 spin columns (Bio-Rad). First, the size exclusion chromatography columns were placed in a clean 2mL tube and centrifuged at 1,000g for 2 minutes to remove the packing buffer. Next, the P30 columns were transferred to a new, sterile 2mL collection tube and 100 $\mu$ L of the sample was then transferred by pipette to the column. These were centrifuged again at 1,000 G for 4 minutes. The eluate was stored in the 2mL collection tube and the columns were discarded. Columns that were dried out were resuspended in TNE buffer when needed.

### **2.2.3 Amplification**

Amplification was carried out using PCR. All reactions were done in sterile 0.25mL PCR tubes on a MyCycler thermocycler (BioRad). Reactions were comprised of the reagents following Table 2 and 3. The DNA polymerase used was either Accustart II (Qiagen) or GoTaq Green (Promega). Primers were obtained from Qiagen following the oligonucleotide sequences used in

**Table 2:** Reagent conditions for amplification by PCR

<b>Reagent</b>	<b>Amount (<math>\mu</math>L)</b>
2x Polymerase Supermix	12.5
Forward Primer (10 $\mu$ M)	0.5
Reverse Primer (10 $\mu$ M)	0.5
Template	5.0
Water	6.5
<b>Total</b>	<b>25</b>

Shaprio et al. (2004) (see Table 12 for primer sequences). Amplified product was stored at 4°C until detection by gel electrophoresis.

### Amplification with Human Serum Albumin

Amplification for testing the efficacy of HSA were comprised of the reagents following Table 3.

**Table 3:** Reagent conditions for the amplification by PCR with the addition of HSA

Reagent	Amount (µL)
2x Polymerase Supermix	12.5
Forward Primer (10µM)	0.5
Reverse Primer (10µM)	0.5
Template	5.0
Human Serum Albumin (50g/L)	0.5
Water	6.0
<b>Total</b>	<b>25</b>

### Thermal Cycling Conditions

All PCRs were carried out under the cycling conditions in Table 4.

**Table 4:** Cycling conditions for all polymerase chain reactions.

Temperature (°C)	Time (minutes)	Number of Cycles
95.0	1:00	1
95.0	0:30	40
55.0	0:30	
72.0	1:00	
72.0	5:00	1
4.0	Hold	-

## **2.2.4 Detection**

Detection was performed using agarose gel electrophoresis. A 2% gel was prepared by mixing approximately 0.7400g of powdered agarose with 37mL of 1x Tris-Borate-EDTA (TBE) buffer in an Erlenmeyer flask and heating it in a microwave at 30s intervals until the agarose had completely dissolved, agitating gently as needed. A volume of 2 $\mu$ L of ethidium bromide (EtBr) was added to the liquid agarose solution before pouring the solution into a gel tray and being left to cool until solid. The gels were then submerged in 1x TBE buffer before the samples and 3 $\mu$ L of ladder (GeneRuler 100bp, Fermentas) were loaded into the wells. Prior to loading, 5 $\mu$ L of amplified product was mixed with 1 $\mu$ L of loading dye (Fermentas). Gels were then subjected to approximately 100V of electricity for approximately 45 minutes, or until the loading dye had sufficiently migrated.

Images were captured using UV light in a Universal Hood II (Bio-Rad) with QuantOne software running on a Windows XP desktop computer. Image processing and annotation was performed using the Fiji package for ImageJ software (National Institute of Health) on a computer running Macintosh OSX Sierra.

## **2.2.5 Sequencing**

Sequencing was performed off-site at the Lakehead University Paleo-DNA laboratory. The sequencing method used was the Sanger capillary electrophoresis dye-termination method, taking both forward and reverse sequences. BigDye termination kits (Fisher) were used for the sequencing reactions.



## **2.2.6 Sequence Analysis**

Raw sequence files were obtained from the Lakehead University Paleo-DNA laboratory as AB1 files. These raw files were imported into BioEdit v7.0.5 (Hall 2005) on an Insignia Flex laptop running Windows 10. BioEdit was used to clip, clean, and edit sequences, eliminating unclear sections at the beginning and end of each sequence and making manual base calls where the sequencing software could not determine the correct base. Once sequences were determined, they were aligned using BioEdit's ClustalW client against a modern bison reference sequence. This sequence alignment was used to create consensus sequences from individual amplicons.

The BLAST was used for comparison. Consensus sequences generated from BioEdit were searched using the nucleotide BLAST (nBLAST) search, with parameters set to search against "other" sequences.

## **2.2.7 Phylogeny**

Phylogenetic analysis was performed using the MEGA6 program (Tamura 2013) on a computer running Windows 10. A total of 328 sequences previously published by Shapiro et al. (2004) were obtained from GenBank and added to a file containing the sequence generated from ISM11. Sequences from Shapiro et al. (2004) were trimmed to match the length and location of the generated sequence. Genetic distances between individuals were computed using a Maximum Composite Likelihood algorithm (Tamura et al. 2004) and otherwise default parameters. A phylogenetic tree was generated using the Neighbour-Joining method (Saitou et al. 1987), as had been done in the Shapiro et al. (2004) study to ensure the best continuity in results.

A simplified version of this tree was also generated to include only the ISM11 individual, the closest genetic match from the first tree, and one individual from each of the four principle clades identified in previous studies.

## 2.3 Proteomic Methods

### 2.3.1 Alkylation, Denaturation, and Digestion

Following the EDTA demineralisation, pellets were washed by adding 500 $\mu$ L of sterile water, vortexing thoroughly, then centrifuging at 13,000rpm for 15 minutes to pellet, repeating once this whole process once. The protein pellet was then suspended in 800 $\mu$ L of 50mM ABC. The protein was then reduced by adding DTT to a final concentration of 5mM incubating for 1 hour at 60°C. The mixture was then alkylated by adding IAA to a final concentration of 15mM and incubating for 45 minutes at room temperature in the dark. The extracted proteins were digested by adding 1 $\mu$ L of 1 $\mu$ g/ $\mu$ L of trypsin and ACN to a final concentration of 8% v/v and incubating for 18 hours. This digestion was quenched by acidifying with 8 $\mu$ L of 1% TFA and incubating at 37°C for 30 minutes and centrifuging at 13,000rpm for 30 minutes.

### 2.3.2 Purification

#### Bind-Elute Tips

Following digestion, protein mixtures were purified using commercially available C18 Millipore ZipTip pipette tips (Thermo-Scientific). Tips were first prepared by adding and discarding one bed volume of 50% ACN followed by one bed volume of 0.1% TFA. Protein samples were aspirated through the tip for 10 cycles. The tip was washed with 10 $\mu$ L of a mixture of 0.1% TFA/5% ACN. The tip was washed a second time using the same parameters. Protein samples were eluted in 10 $\mu$ L of 0.1% formic acid into a clean gas chromatography (GC) vial with an insert that had been sterilized with an acid wash prior to use. Tubes were sealed and capped before mass spectrometry.

### **Organic Phase Separation**

Two organic phase separations were attempted in this project: one with phenol:chloroform and one with only chloroform.

For the phenol:chloroform separation, a stock solution was first prepared by mixing 1000 $\mu$ L of phenol with 1000 $\mu$ L of chloroform in a sterile 2mL tube, vortexing to combine and create a 1:1 mixture. A volume of 750 $\mu$ L of the phenol:chloroform solution was then added to 750 $\mu$ L of sample and that mixture was vortexed for 0.5 minutes. at high speed. The mixture was centrifuged at 13,000rpm and 4°C for 1 min. to create a clear phase separation. The upper aqueous phase was transferred, via pipette, to a clean 1.5 $\mu$ L tube for further analysis. The organic phase was left to evaporate overnight. This same procedure was performed with 750 $\mu$ L of chloroform alone. Once the organic phase had completely dried off, the protein pellet was resuspended in 75 $\mu$ L of sterile water and incubated at 37°C for 15 min.

### **2.3.3 Liquid Chromatography - Tandem Mass Spectrometry**

Samples were analyzed using the facilities at the Lakehead University Instrumentation Lab. A Dionex Ultimate 3000 UHPLC and Bruker amaZon X Ion Trap spectrometer were used for the LC-MS/MS. For the HPLC procedure, 5 $\mu$ L of each sample was run through a C18 3 $\mu$ m, 120 Å column (Acclaim 120) at a flow rate of 0.200 mL/min with an ACN gradient from 45% to 85%. The MS was set to operate within a range of 50-3000amu, with Auto MS/MS and SmartFrag enabled, with a cut off selection at 20% and an absolute threshold of 25000. Molecules were dissociated under CID between sections of tandem MS runs.

### **2.3.4 Protein Analysis**

Mass spectrometry files were obtained as MGF files for analysis. Files were uploaded to the GPM using their browser-based server-side client, found at [www.thegpm.org](http://www.thegpm.org). Using the

X!P3 algorithm, files were set to search against a male *Homo sapiens* database (there was no bison database available, and this represented the largest mammalian library to compare against). Fixed modifications were added for carboxymethylation, with potential modifications for deamidation at N and Q and at hydroxyproline (15.994915@P). Trypsin was selected as the digestion enzyme with semi-style cleavage enabled and the detection instrument parameter was set to Ion Trap. Once peptide sequences were obtained from the GPM, they were searched using the protein Basic Local Alignment Search Tool (BLASTp), with default parameters.

### 3. Results

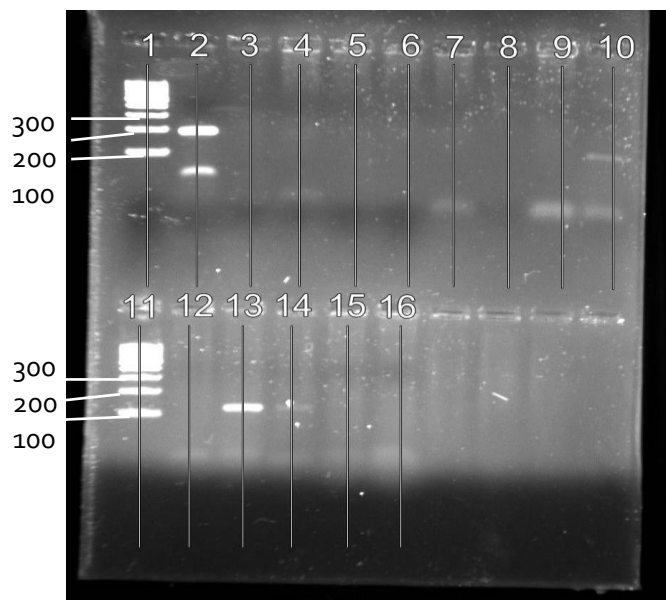
A DNA sequence 156bp long was obtained from one individual (ISM 11), while protein sequences were obtained from 4 individuals (ISM 2, 5, 6, and 11) (Table 5). Notably, both DNA and protein sequences were recovered for the same individual, ISM 11, using the same extract (see methods section).

**Table 5:** Summary of results from all samples worked with, including amount of sample used, results of nucleotide and peptide sequencing, and quantification data. Quantification was performed on an Epoch Spectrophotometer (Biotek) using a Take 3 microwell plate and Gen5 software, using 2 $\mu$ L samples performed in duplicate. Protein values represent amount detected following digestion. Nucleic acid values represent numbers following ethanol precipitation purification.

Sample	Amount of bone powder (g)	Length of Sequence (bp)	No. of Proteins Identified	Protein Concentration (mg/mL)	DNA Concentration (ng/ $\mu$ L)	Purity (260/280)
ISM 1	0.1950	-	-	4.54	28.49	1.69
ISM 2	0.2245	-	26	6.46	35.52	1.68
ISM 3	0.1916	-	-	6.04	0.82	1.68
ISM 4	0.1904	-	-	5.14	30.33	1.64
ISM 5	0.1803	-	2	7.60	30.45	1.60
ISM 6	0.1822	-	2	6.68	15.27	1.59
ISM 7	0.2001	-	-	6.55	22.78	1.57
ISM 8	0.1922	-	-	3.49	2.55	1.38
ISM 9	0.2200	-	-	3.19	2.23	1.22
ISM10	-	-	-	-	-	-
ISM 11	0.2205	156	4	2.75	2.35	1.67
Kenora	-	-	-	-	-	-

### 3.1 Nucleic Acids

Amplification by PCR was successful on one of the bison samples, ISM11 (Figure 6, Table 6). Two amplicons were obtained from the 12 primer sets used, which were then sequenced by Sanger sequencing (Figures 7 and 8). The consensus sequence was then searched with the Basic Local Alignment Search Tool (BLAST, NCBI), matching closest with *Bison priscus* and *Bison bison* individuals, with a high degree of confidence (Figures 9, 10, and 11).

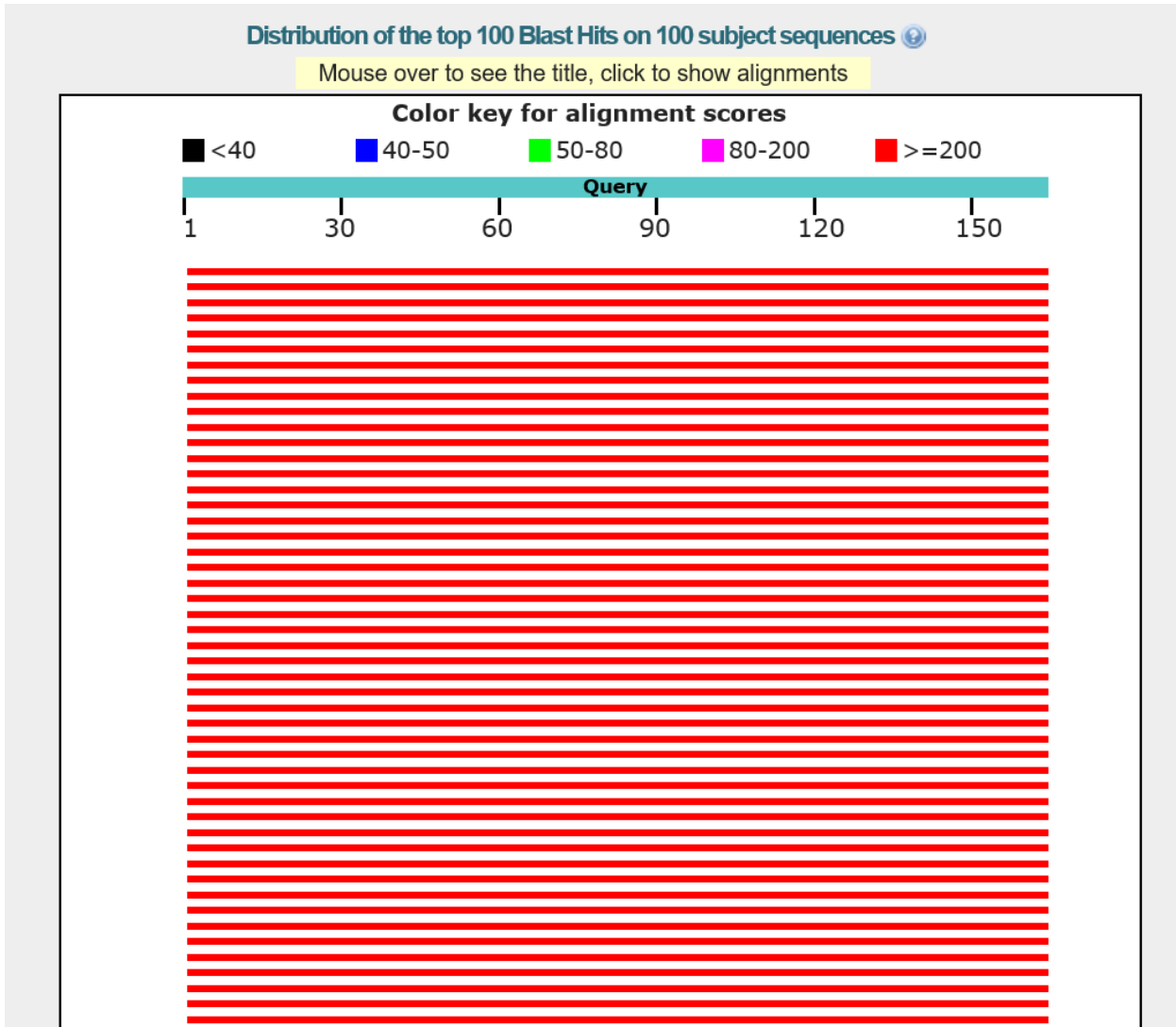


**Figure 6:** Amplified product of ISM 11 visualized on a 2% agarose gel made in TBE buffer with 2 $\mu$ L ethidium bromide. Image taken under UV light with a Universal Hood II (BioRad) and QuantityOne software. Lane identities are summarized in Table 6.

Lane #	Identity	Lane #	Identity	Lane #	Identity	Lane #	Identity
1	GeneRuler 100bp	5	PS2	9	PS6	13	PS9*
2	Positive Control	6	PS3	10	PS7*	14	PS10
3	Negative Control	7	PS4	11	GeneRuler 100bp	15	PS11
4	Primer Set (PS) 1	8	PS5	12	PS8	16	PS12

**Table 6:** Summary of lane identities from gel electrophoresis of amplified ISM 11 sample (Figure 6). Asterix denotes amplified product.






**Figure 9:** Graphical view of the top scoring alignments from a Basic Local Alignments Search Tool (BLAST) nucleotide search of the sequence from ISM11 seen in Figure 8.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected:0

[Alignments](#) [Download](#) [GenBank](#) [Graphics](#) [Distance tree of results](#) 

	Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/>	<a href="#">Bison priscus isolate AF028 voucher PMA P99.2.28 mitochondrion, complete genome</a>	266	266	99%	1e-67	93%	<a href="#">KX269121.1</a>
<input type="checkbox"/>	<a href="#">Bison priscus isolate AE033 voucher SFU 3429 mitochondrion, complete genome</a>	266	266	99%	1e-67	93%	<a href="#">KX269117.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate PH126 tRNA-Pro gene and D-loop, partial sequence; mitochondrial</a>	266	266	99%	1e-67	93%	<a href="#">KU705809.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate PH125 tRNA-Pro gene and D-loop, partial sequence; mitochondrial</a>	266	266	99%	1e-67	93%	<a href="#">KU705808.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate AF012 tRNA-Pro gene and D-loop, partial sequence; mitochondrial</a>	266	266	99%	1e-67	93%	<a href="#">KU705771.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate AF009 tRNA-Pro gene and D-loop, partial sequence; mitochondrial</a>	266	266	99%	1e-67	93%	<a href="#">KU705768.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate AE033 tRNA-Pro gene and D-loop, partial sequence; mitochondrial</a>	266	266	99%	1e-67	93%	<a href="#">KU705766.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate BTSBH1005 mitochondrion, complete genome</a>	266	266	99%	1e-67	93%	<a href="#">GU947003.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate B1428 mitochondrion, complete genome</a>	266	266	99%	1e-67	93%	<a href="#">GU946999.1</a>
<input type="checkbox"/>	<a href="#">Bison bison isolate B1191 mitochondrion, complete genome</a>	266	266	99%	1e-67	93%	<a href="#">GU946998.1</a>

**Figure 10:** Top ten alignment identities from a Basic Local Alignments Search Tool (BLAST) nucleotide search of the sequence from ISM11 found in Figure 8. Alignment scores correspond to the graphic view in Figure 9.



Bison priscus isolate AF028 voucher PMA P99.2.28 mitochondrion, complete genome  
 Sequence ID: [KX269121.1](#) Length: 16320 Number of Matches: 1

[Related Information](#)

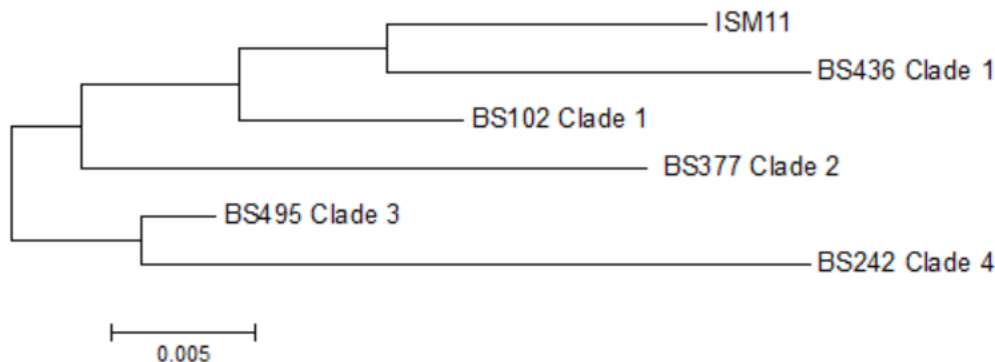
Range 1: 16075 to 16250 [GenBank](#) [Graphics](#) ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
266 bits(294)	1e-67	163/176(93%)	13/176(7%)	Plus/Plus
Query 2		GTACATAGCACATTATGTCAAATCTACCC TTGGCAACATGCATATCCCTTCCATTAGATC		61
Sbjct 16075		GTACATAGCACATTATGTCAAATCTACCC TTGGCAACATGCATATCCCTTCCATTAGATC		16134
Query 62		ACGAGCTTAATTACCATGCCGCGTGAAC CAGCAACCCGCTAGGCA-----C		108
Sbjct 16135		ACGAGCTTAATTACCATGCCGCGTGAAC CAGCAACCCGCTAGGCAAGGATCCCTCTTC		16194
Query 109		TCGCTCCGGGCCATGAACCGTGGGGT CGCTATTTAATGAAC TTATCAGACATC		164
Sbjct 16195		TCGCTCCGGGCCATGAACCGTGGGGT CGCTATTTAATGAAC TTATCAGACATC		16250

**Figure 11:** Alignment view of the top alignment from the Basic Local Alignments Search Tool (BLAST) nucleotide search of the sequence from ISM11 found in Figure 10.

## 3.2 Phylogeny

A phylogenetic tree was generated using sequences from Figure 8, computed using the Neighbour-Joining algorithm in MEGA6 software (Tamura et al. 2013). The sequence obtained from sample ISM11 was compared with 326 aDNA bison sequences (Shapiro et al. 2004 Wilson et al. 2008). A simplified phylogenetic tree was generated comparing sample ISM11 with the sequences representing the four main bison clades (Figure 12). Clade 1 corresponds primarily to ancient bison from southern North America (i.e. south of the Laurentide ice sheet), Clades 2 and 3 primarily represent bison from Eastern Beringia, and Clade 4 is primarily occupied with Western Beringian ancient bison.



**Figure 12:** The evolutionary history was inferred using the Neighbour-Joining method (Saitou et al. 1987). The optimal tree with the sum of branch length = 0.09598157 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004) and are in the units of the number of base substitutions per site. The analysis involved 6 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 80 positions in the final dataset. Evolutionary analysis were conducted in Mega6 (Tamura et al. 2013). This is a simplified version of the full tree found in the appendix (Figure 25).

### 3.3 Proteomics

Out of the twelve samples, four were successfully analyzed via LC-MS/MS. A summary table of results from two individuals are presented: ISM11 (Table 7 and Figure 13) and ISM2 (Table 8). ISM2 represented the highest number of proteins identified (26 in total; Table 5), while ISM11 was the individual from which both nucleotide and protein sequences were obtained (Table 5).

**Table 7:** Summary of all Peptide matches for ISM11 from a GPM search following LC-MS/MS using the X!P3 algorithm and taxa classification based on subsequent protein BLAST search.

Protein	Gene	Peptides	Taxa
Collagen type 1 alpha 1	COL1A1	GKHGNRGETGPSGPVGPAGAVGRGPSQPQIR	Placentals
Steroid Receptor RNA Activator 1	SRA1	TSPGPPPMGPPPPSSKAPR	Placentals
Dedicator of Cytokinesis 7	DOCK7	LLRHCSSTIGTIRSHASASLY	Vertebrates
Ubiquitin-like Modifier Activating Enzyme 5	UBA5	REGVCAASLPTTMGVVAGILVQNVL	Vertebrates

Models from 'ISM11\_150\_July\_27\_2017.mgf': Main model display

**gpm** GPM33000020758 [get annotation](#)

Contributor: anonymous

Note: Overall data quality low. p-score = 23

enter keywords  log(e) < -1 # > 0 Display:     p = 23, FPR=4.50% ?

rank	log(e)	log(l)	%/%	#	total	Mr	Accession
1	-1.9	5.29	2.3/3	1	1	129.2	ENSP00000297268 gpmD8   psyt   snap <sup>+</sup> protein peptide COL1A2 <sup>np</sup> , collagen type I alpha 2 chain [Source:HGNC Symbol;Acc:HGNC:2198] IPR008160 (*6) Collagen triple helix repeat IPR000885 (*4) Fibrillar collagen, C-terminal
2	-1.4	5.18	7.6/11	1 <sub>4</sub>	1 <sub>4</sub>	25.7	ENSP00000337513 gpmD8   psyt   snap <sup>+</sup> protein peptide SRA1:p, Steroid receptor RNA activator 1 (Steroid receptor RNA activator protein) (SRAP). Source: Uniprot/SWISSPROT Q9HD15 Annotated Domains: IPR006077 Vinculin/alpha-catenin IPR000694 Proline-rich region IPR009917 Steroid receptor RNA activator IPR006030 Molluscan rhodopsin C-terminal tail
3	-1.4	5.16	0.9/1	1	1	239.3	ENSP00000340742 gpmD8   psyt   snap <sup>+</sup> homo (1/1) protein peptide DOCK7 <sup>sp</sup> , dedicator of cytokinesis 7 [Source: HGNC 19190] IPR016024 ABIM-type fold IPR010703 DOCK C IPR021816 DOCK C/D N
4	-1.1	5.16	5.9/9	1	1	44.8	ENSP00000348565 gpmD8   psyt   snap <sup>+</sup> homo (4/4) protein peptide UBA5:p, ubiquitin-like modifier activating enzyme 5 [Source: HGNC 23230] IPR009036 Molybdenum cofactor synthase MoeB IPR000594 Thif-NAD FAD-bd

3.3/4.8 #2 #2

**Figure 13:** Typical Global Proteome Machine search results, shown here from sample ISM11 following Tandem Mass Spectrometry using the X! P3 algorithm against a *Homo sapiens* database with complete modifications for carbamidomethylation at C and U, potential modifications for oxidation at M, deamidation at N and Q, modified hydroxyproline, and oxidation at M and W, with semi-style cleavage with trypsin and Ion Trap detection.

**Table 8:** Summary of all peptide matches for ISM 2 from a GPM search following LC-MS/MS using the X!P3 algorithm and taxa classification based on subsequent protein BLAST search (Page 1 of 2).

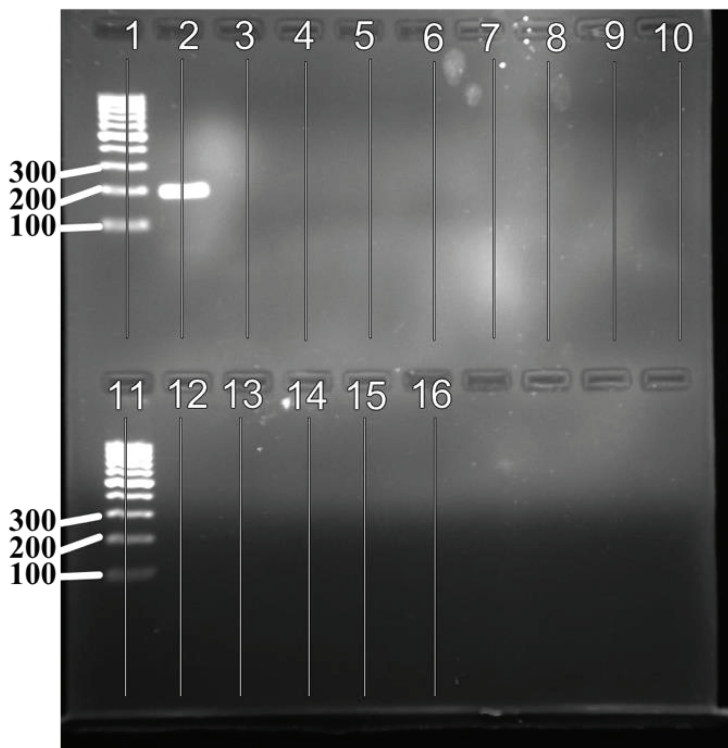
Protein	Gene	Peptides	Taxa
Collagen 1 alpha 1	COL1A1	RGLPGPPGAPGPPQGFQGGPPGEPGEPGASGPMGPR EPGDAGAKGDAGPPGPAGPAGPGPIGNVGAPGAK GETGPAGPAGPVGPVVGARGPAGPQ	Placentals
Lysosomal a-mannosidase	MAN2B1	PQVVLAPGGGAAYNLGAPPR	Placentals
Cell migration inducing hyaluronan	KIAA1199	GVIVHVIDPKSGTVIH	Placentals
Mannosyl glycoprotein-b	MGAT5	ESGFKIAETASGGPLGELVQW	Vertebrates
DNA methyltransferase 3A	DNMT3A	SEPQPEEGSPAGGQKGGAPAEGEGAA	Placentals
Synaptic Ras GTPase activating protein 1	SYNGAP1	SLSKEGSIGGSGGSGGGGGGGLKPSITK	Placentals
Collagen type 16 alpha 1	COL16A1	PGQPGYPGATGPPGLPGIKGER	Vertebrates
Transcription factor SOX-4	SOX4	GGSGGGGHGGGGGGSSNAGGGGGGASGGGANS KPAQKK	Cellular organisms
Unknown	Unknown	SAHLPGHQGPSSPASPPR	Cellular organisms
XK-Related protein	IFFO2	GEAAAAAGPPGVVGAGGPGPRYE	Cellular organisms
Plekstin homology D-containing family N iso 1	PLEKHN1	EHAFAQITGPLPAPLLVLCPSRAEL	Placentals

**Table 8:** Summary of all peptide matches for ISM 2 from a GPM search following LC-MS/MS using the XIP3 algorithm and taxa classification based on subsequent protein BLAST search (Page 2 of 2).

Protein	Gene	Peptides	Taxa
Reticulon 1	RTN1	GPGPLGPGAPP	Cellular organisms
Radphilin 3A isoform 1	RPH3A	TGPDPASAPGRGNY	Placentals
Microtubule cross-linking factor 1	SOGA2	GAPPGSPEPPALLAAPLAAGACPGGRSI	Placentals
Collagen type 6 alpha 3	COL6A3	QGTRGAQGPAGPAGPPGLIGEQQISGPR	Placentals
Spectrin b-chain, brain 2	SPTBN2	AQQFYRDAEAEAWMGEQEL	Mammals
CUB and Sushi multiple D1	CSMD1	IRYSCLPGYILEGHAILTCIVSPGNGA	Vertebrates
E1A-binding protein p400 isoform X1	EP400	RPYLSSPLRAPSEESQDYHKK	Primates
B-cell antigen receptor	CD79A	FLGPGEDPNGTLIIQNVNK	Cellular organisms
Bioorientation of chromosomes/cell	BOD1L2	MADGGGGGGSGGAGPASTRASGGGGPINPASLPPGD	Placentals
Cysteine protease ATG4D	ATG4D	GASGPALGSPGAGPSEPDEVDFKFK	Placentals
G-nucleotide binding	GNB1L	WRMQPLAVLAFHSAAVQCVAFTADGLLAAGSK	Placentals
Collagen 7 alpha 1	COL7A1	GPPGSVGPAGASGLKGDK	Vertebrates
Testicular protein kinase	TESK1	METALPGPGPPAVGPS	Placentals
Chondroitin poly factor	CHPF	QAFHPAVAPPQGGPPPELGR	Mammals
Chromosome 11 open reading frame 48	C11orf48	ACTASGAPTGAPR	Cellular organisms

## Human Serum Albumin (HSA)

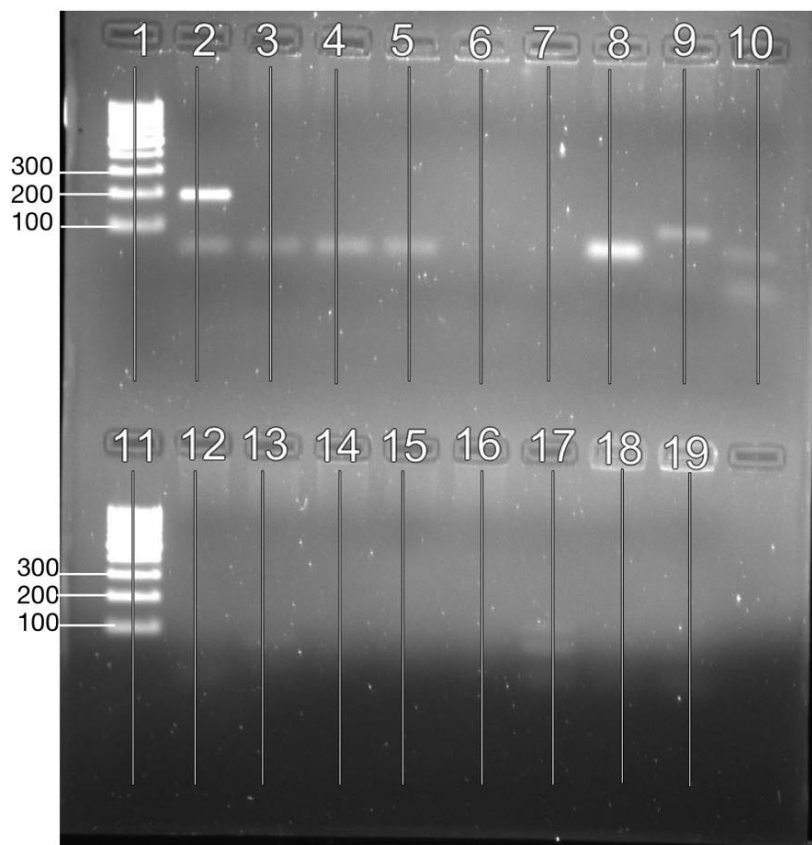
The use of HSA was evaluated for its effectiveness as a reliever of PCR inhibition. Attempts to amplify sample ISM 3 has shown inhibition (Figure 14). Repeated amplification with the use of HSA has ameliorated the inhibition (Figure 15).



**Figure 14:** Attempted amplification of ISM 3 visualized on a 2% agarose gel made in 1x TBE buffer with 2 $\mu$ L ethidium bromide. Image taken under UV light with a Universal Hood II (BioRad) and QuantityOne software. Lane identities are summarized in Table 9.

**Table 9:** Lane identities from gel electrophoresis of amplified ISM 3 (Figure 15). Primer sets correspond to those in Table 12.

Lane #	Identity	Lane #	Identity
1	GeneRuler 100bp (Fermentas)	9	PS6
2	Positive Control	10	PS7
3	Negative Control	11	GeneRuler 100bp (Fermentas)
4	Primer Set (PS) 1	12	PS8
5	PS2	13	PS9
6	PS3	14	PS10
7	PS4	15	PS11
8	PS5	16	PS12



**Figure 15:** Amplified product of ISM 3 with the addition of 1 $\mu$ L of 10mg/mL Human Serum Albumin visualized on a 2% agarose gel made in 1x TBE buffer with 2 $\mu$ L ethidium bromide. Image taken under UV light with a Universal Hood II (BioRad) and QuantityOne software. Lane identities are summarized in Table 10.

**Table 10:** Lane identities from gel electrophoresis of amplified ISM 3 with the addition of HSA (Figure 15). Primer sets correspond to those in Table 12.

Lane #	Identity	Lane #	Identity
1	GeneRuler 100bp (Fermentas)	11	GeneRuler 100bp (Fermentas)
2	Positive Control	12	PS8
3	Negative Control	13	PS9
4	Primer Set (PS) 1	14	PS10
5	PS2	15	PS11
6	PS3	16	PS12
7	PS4	17	PS7 with no HSA
8	PS5	18	PS7 with 0.5 $\mu$ L HSA
9	PS6	19	PS7 with 1.5 $\mu$ L HSA
10	PS7		



### 3.4 Chloroform Separation

The use of chloroform as a means of purifying protein was evaluated in this research. It was shown that chloroform can be used to increase protein purity, while losing 17.92% of the sample. This loss can be mitigated by incubating the sample for 15 minutes at 37°C (Table 11).

**Table 11:** Spectrophotometric readings from Human Serum Albumin (HSA) dissolved in 2mL of water before and after purification with chloroform and after purification and incubation for 15 minutes at 37°C. Percent loss is relative to unpurified HSA

HSA Unpurified	
Absorbance at 260nm	0.51225
Absorbance at 280nm	0.31925
260/280	0.6235
mg/mL of protein	10.2425
HSA Purified	
Absorbance at 260nm	0.4205
Absorbance at 280nm	0.2755
260/280	0.65525
mg/mL of protein	8.40725
Percent loss	17.92
HSA Purified Post-Incubation	
Absorbance at 260nm	0.448
Absorbance at 280nm	0.2975
260/280	0.6645
mg/mL of protein	8.953
Percent loss	12.59

## 4. Discussion

### 4.1 Nucleotide Sequencing

Successful amplification and sequencing was achieved on one ancient bison sample, ISM11. This produced two amplicons with two associated sequences (forward and reverse) each (Figure 8). These were then combined to a consensus sequence used in a BLAST search to establish identity. The sequence matched unambiguously to bison, matching database entries from both *B. bison* and *B. priscus* species (Figure 10). These results were statistically robust, with all top results returning a very low e-value ( $1e-67$ ) and high max score (all  $>200$ ) (Figure 9 and Figure 10). Confidence in authenticity is also aided by the methodological blanks consistently showing no amplification (Figure 6) and all primer sets being species-specific (i.e. will only amplify bison). Therefore, it can be said with confidence that the sequences obtained in this study are authentic, and that they align well with bison DNA.

### 4.2 Phylogeny

When considering phylogeny, the ISM11 individual was confidently placed into Clade 1 from the earlier determination of ancient bison relationship presented by Shapiro et al. (2004) and Wilson et al. (2008) and was distinctly separate from Clades 2, 3, and 4, as shown in Figure 12. This shows a clear affinity between this ancient bison sample and the bison that populated North America south of Beringia. The position of this bison within the southern clade is consistent with the previous interpretation of a genetically distinct and geographically isolated bison population ubiquitous across Holocene North America.

The Shapiro et al (2004) and Wilson et al (2008)-led studies suggested that since the *B. occidentalis* type specimen was from Northern Alaska (i.e. north of the Laurentide ice sheet during the last glacial maximum (LGM)), and since the bison from south of the Laurentide ice sheet represent a distinct and separate population with no mixture from northern populations, the species assignment of *B. occidentalis* was likely invalid in the North American mid-continent. In light of this interpretation, some archaeologists have revisited earlier archaeological site reports and re-classified individuals identified as *B. occidentalis* as the more genetically-consistent *B. bison* (Widga 2014). The phylogenetic results from this study provide direct evidence in support of this re-classification. However, the morphological differences that led to the original classification of these individuals cannot be ignored, and there is the distinct possibility that there is a relationship between Beringian bison and those from the southern portion of the continent that cannot be revealed by the DNA sequences studied here and previously.

The HV1 and HV2 segments of mtDNA explored here and previously do have a high capacity for resolution due to their affinity for mutations and are deservedly preferred targets in ancient population studies. However, they obviously do not represent the whole genome, and so the markers for relationship (although unlikely) may exist in another as-yet-unstudied portion of ancient bison genome. Therefore, an appropriate future direction would be to explore whole-genome sequencing across many ancient bison samples to establish a better understanding of genetic relationships and to determine if there is a link between the apparent phenotypic similarities between northern and southern ancient bison populations and their genotypes.

These initial results also suggest some potential distinction between Great Lakes-area bison and the others from southern North American, as seen by the ISM11 populating its own branch in Figure 25. This study did not reveal enough information to elucidate regional

distinction with any confidence, but further studies should explore the potential for establishing what a Great Lakes ancient bison population might have looked like genetically.

Though successful in some respects, this study did not obtain DNA sequences from all individuals. There are many potential reasons for this outcome. This is most likely due to taphonomic factors of samples and their depositional environment. Northern environments present soil conditions that are typically wet, acidic, and having a high concentration of humic and tannic materials. All these contribute to DNA degradation and amplification inhibition, when present in combination can make DNA sequencing extremely challenging. Several steps were taken to mitigate these effects. Multiple purification steps were used (e.g. ethanol precipitation, size-exclusion chromatography), the use of PCR additives (serum albumin) was explored, and best-practices were undertaken during sample preparation to minimize deleterious methodological effects (e.g. sample pulverizing in liquid-nitrogen). Despite their effectiveness in the literature, these did not produce DNA sequence results in all samples. This could be due to significant template degradation, providing no starting point for amplification, as a result of taphonomic conditions in the environment and storage conditions in curation. Future studies should consider a wider range of inhibition relievers, such as the addition of N-phenacyl thiazolium bromide (PTB) to mitigate DNA cross-linking (Poinar 2001), one form of inhibition not explored here.

The storage condition of bones following excavation has also been studied relative to DNA preservation. It has been shown that even storage for several decades in museum conditions can be as detrimental to DNA preservation as several millennia in the context of original deposition (Pruvost et al. 2007). These samples have been in storage for as long as nearly 80 years, and so there exists the strong possibility that effects of museum curation have

contributed to failed DNA sequencing. In nature, a cool, dry, and slightly alkaline environment is best for DNA preservation (Bollongino et al. 2008). Likewise, storage after excavation should mimic these conditions as closely as possible, otherwise the genetic information would face further or accelerated damage that could inhibit analysis.

## 4.3 Protein Sequencing

Peptide sequences were successfully obtained from four individuals (Figures 13, 21, 22, and 23). Collagen was positively identified, in some form, in each sample, with collagen type 1 being the most common (ISM 5 and ISM 11). Global Proteome Machine results were returned with consistently high confidence (False Positive Rate of ~5%), with BLAST results also returning results with a high degree of certainty (E-value of  $7e-11$  and max score of 63.9 in Figure 19). Confidence is added by the use of negative controls, which produced no results (Figure 24).

The protein identified with most confidence was collagen Type 1 Alpha 1 from individual ISM2 (Table 8). This identification was based on three polypeptides totaling 92 amino acids in length, and when searched in a BLASTp revealed an affinity only as specific as placental mammals (Table 8). This level of resolution was similar across all peptide searches using protein BLAST.

In light of these results it is possible to conclude that this method will consistently obtain a variety of authentic protein sequences, particularly collagen. This is consistent with the previous studies in the literature that have used similar methods (e.g. Collins et al. 2011; Cappellini et al. 2011). However, this study has not achieved the same level of resolution of those previous publications.

There are a few possible sources for these limitations. The most likely reason is sub-optimal use of the database search tool (The GPM). The GPM relies on comparison of mass spectra to a database (the GPMdb) of other mass spectra, and so a protein similar to the analyte must be present for an identity to be established. The GPMdb does not contain a library of bison peptides, and so there is the possibility that some proteins were not identified even if they were successfully recovered simply because there did not exist a close match in the database. This limitation could be overcome using *de novo* peptide sequencing in the future. Another likely source of limitation is user error. Small changes in search parameters can produce drastically different results, and though all reasonable steps were taken to ensure best use of this tool, the possibility for better optimization still exists. Limitation as a result of sub-optimal use of the GPM is also encouraged through examination of raw mass spectra by an external reviewer, who was able to confirm that results up to database searching appeared typical of protein research (Dr. Matt Willets, personal communication). Better use of the GPM, or use of a different, more accessible peptide sequencing program may be worth exploring in the future.

## 4.4 Multiomics

Aside from interpretations on phylogeny, genetic nature, and protein nature of ancient bison, this study has important implications for multiomic approaches to future studies. Multiomics (research that combines multiple areas of global macromolecule analysis, such as genomics and proteomics) have great potential for more efficient projects with higher informational yield. This study has shown that it is possible to obtain both genomic and proteomic information from the same extract using a relatively simple procedure on ancient samples. This not only increases output, but conserves sample, allowing for more detailed studies

while using no more resources. This is significant for studies working on ancient material, where sample resources are often very limited.

At the time of writing, no published study has used a simultaneous extraction for both aDNA and proteins. Even studies directly comparing proteins and DNA from the same samples are electing to use separate extraction processes on precious archaeological samples (Wadsworth et al. 2017). This study provides a proof-of-concept and feasible workflow for a single-step multiple extraction for both nucleic and amino acids from archaeological bone. This can potentially save time, resources, and simplify procedures for ancient biomolecular studies.

## **4.5 Authenticity**

Verification of authenticity is significant whenever biological samples are concerned. The standard for authenticity when considering ancient genetics has been established by Poinar (2003), outlining a “top ten” list of criteria that researchers should follow to defend the validity of their results.

The first of these criteria recommends the use of a physically isolated work area for all DNA work. This project was able to utilize the facilities at Lakehead University to the best of its abilities, working in a lab that had not processed bison samples previously to reduce the risk of contamination. Amplified product did receive its own work and storage area, though it did not have its own physically isolated post-PCR area, as recommended by Poinar (2003).

The second criterion is the use of negative controls, which this project used rigorously. An experimental blank was carried through every stage of the experimental process, and results were not reported if any indication of contamination exogenous contamination appeared.

The molecular behaviour of the results is a key point supporting the authenticity of the

ancient results presented here. Authentically ancient samples tend to carry with them biochemical sources of inhibition, and that is indeed what was seen in the samples in this study.

Related to this, Poinar (2003) notes the importance of observing is the biochemical preservation of samples when considering authenticity. Ancient samples tend to be heavily degraded, and DNA strands will be present only in strands below 200 base pairs in length due to histone behaviour (Poinar 2003). This was consistent with the results of this study.

Though not noted by Poinar (2003) in his “top ten” paper, the use of species-specific primers has been cited as a way to confidently exclude the chance for contamination from exogenous sources (i.e. the humans conducted the experiment) (Yang and Watt 2004). Highly specific primer sets that will only amplify authentic bison samples were used in this study.

The results from this study also fall in line with previous ancient bison studies (e.g. Shapiro et al. 2004), adding authenticity by making reasonable phylogenetic sense. The phylogenetic results - generated using the same structure as the Shapiro et al. (2004) study - showed the bison individual ISM11 matched closely with other bison from south of the Laurentide ice sheet during the last glacial maximum. It is highly unlikely that these conclusions would have been reached if they had been introduced from an external source.

Though the above criteria were met and present a strong case for the authenticity of the results, there are some of Poinar’s (2003) criteria that were not met in this study. Cloning was not performed and samples were not sent for independent replication at a separate lab. This study is conscientious of these criteria, though they were not all performed due to practicality and cost concerns. Nevertheless, the results from this project are presented here with strong confidence in their authenticity.



## 4.6 Human Serum Albumin for Inhibition

### Relief

The feasibility of HSA as an alternative to BSA was explored in this study. Although, BSA has been used for over two decades when working with amplification inhibition, and has become a common component of the molecular biologist's toolkit (Kreader 1996). Since BSA is a component of bovid biology, there are obvious concerns when considering contamination in bovid and bovid-like species, and so homologous alternatives derived from a different biological source are worth evaluating.

The addition of 1 $\mu$ L of 10g/L HSA to a 25 $\mu$ L PCR reaction was found to alleviate inhibition, consistent with the concentrations of BSA usually used (Kreader 1996). This can be seen clearly by comparing Figures 14 and 15. In an uninhibited reaction, one would expect to find either targeted or non-specific product, indicating proper enzyme activity. However, neither of these are observed in the reaction represented in Figure 14, despite proper PCR parameters, indicated by the successful positive control. This shows that the DNA polymerase enzyme could not act as needed, indicating inhibition.

Following the addition of 1 $\mu$ L of 10g/L of HSA, there was a clear presence of non-specific product (Figure 15), and though target product did not amplify, this shows that the addition of HSA can relieve inhibition and return enzyme activity. This finding is consistent with previous studies that have used alternate sources of serum albumin (rabbit serum albumin, for example) to relieve PCR inhibition when studying bovid DNA (Shapiro et al. 2004).

## 4.7 Chloroform Separation for Protein

### Purification

In addition to the conventional separation methods tested, this study explored the use of a single-solvent organic phase separation for the purification of proteins with the use of chloroform. This examination was evaluated on two fronts. First, protein concentration was quantified spectrophotometrically for concentration and purity. It was found that the chloroform separation resulted in a high purity (260/280 ratio of 0.66) and 17.92% loss of analyte (Table 11). This percent loss was further mitigated by incubating the sample for 15 minutes at 37°C, reducing this value to 12.59% (Table 11). This improved recovery is likely due to sample being bound to the inside surface of the Eppendorf tube, which is then released back into solution when incubated.

For comparison, a phenol:chloroform mixture was also used in an attempt to purify this protein solution. However, this workflow relies on the liquid fraction of the protein-containing organic phase drying out, leaving the solid protein behind. The phenol did not evaporate at all (in fact, solidifying) and proved unsuitable for the scope of this evaluation.

The feasibility of this purification was also tested practically on an ancient sample (ISM5). The use of the relatively simple chloroform purification was successful, as peptide sequences were obtained following purification (Figure 22). This organic phase separation presents a simpler alternative for protein purification, has the potential for a one-step purification of both nucleotides and proteins.

## 5. Conclusion and Future Directions

This project was able to establish a feasible workflow for a multiomic approach to ancient material. It was shown that a single powdered bone sample of approximately 0.200g could be treated with EDTA before continuing on to separate genomic and proteomic analysis, and that usable information could be gathered from both of these phases.

Using this information, indications of relationship between ancient bison from the North American mid-continent are presented. It was found that for the samples from which sequences could be obtained, the bison in this project genetically resembled those from southern North America, in agreement with the previous studies published by Shapiro et al. (2004) and Wilson et al. (2008). This in turn augments the robustness of recent reclassification of ancient bison remains from *B. occidentalis* to *B. bison* made in recent re-examinations of curated material (Widga 2014).

In the proteomics stream, multiple proteins and associated peptide sequences were identified from four ancient bison samples. This was most notably collagen - the targeted protein in this study, though many other non-collagenous proteins were also found. Though limited at the present time, there exists great potential for further proteomic findings from these samples.

This project has also presented two evaluations of methodological enhancements. For one, the use of Human Serum Albumin (HSA) has been clearly shown to relieve inhibition in PCR using concentrations similar to those typically used with BSA, showing the feasibility of HSA as a PCR additive. Secondly, the effectiveness of chloroform alone as a way to purify proteins was evaluated, and the addition of an incubation step of 15 minutes at 37°C was shown to quantifiably improve recovery by about 5%. This purification was also tested practically, and

was used to purify ancient protein prior to mass spectral analysis.

Some larger questions that emerge from this study are worth further exploration. For one, more extensive work to obtain genetic sequences from more bison individuals from this region would be beneficial to further resolve the nature of ancient bison relationship. More sequencing work can be done with the 12 bison samples in this study, with the potential to obtain individuals from other sites in the future.

The proteomic techniques used here still have room for optimization. In particular, the database searching presents the area most likely in need of improvement. This can come in two ways. A better understanding of the GPM and how best to pull information from it would be beneficial, as the multitude of search functions present both the opportunity for great information and the chance for error and clouding of information. Other database search options (e.g. Mascot) can be explored as well. The other option would be to explore *de novo* peptide sequencing. This would overcome the limitations of database searching, as there would be no need for a comparable individual to exist within that database. The *de novo* sequences could then be used in a BLAST search to reveal protein (and ideally species or individual) identity and relationship.

## 6. Works Cited

- Abbaszadegan, M., Huber, M. S., Gerba, C. P., & Pepper, I. L. (1993). Detection of enteroviruses in groundwater with the polymerase chain reaction. *Applied and Environmental Microbiology*, 59(5), 1318–1324.
- Abramsky, O., & London, Y. (1975). Purification and partial characterization of two basic proteins from human peripheral nerve. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 393(2), 556–562.
- Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E., & Kent, S. B. (1987). Internal amino acid sequence analysis of proteins separated by one-or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proceedings of the National Academy of Sciences*, 84(20), 6970–6974.
- Aggarwal, S., & Yadav, A. K. (2016). False discovery rate estimation in proteomics. *Statistical Analysis in Proteomics*, 119–128.
- Agogino, G., & Frankforter, W. (1960). A Paleo-Indian Bison-Kill in Northwestern Iowa. *American Antiquity*, 25(3), 414-415.
- Ahn, S. J., Costa, J., & Rettig Emanuel, J. (1996). PicoGreen quantitation of DNA: effective evaluation of samples pre-or post-PCR. *Nucleic Acids Research*, 24(13), 2623–2625.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410.
- Aston, F. W. (1919). The constitution of the elements. *Nature*, 104(393), 2616.
- Baron, H., Hummel, S., & Herrmann, B. (1996). Mycobacterium tuberculosis complex DNA in ancient human bones. *Journal of Archaeological Science*, 23(5), 667–671.

- Barth, H. G., Jackson, C., & Boyes, B. E. (1994). Size exclusion chromatography. *Analytical Chemistry*, 66(12), 595R–620R.
- Bollongino, R., Tresset, A., & Vigne, J.-D. (2008). Environment and excavation: Pre-lab impacts on ancient DNA analyses. *Comptes Rendus Palevol*, 7(2), 91–98.
- Brown, W. E., & Wold, F. (1973). Alkyl isocyanates as active-site-specific reagents for serine protease. Identification of the active-site serine as the site of reaction. *Biochemistry*, 12(5), 835–840
- Brodsky, Barbara. (2014). Collagen. In *AccessScience*. McGraw-Hill Education. <https://doi.org/10.1036/1097-8542.148600>.
- Buehler, M. J. (2006). Nature designs tough collagen: explaining the nanostructure of collagen fibrils. *Proceedings of the National Academy of Sciences*, 103(33), 12285–12290.
- Butler, M. F., & Heppenstall-Butler, M. (2001). Phase separation in gelatin/maltodextrin and gelatin/maltodextrin/gum arabic mixtures studied using small-angle light scattering, turbidity, and microscopy. *Biomacromolecules*, 2(3), 812–823.
- Cappellini, E., Jensen, L. J., Szklarczyk, D., Ginolhac, A., da Fonseca, R. A. R., Stafford Jr, T. W., ... Willerslev, E. (2011). Proteomic analysis of a pleistocene mammoth femur reveals more than one hundred ancient bone proteins. *Journal of Proteome Research*, 11(2), 917–926.
- Chuprina, V. P., Heinemann, U., Nurislamov, A. A., Zielenkiewicz, P., Dickerson, R. E., & Saenger, W. (1991). Molecular dynamics simulation of the hydration shell of a B-DNA decamer reveals two main types of minor-groove hydration depending on groove width. *Proceedings of the National Academy of Sciences*, 88(2), 593–597.

- Cohen, A. S., Najarian, D. R., & Karger, B. L. (1990). Separation and analysis of DNA sequence reaction products by capillary gel electrophoresis. *Journal of Chromatography A*, 516(1), 49–60.
- Cohen, A. S., Najarian, D. R., Paulus, A., Guttman, A., Smith, J. A., & Karger, B. L. (1988). Rapid separation and purification of oligonucleotides by high-performance capillary gel electrophoresis. *Proceedings of the National Academy of Sciences*, 85(24), 9660–9663.
- Collins, M., Buckley, M., Grundy, H. H., Thomas-Oates, J., Wilson, J., & van Doorn, N. (2010). ZooMS: the collagen barcode and fingerprints. *Spectroscopy Europe*, 22, 2.
- Cooper, C. L. (1937). *Bison Occidentalis* at Interstate Park, Wisconsin. *Proceedings of the Geological Society of America*, 368
- Cowan, P. M., McGavin, S., & North, A. C. T. (1955). The polypeptide chain configuration of collagen. *Nature*, 176(4492), 1062–1064.
- Craig, R., & Beavis, R. C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry*, 17(20), 2310–2316.
- Desjardins, P., & Conklin, D. (2010). NanoDrop microvolume quantitation of nucleic acids. *Journal of Visualized Experiments: JoVE*, (45).
- Di Lullo, G. A., Sweeney, S. M., Körkkö, J., Ala-Kokko, L., & San Antonio, J. D. (2002). Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *Journal of Biological Chemistry*, 277(6), 4223–4231.
- Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29(31), 7133–7155.

- Eriksson, J., & Fenyő, D. (2004). Probit: a protein identification algorithm with accurate assignment of the statistical significance of the results. *Journal of Proteome Research*, 3(1), 32–36.
- Franca, L. T. C., Carrilho, E., & Kist, T. B. L. (2002). A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics*, 35(2), 169–200.
- Frankforter, W. D., & Agogino, G. A. (1960). THE SIMONSEN SITE: REPORT FOR THE SUMMER OF 1959. *Plains Anthropologist*, 5(10), 65–70.
- Gale, M. R., Grigal, D. F., & Harding, R. B. (1991). Soil productivity index: predictions of site quality for white spruce plantations. *Soil Science Society of America Journal*, 55(6), 1701–1708.
- Getz, E. B., Xiao, M., Chakrabarty, T., Cooke, R., & Selvin, P. R. (1999). A comparison between the sulfhydryl reductants tris (2-carboxyethyl) phosphine and dithiothreitol for use in protein biochemistry. *Analytical Biochemistry*, 273(1), 73–80.
- Graves, P. R., & Haystead, T. A. J. (2002). Molecular biologist's guide to proteomics. *Microbiology and Molecular Biology Reviews*, 66(1), 39–63.
- Grecz, N., Hammer, T. L., Robnett, C. J., & Long, M. D. (1980). Freeze-thaw injury: Evidence for Double strand breaks in *Escherichiacoli* DNA. *Biochemical and Biophysical Research Communications*, 93(4), 1110–1113.
- Guinot, P., Rogé, A., Gargadennec, A., Garcia, M., Dupont, D., Lecoœur, E., ... Andary, C. (2006). Dyeing plants screening: an approach to combine past heritage and present development. *Coloration Technology*, 122(2), 93–101.



- Gustavsson, S. Å., Samskog, J., Markides, K. E., & Långström, B. (2001). Studies of signal suppression in liquid chromatography–electrospray ionization mass spectrometry using volatile ion-pairing reagents. *Journal of Chromatography A*, 937(1), 41–47.
- Hagelberg, E. (1994). Mitochondrial DNA from ancient bones. In *Ancient DNA* (pp. 195–204). Springer.
- Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In *Nucleic acids symposium series* (Vol. 41, pp. 95–98). [London]: Information Retrieval Ltd., c1979-c2000.
- Han, X., Aslanian, A., & Yates, J. R. (2008). Mass spectrometry for proteomics. *Current Opinion in Chemical Biology*, 12(5), 483–490.
- Han, X., Jin, M., Breuker, K., & McLafferty, F. W. (2006). Extending top-down mass spectrometry to proteins with masses greater than 200 kilodaltons. *Science*, 314(5796), 109–112.
- Hänni, C., Laudet, V., Stehelin, D., & Taberlet, P. (1994). Tracking the origins of the cave bear (*Ursus spelaeus*) by mitochondrial DNA sequencing. *Proceedings of the National Academy of Sciences*, 91(25), 12336-12340.
- Heintzman, P. D., Froese, D., Ives, J. W., Soares, A. E. R., Zazula, G. D., Letts, B., ... Hare, P. G. (2016). Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada. *Proceedings of the National Academy of Sciences*, 113(29), 8057–8063.
- Henle, E. S., & Linn, S. (1997). Formation, prevention, and repair of DNA damage by iron/hydrogen peroxide. *Journal of Biological Chemistry*, 272(31), 19095–19098.

- Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. *Nature*, *312*(5991), 282.
- Höss, M., Dilling, A., Currant, A., & Pääbo, S. (1996). Molecular phylogeny of the extinct ground sloth *Mylodon darwini*. *Proceedings of the National Academy of Sciences*, *93*(1), 181-185.
- Jenks, Albert E. (1937) A Minnesota Kitchen Midden with Fossil Bison. *Science* *86*:243–244.
- Johnson, P. H., Olson, C. B., & Goodman, M. (1985). Isolation and characterization of deoxyribonucleic acid from tissue of the woolly mammoth, *Mammuthus primigenius*. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, *81*(4), 1045-1051.
- Jonscher, K. R., & Yates, J. R. (1997). The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Analytical Biochemistry*, *244*(1), 1–15.
- Karger, B. L. (1997). HPLC: Early and recent perspectives. *J. Chem. Educ.*, *74*(1), 45.
- Karlin, S., & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, *87*(6), 2264–2268.
- Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, *14*, 1–63.
- Killelea, T., Ralec, C., Bossé, A., & Henneke, G. (2014). PCR performance of a thermostable heterodimeric archaeal DNA polymerase. *Frontiers in Microbiology*, *5*.
- Köster, C. (2015). Twin trap or hyphenation of a 3D Paul-and a Cassinian ion trap. *Journal of The American Society for Mass Spectrometry*, *26*(3), 390-396.

- Krokhin, O. V, Antonovici, M., Ens, W., Wilkins, J. A., & Standing, K. G. (2006). Deamidation of-Asn-Gly-sequences during sample preparation for proteomics: Consequences for MALDI and HPLC-MALDI analysis. *Analytical Chemistry*, 78(18), 6645–6650.
- Kunitz, M., & Northrop, J. H. (1934). Inactivation of crystalline trypsin. *The Journal of General Physiology*, 17(4), 591–615.
- Liang, S. B., Hu, D. P., Zhu, C., & Yu, A. B. (2002). Production of fine polymer powder under cryogenic conditions. *Chemical Engineering & Technology*, 25(4), 401–405.
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362(6422), 709–715.
- Madden, T. (2013). The BLAST Sequence Analysis Tool. 2<sup>nd</sup> edition, Internet Access. *The National Centre of Biotechnology Information*. Bethesda, MD.
- Matheson, C. D., Gurney, C., Esau, N., & Lehto, R. (2010). Assessing PCR inhibition from humic substances. *The Open Enzyme Inhibition Journal*, 3(1).
- Matheson, C. D., Marion, T. E., Hayter, S., Esau, N., Fratpietro, R., & Vernon, K. K. (2009). Removal of metal ion inhibition encountered during DNA extraction and amplification of copper-preserved archaeological bone using size exclusion chromatography. *American Journal of Physical Anthropology*, 140(2), 384–391.
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74(2), 560–564.
- McAndrews, J. H. (1982). Holocene environment of a fossil bison from Kenora, Ontario. *Ontario Archaeology*, 37, 41-51.

- McDonald, W. H., & Yates 3rd, J. R. (2003). Shotgun proteomics: integrating technologies to answer biological questions. *Current Opinion in Molecular Therapeutics*, 5(3), 302–309.
- McLafferty, F. W., Breuker, K., Jin, M., Han, X., Infusini, G., Jiang, H., ... Begley, T. P. (2007). Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. *The FEBS Journal*, 274(24), 6256–6268.
- McPherson, M., & Møller, S. (2000). *Pcr*. Taylor & Francis.
- Melton, Terry. (2003). Forensic mitochondrial DNA analysis. In *AccessScience*. McGraw-Hill Education.
- Meselson, M., & Stahl, F. W. (1958). The replication of DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 44(7), 671–682.
- Montier, L. L. C., Deng, J. J., & Bai, Y. (2009). Number matters: control of mammalian mitochondrial DNA copy number. *Journal of Genetics and Genomics*, 36(3), 125–131.
- Moretti, T., Koons, B., & Budowle, B. (1998). Enhancement of PCR Amplification Yield and Specificity Using AmpliTaq Gold [TM] DNA Polymerase. *Biotechniques*, 25(4), 716–723.
- Mullay, J. (1987). Estimation of atomic and group electronegativities. In *Electronegativity* (pp. 1–25). Springer.
- Nielsen-Marsh, C., Gernaey, A., Turner-Walker, G., Hedges, R., Pike, A. W. G., & Collins, M. (2000). The chemical degradation of bone.
- Nyrén, P., & Lundin, A. (1985). Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. *Analytical Biochemistry*, 151(2), 504–509.
- O'Farrell, P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *Journal of Biological Chemistry*, 250(10), 4007–4021.

- Opel, K. L., Chung, D. T., Drábek, J., Tatarek, N. E., Jantz, L. M., & McCord, B. R. (2006). The application of miniplex primer sets in the analysis of degraded DNA from human skeletal remains. *Journal of Forensic Sciences*, 51(2), 351–356.
- Pääbo, S. (1984). Molecular cloning of Egyptian DNA. *Nature*, 314, 644-645.
- Pääbo, S. (2014). The human condition—a molecular approach. *Cell*, 157(1), 216–226.
- Pääbo, S. (1985). Molecular cloning of ancient Egyptian mummy DNA. *Nature*, 314(6012), 644–645.
- Pääbo, S., Gifford, J. A., & Wilson, A. C. (1988). Mitochondrial DNA sequences from a 7000-year old brain. *Nucleic Acids Research*, 16(20), 9775–9787.
- Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., ... Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annu. Rev. Genet.*, 38, 645–679.
- Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., ... Scholtz, J. M. (2011). Contribution of hydrophobic interactions to protein stability. *Journal of Molecular Biology*, 408(3), 514–528.
- Pace, C. N., Shirley, B. A., McNutt, M., & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *The FASEB Journal*, 10(1), 75–83.
- Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., ... & Pääbo, S. (1998). Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. *Science*, 281(5375), 402-406.
- Poinar, H. N., & Stankiewicz, B. A. (1999). Protein preservation and DNA retrieval from ancient tissues. *Proceedings of the National Academy of Sciences*, 96(15), 8426–8431.

- Pruvost, M., Schwarz, R., Correia, V. B., Champlot, S., Braguier, S., Morel, N., ... & Geigl, E. M. (2007). Freshly excavated fossil bones are best for amplification of ancient DNA. *Proceedings of the National Academy of Sciences*, 104(3), 739-744
- Rawlings, N. D., & Barrett, A. J. (1994). [2] Families of serine peptidases. *Methods in Enzymology*, 244, 19–61.
- Rich, A., & Crick, F. (1955). *The structure of collagen*. Macmillan Journals Limited.
- Riede, U. N., Jonas, I., Kirn, B., Usener, U. H., Kreutz, W., & Schlickewey, W. (1992). Collagen stabilization induced by natural humic substances. *Archives of Orthopaedic and Trauma Surgery*, 111(5), 259–264.
- Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, 242(1), 84–89.
- Rosenfeld, J., Capdevielle, J., Guillemot, J. C., & Ferrara, P. (1992). In-gel digestion of proteins for internal sequence analysis after one-or two-dimensional gel electrophoresis. *Analytical Biochemistry*, 203(1), 173–179.
- Ross, K. S., Haites, N. E., & Kelly, K. F. (1990). Repeated freezing and thawing of peripheral blood and DNA in suspension: effects on DNA yield and integrity. *Journal of Medical Genetics*, 27(9), 569–570.
- Ruiz-Martinez, M. C., Berka, J., Belenkii, A., Foret, F., Miller, A. W., & Karger, B. L. (1993). DNA sequencing by capillary electrophoresis with replaceable linear polyacrylamide and laser-induced fluorescence detection. *Analytical Chemistry*, 65(20), 2851–2858.

- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., ... Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839), 487–491.
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), 441–448.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463–5467.
- Sechi, S., & Chait, B. T. (1998). Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Analytical Chemistry*, 70(24), 5150–5158.
- Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., ... Binladen, J. (2004). Rise and fall of the Beringian steppe bison. *Science*, 306(5701), 1561–1565.
- Shay, C. T. (1971). *Itasca Bison Kill Site: An Ecological Analysis*.
- Shoulders, M. D., & Raines, R. T. (2009). Collagen structure and stability. *Annual Review of Biochemistry*, 78, 929–958.
- Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., ... Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071), 674–679.
- Snyder, P. S., & Thomas, J. F. (1968). Solute activity coefficients at infinite dilution via gas-liquid chromatography. *Journal of Chemical & Engineering Data*, 13(4), 527–529.

- Söderhäll, C., Marenholz, I., Kerscher, T., Rüschemdorf, F., Esparza-Gordillo, J., Worm, M., ... Rohde, K. (2007). Variants in a novel epidermal collagen gene (COL29A1) are associated with atopic dermatitis. *PLoS Biology*, 5(9), e242.
- Stewart, J. R. M., Allen, R. B., Jones, A. K. G., Penkman, K. E. H., & Collins, M. J. (2013). ZooMS: making eggshell visible in the archaeological record. *Journal of Archaeological Science*, 40(4), 1797–1804.
- Stoneking, M. (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *The American Journal of Human Genetics*, 67(4), 1029–1032.
- Stoscheck, C. M. (1990). [6] Quantitation of protein. *Methods in Enzymology*, 182, 50–68.
- Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., & Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26), 9528–9533.
- Tamura, K., Nei, M., & Kumar, S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America*, 101(30), 11030–11035.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30(12), 2725–2729.
- Tuross, N., Fogel, M. L., & Hare, P. E. (1988). Variability in the preservation of the isotopic composition of collagen from fossil bone. *Geochimica et Cosmochimica Acta*, 52(4), 929–935.



- Vellaichamy, A., Lin, C. Y., Aye, T. T., Kunde, G. R., & Nesvizhskii, A. I. (2010). A Chloroform-Assisted Protein Isolation Method Followed by Capillary NanoLC-MS Identify Estrogen-Regulated Proteins from MCF7 Cells. *J Proteomics Bioinform*, 3, 212–220.
- Wadsworth, C., & Buckley, M. (2014). Proteome degradation in fossils: investigating the longevity of protein survival in ancient bone. *Rapid Communications in Mass Spectrometry*, 28(6), 605–615.
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356), 737–738.
- Webb, S. D., Milanich, J. T., Alexon, R., & Dunbar, J. S. (1984). A Bison antiquus kill site, Wacissa River, Jefferson County, Florida. *American Antiquity*, 49(2), 384-392.
- Wilson, A. M., Work, T. M., Bushway, A. A., & Bushway, R. J. (1981). HPLC determination of fructose, glucose, and sucrose in potatoes. *Journal of Food Science*, 46(1), 300–301.
- Wilson, I. G. (1997). Inhibition and facilitation of nucleic acid amplification. *Applied and Environmental Microbiology*, 63(10), 3741.
- Wilson, M. C., Hills, L. V., & Shapiro, B. (2008). Late Pleistocene northward-dispersing Bison antiquus from the Bighill Creek Formation, Gallelli gravel pit, Alberta, Canada, and the fate of Bison occidentalis. *Canadian Journal of Earth Sciences*, 45(7), 827–859.
- Wold, Marc S., Rich, Alexander, Weeks, Daniel, and Lutter, Leonard C. (2016). Deoxyribonucleic acid (DNA). In *AccessScience*. McGraw-Hill Education
- Wolff, S. P., & Dean, R. T. (1986). Fragmentation of proteins by free radicals and its effect on their susceptibility to enzymic hydrolysis. *Biochemical Journal*, 234(2), 399–403.

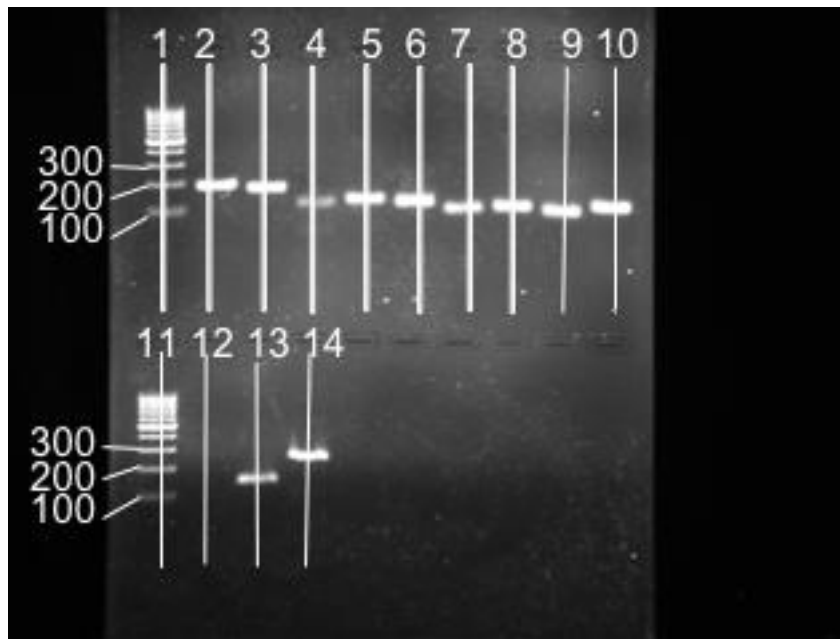
- Yang, H., Golenberg, E. M., & Shoshani, J. (1997). Proboscidean DNA from museum and fossil specimens: an assessment of ancient DNA extraction and amplification techniques. *Biochemical Genetics*, 35(5-6), 165-179.
- Yang, L., Arora, K., Beard, W. A., Wilson, S. H., & Schlick, T. (2004). Critical role of magnesium ions in DNA polymerase  $\beta$ 's closing and active site assembly. *Journal of the American Chemical Society*, 126(27), 8441–8453.
- Yates III, J. R. (2004). Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.*, 33, 297–316.
- Zubarev, R. A., Kelleher, N. L., & McLafferty, F. W. (1998). Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American Chemical Society*, 120(13), 3265–3266.

# Appendix

**Table 12:** Primers used in this study. 1st forward primer is paired with first reverse primer (i.e. E01 with F01, and so on). Primer design originally by Shapiro et al. (2004).

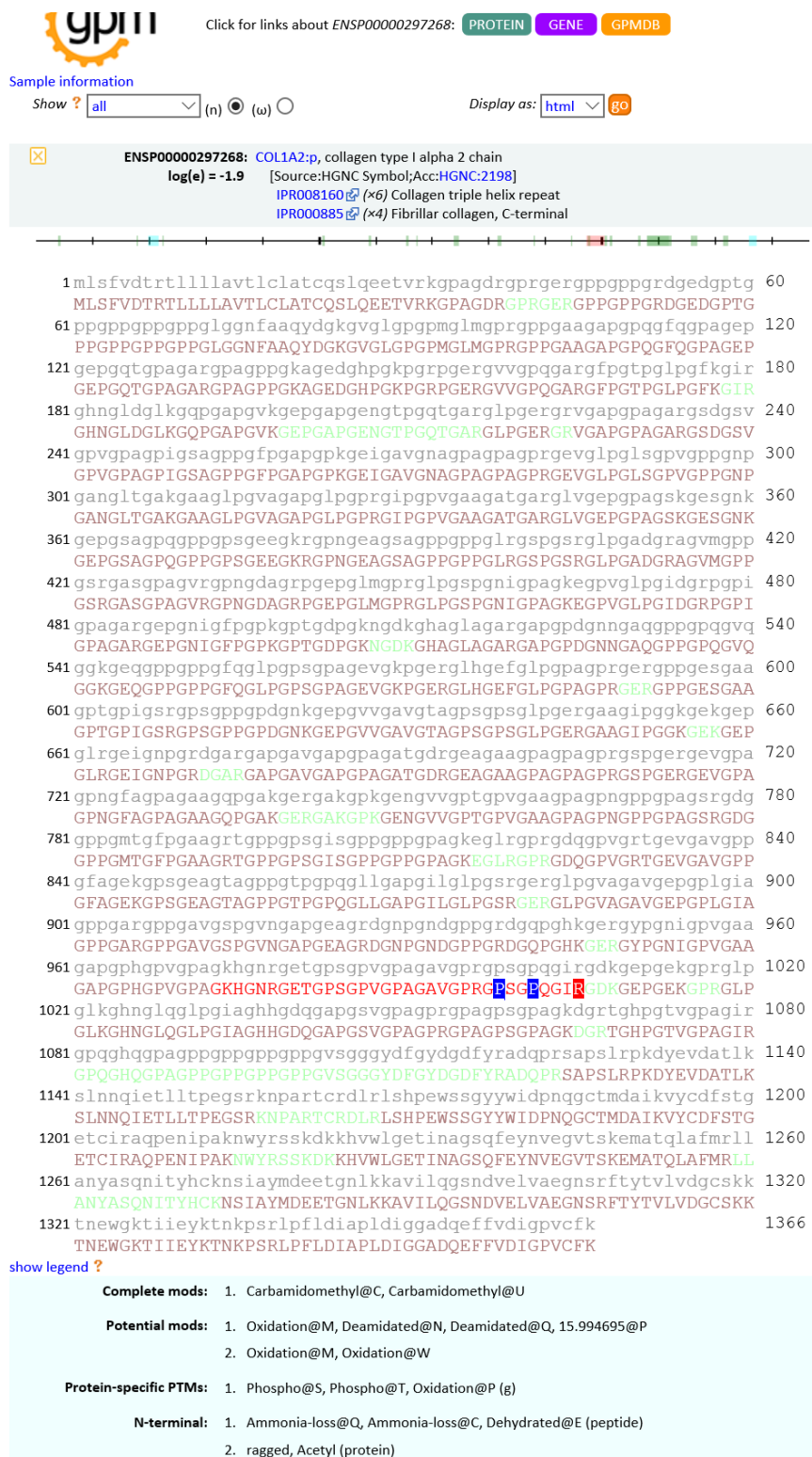
	Name	Sequence	Tm°C	CG%	Length (bp)
Forward Primers	E01	ctacagtctcaccgtcaaccc	64.0	57.1	21
	E02	acccccaagctgaagtct	64.2	50.0	20
	E03	ccataaatgcaaagagcctcac	64.9	45.5	22
	E04	actagctaacgtcactcacccc	63.7	54.5	22
	E05	ccactagctaacgtcactcacccc	68.9	58.3	24
	E06	gccccatgcatataagcaag	64.5	50.0	20
	E07	cgtaacatagcacattatgtc	53.6	40.0	20
	E08	gcacattatgtcaaactaccctgacaac	69.4	40.0	30
	E09	cacgagcttaactaccatgc	60.4	50.0	20
	E10	gagcttaaytaccatgccg	60.3	50.0	19
	E11	ggggtgagtgacgttagctagt	63.7	54.5	22
	E12	taagtacttgcttatatgcatggggc	66.6	42.3	26
Reverse Primers	F01	ggggtgagtgacgttagctagtg	66.1	56.5	23
	F02	cttgcttatatgcatggggc	64.5	50.0	20
	F03	gcatggggcatataaatctaattac	63.3	40.0	25
	F04	gattgacataatgtgctatg	55.2	33.3	21
	F05	caagggtagatttgacataatgtg	61.5	37.5	24
	F06	gcctagcgggttgctggtttcacgc	78.5	64.0	25
	F07	gatgtctgataaagttcattaaatagcgacccc	71.0	39.4	33
	F08	ccagatgtctgataaagtcca	57.3	38.1	21
	F09	gatgagatggccctgaagaa	64.5	50.0	20
	F10	ccaaatgtatgacagcacag	59.7	45.0	20

**Table 13:** Summary gel electrophoresis of amplified modern control (Figure 16)



**Figure 16:** Amplified product of a modern bison extract visualized on a 2% agarose gel made in TBE buffer with 2 $\mu$ L ethidium bromide following the protocol in Tables 3 and 4. Image taken under UV light with a Universal Hood II (BioRad) and QuantityOne software. Lane identities are summarized in Table 13

Lane #	Identity	Lane #	Identity
1	GeneRuler 100bp (Fermentas)	9	GeneRuler 100bp (Fermentas)
2	Primer Set (PS) 1	10	PS8
3	PS2	11	PS9
4	PS3	12	PS10
5	PS4	13	PS11
6	PS5	14	PS12
7	PS6		
8	PS7		



**Figure 17:** Amino acid alignment from the top-scoring results from Figure 9. Sequenced ISM11 protein (red capital letters) against a reference sequences from the GPMDB (lower case). Detected amino sequences are in red text, modified bases are highlighted in blue, polymorphic sites are highlighted in red, and sections believed to be difficult to sequence are in green text.



**Figure 18:** Graphical view of the top scoring alignments from a Basic Local Alignments Search Tool (BLAST) protein search of the amino acid sequence from ISM11 in Figure 17.

Sequences producing significant alignments:

Select: [All](#) [None](#) Selected: 0

Description	Max score	Total score	Query cover	E value	Ident	Accession
<input type="checkbox"/> type I collagen [Homo sapiens]	63.9	63.9	100%	7e-11	100%	<a href="#">CAA39142.1</a>
<input type="checkbox"/> collagen 1 pro-alpha-2 chain [Homo sapiens]	58.2	58.2	100%	2e-10	97%	<a href="#">AAA51996.1</a>
<input type="checkbox"/> unnamed protein product [Homo sapiens]	61.6	61.6	100%	3e-10	100%	<a href="#">CAA23761.1</a>
<input type="checkbox"/> PREDICTED: LOW QUALITY PROTEIN: collagen alpha-2(I) chain [Rhinopithecus bieti]	61.2	61.2	100%	5e-10	100%	<a href="#">XP_017710454.1</a>
<input type="checkbox"/> PREDICTED: collagen alpha-2(I) chain [Gorilla gorilla gorilla]	61.2	61.2	100%	5e-10	100%	<a href="#">XP_004045817.1</a>
<input type="checkbox"/> PREDICTED: collagen alpha-2(I) chain [Pan paniscus]	61.2	61.2	100%	5e-10	100%	<a href="#">XP_003809763.1</a>
<input type="checkbox"/> hypothetical protein EGM_12766 [Macaca fascicularis]	61.2	61.2	100%	5e-10	100%	<a href="#">EHH52338.1</a>
<input type="checkbox"/> hypothetical protein EGK_13930 [Macaca mulatta]	61.2	61.2	100%	5e-10	100%	<a href="#">EHH17509.1</a>
<input type="checkbox"/> collagen alpha-2(I) chain [Papio anubis]	61.2	61.2	100%	5e-10	100%	<a href="#">XP_003896357.1</a>
<input type="checkbox"/> PREDICTED: collagen alpha-2(I) chain isoform X1 [Colobus angolensis palliatus]	61.2	61.2	100%	5e-10	100%	<a href="#">XP_011810470.1</a>

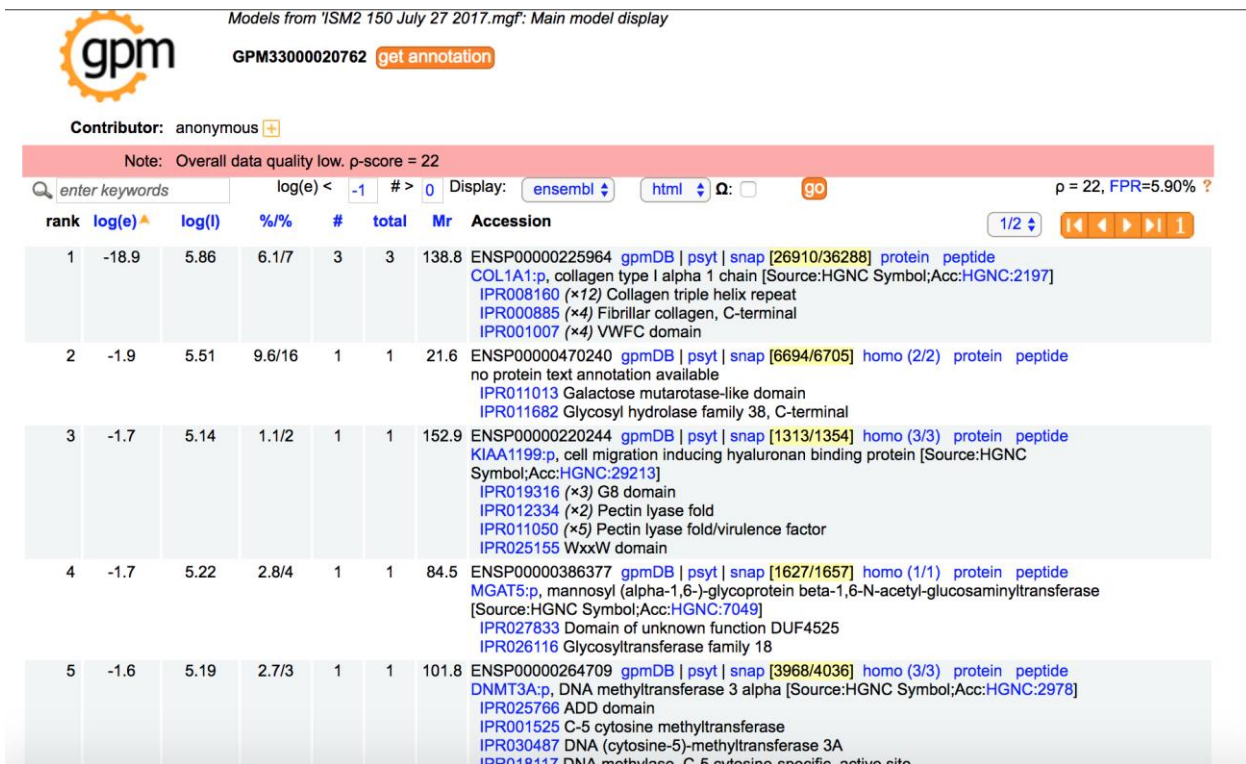
**Figure 19:** Top ten alignment identities from a Basic Local Alignments Search Tool (BLAST) protein search of the sequence from ISM11 found in figure 10. Alignment scores correspond to the graphic view in Figure 18.

Score	Expect	Method	Identities	Positives	Gaps
63.9 bits(154)	7e-11	Compositional matrix adjust.	33/33(100%)	33/33(100%)	0/33(0%)

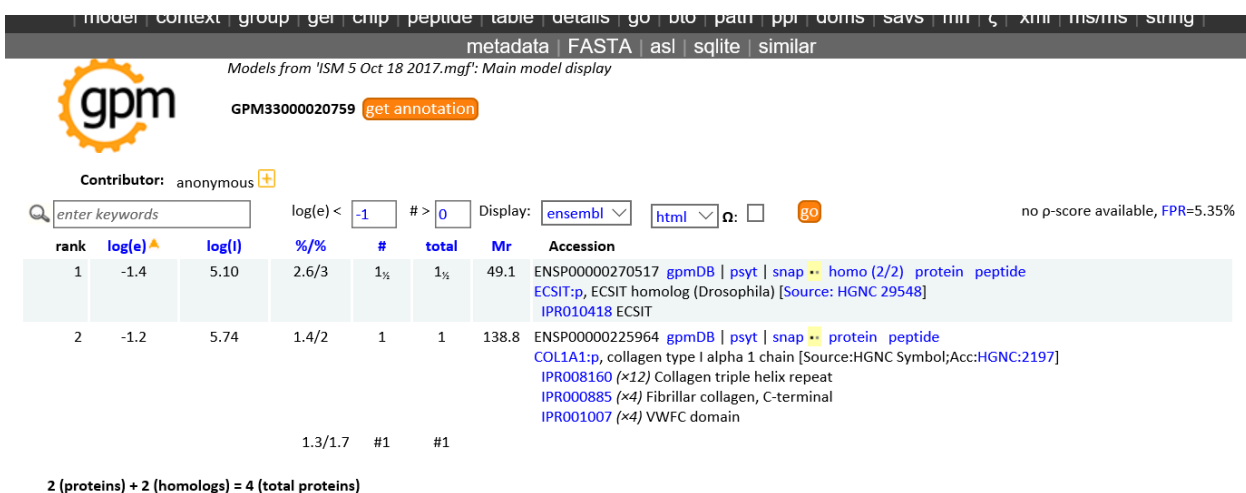
```

Query 1  GKHGNGRGTGSPGVPAGAVGPRGSPGQIR 33
          GKHGNGRGTGSPGVPAGAVGPRGSPGQIR
Sbjct 14 GKHGNGRGTGSPGVPAGAVGPRGSPGQIR 46
  
```

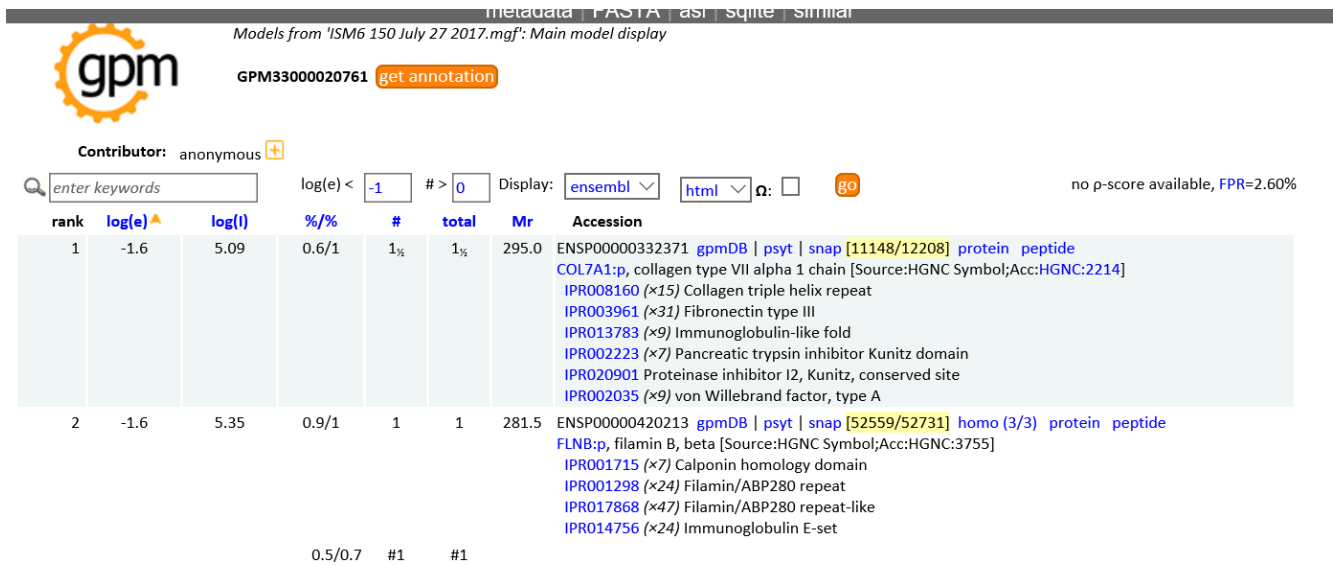
**Figure 20:** Alignment view of the top alignment from the Basic Local Alignments Search Tool (BLAST) protein search of the sequence from ISM11 found in Figure 18.



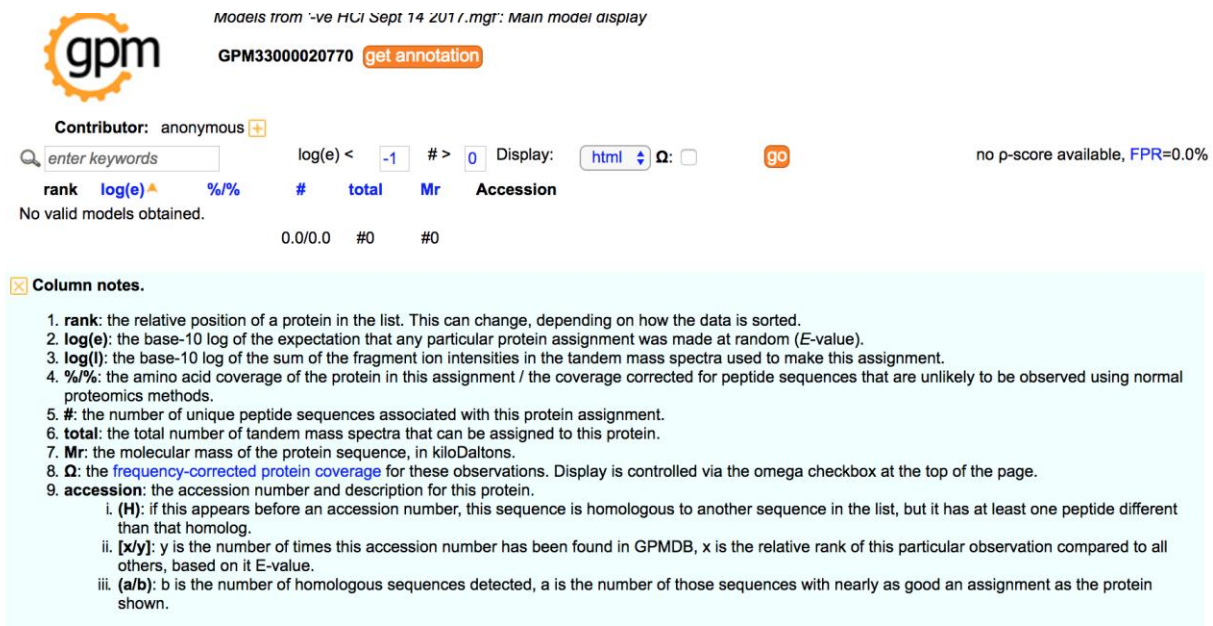
**Figure 21:** Top 5 Global Proteome Machine search results from sample ISM2 following Tandem Mass Spectrometry using the X! P3 algorithm against a *Homo sapiens* database with complete modifications for carbamidomethylation at C and U, potential modifications for oxidation at M, deamidation at N and Q, modified hydroxyproline, and oxidation at M and W, with semi-style cleavage with trypsin and Ion Trap detection. 33 protein matches were identified in total from this individual.



**Figure 22:** Global Proteome Machine search results from sample ISM5 following Tandem Mass Spectrometry using the X! P3 algorithm against a *Homo sapiens* database with complete modifications for carbamidomethylation at C and U, potential modifications for oxidation at M, deamidation at N and Q, modified hydroxyproline, and oxidation at M and W, with semi-style cleavage with trypsin and Ion Trap detection.



**Figure 23:** Global Proteome Machine search results from sample ISM6 following Tandem Mass Spectrometry using the X! P3 algorithm against a *Homo sapiens* database with complete modifications for carbamidomethylation at C and U, potential modifications for oxidation at M, deamidation at N and Q, modified hydroxyproline, and oxidation at M and W, with semi-style cleavage with trypsin and Ion Trap detection.



**Figure 24:** GPM search results from the negative experimental control following Tandem Mass Spectrometry using the X! P3 algorithm against a *Homo sapiens* database with complete modifications for carbamidomethylation at C and U, potential modifications for oxidation at M, deamidation at N and Q, modified hydroxyproline, and oxidation at M and W, with semi-style cleavage with trypsin and Ion Trap detection. 33 protein matches were identified in total from this individual











AY748784.1 Bison priscus isolate BS693 control region partial sequence mitochondrial  
AY748782.1 Bison priscus isolate BS691 control region partial sequence mitochondrial  
AY748773.1 Bison priscus isolate BS672 control region partial sequence mitochondrial  
AY748762.1 Bison priscus isolate BS655 control region partial sequence mitochondrial  
AY748760.1 Bison priscus isolate BS609 control region partial sequence mitochondrial  
AY748753.1 Bison priscus isolate BS582 control region partial sequence mitochondrial  
AY748749.1 Bison priscus isolate BS571 control region partial sequence mitochondrial  
AY748735.1 Bison priscus isolate BS533 control region partial sequence mitochondrial  
AY748698.1 Bison priscus isolate BS459 control region partial sequence mitochondrial  
AY748666.1 Bison priscus isolate BS413 control region partial sequence mitochondrial  
AY748623.1 Bison priscus isolate BS353 control region partial sequence mitochondrial  
AY748576.1 Bison priscus isolate BS284 control region partial sequence mitochondrial  
AY748574.1 Bison priscus isolate BS282 control region partial sequence mitochondrial  
AY748539.1 Bison priscus isolate BS212 control region partial sequence mitochondrial  
AY748506.1 Bison priscus isolate BS150 control region partial sequence mitochondrial

0.0050

**Figure 25:** The evolutionary history was inferred using the Neighbour-Joining method (Saitou et al. 1987). The optimal tree with the sum of branch length = 0.09598157 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura et al. 2004) and are in the units of the number of base substitutions per site. The analysis involved 326 nucleotide sequences: 325 from Shapiro et al. (2004) and one produced in this project (highlighted in yellow). All positions containing gaps and missing data were eliminated. There were a total of 80 positions in the final dataset. Evolutionary analysis were conducted in Mega6 (Tamura et al. 2013).