A Metagenomic Analysis of Ancient DNA at the Holocene Epoch

A METAGENOMIC ANALYSIS OF ANCIENT SEDIMENTARY DNA ACROSS THE PLEISTOCENE-HOLOCENE TRANSITION

BY

TARA ROSE SADOWAY, B.SC., HONS.

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOLOGY
AND THE SCHOOL OF GRADUATE STUDIES
OF MCMASTER UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

© Copyright by Tara Rose Sadoway, April 2014
All Rights Reserved

Master of Science (2014)	McMaster University
(Biology)	Hamilton, Ontario, Canada

TITLE:	A Metagenomic Analysis of Ancient Sedimentary DNA Across the Pleistocene-Holocene Transition
AUTHOR:	Tara Rose Sadoway B.Sc., Hons. (Cellular, Molecular, & Microbial Biology) Thompson Rivers University, Kamloops, B.C., Canada
SUPERVISORS:	Dr. G. B. Golding & Dr. H. N. Poinar
NUMBER OF PAGES:	x, 66

Abstract

Ancient DNA has the power to elucidate ecological and evolutionary relationships that were previously only quantifiable by proxy. This work details both a metagenetic and a targeted metagenomic study of ancient sedimentary DNA. By using DNA to investigate the plants and animals present in twelve different time points, we describe the nature of the ecological change over the Pleistocene-Holocene transition. We show that as the stability of the habitat degraded due to climate change, the dominant plant communities exhibited a shift from functional groups such as forbs to shrubs and trees. As this cascading change consequently affected the animal communities, we demonstrate the decline, extinction, and replacement of a variety of megafaunal species and mammoths. As well, we provide a proof-of-concept for the targeted oligonucleotide enrichment for ancient sedimentary DNA. By processing the same DNA extracts with targeted enrichment, we show that metagenomic soil DNA can provide the same taxonomic fingerprint unique to each time period even using different genetic loci. This unique pattern can be used as a reference in future studies. Although the oligonucleotide baits did not yield the composition of taxa that we expected, the oligonucleotide baits did increase the eukaryotic fraction of DNA extracts by up to 50%. Overall, this technique is open to further study and has fantastic potential to redefine the metagenomic work of ancient DNA soil cores.

Acknowledgements

I would like to thank my supervisors Dr. G. B. Golding and Dr. H. N. Poinar for their support and for allowing me to work on such an interesting project. As well, I would like to thank W. Sung, H. Ghenu, T. Porter, M. Kuch, and many other colleagues for their supportive comments and thoughtful conversations. A special thank-you to all my collaborators at the University of Alberta (Dr. Froese and grad students) and in the Yukon who allowed these permafrost samples to be collected and studied. I dedicate this thesis to my amazing family, to Sean, and to Kitty.

Contents

Abstract iii

Acknowledgements iv

Contents v

List of Figures vii

List of Tables viii

Abbreviations ix

Declaration of Academic Achievement x

Introduction 1

I References 3

Chapter 1 5

I Introduction 6

II Methods 8
- i Sample Extraction and Amplification . . . 8
- ii Library Preparation, Indexing, and Data Pipeline . . . 9

III Results 10
- i The Floral Record . . . 10
- ii The Faunal Community . . . 13
- iii Metagenetic Profile . . . 16

IV Discussion 19

V References 21

VI Supplementary Materials 29

Chapter 2 **37**

I Introduction **38**

II Methods **39**
i Sample Extraction and Library Preparation 39
ii Oligonucleotide Bait Design 40
iii Enrichment and Sequencing 41
iv Data Analysis 42

III Results **45**
i Initial Quantitative Observations 45
ii Hybridization Temperature Patterns 49
iii Community Analysis 51
iv GC Content 55
v The Blanks and Controls 55

IV Discussion **55**

V References **59**

VI Supplementary Materials **63**

Conclusion **64**

I References **66**

List of Figures

Chapter 1 **5**

1 Relative functional group assignment of plants 12
2 Pollen quantification compared to DNA datasets 14
3 NMDS plot for plant loci 15
4 NMDS plot for vertebrate loci 17
5 Metagenetic NMDS plot for plant and animal loci 18
S1 Difference between BLAST and curated datasets results in the NMDS metagenetic analysis 36

Chapter 2 **37**

1 Raw sequencing reads correlate with the total number of assigned reads 48
2 Assigned taxonomic counts correlate with quantity of initial unique reads 49
3 Hybridization as a correlated factor for total reads or GC content . . 50
4 NMDS of enriched metagenomic samples 52
5 The GC content of enriched reads 56

List of Tables

Chapter 1 5

1 Relative abundance of plant taxonomic assignments 11
S1 DNA permafrost core samples 29
S2 Blanks and controls for ampliconic DNA samples 30
S3 Plant community sequence data by functional group 31
S4 Parameters and sample script for the bioinformatic pipeline. 31
S5 Curated vertebrate reference list - Part I 32
S6 Curated vertebrate reference list - Part II 33
S7 Primers for ampliconic loci amplification 35

Chapter 2 37

1 Theoretically efficacy of oligonucleotide bait sets 41
2 DNA enrichment sample hybridization conditions 43
3 Mis-binned sequences per sequencing direction 44
4 Enrichment sequence read counts. 46
5 Quantification of assigned taxonomy by rank 47
6 Taxonomic identification of enriched reads by domain 54
S1 Enriched DNA permafrost core samples 63

Abbreviations

aDNA - ancient DNA
AT - adenine thymine
baits - oligonucleotide baits
BP - Before Present
bp -base pairs
cal - calibrated
cox1 - cytochrome c oxidase subunit 1
cytb - cytochrome b
DNA - deoxyribonucleic acid
EB - elution buffer
GC - guanine cytosine
GenBank - genetic sequence database
GFP - green fluorescent protein
#k - number in thousands
PCR - polymerase chain reaction
PH - Pleistocene-Holocene
qPCR - quantitative PCR
rbcL - ribulose-bisphosphate carboxylase gene
RNA - ribonucleic acid
rRNA - ribosomal RNA
trnL - transfer RNA gene
UDG - uracil-DNA glycosylase
yrs - years

Declaration of Academic Achievement

The majority of the work on both papers was done by me. For both studies, I was not responsible for the field recovery of the first set of Lucky Lady cores in the Yukon and the homogenization of those samples at the University of Alberta. However, all post processing steps, including DNA extraction, amplification, library preparation, and the data analysis were solely my work. Dr. T. Porter performed the *in silico* oligonucleotide bait design and sampling. M. Kuch assisted with the initial round of enrichment in the 50/55°C hybridization experiment. I obtained the second set of Lucky Lady cores (samples LL1-3) through field work at the study site near Dawson city, Yukon in 2012. For both papers, I wrote my own Perl scripts and processed all the high-throughput data myself. I wrote the entire manuscript in LaTeX.

Introduction

This thesis investigates ancient permafrost DNA through the use of metagenetic and metagenomic techniques. The first study employs three vertebrate and two plant partial genetic loci to reconstruct the flora and fauna over the Holocene epoch. The second paper focuses on the new technique of whole soil metagenomic enrichment to selectively enrich a wide range of whole vertebrate genetic loci from a background of predominantly bacterial DNA.

Permafrost, a perennially frozen type of terrain, is a unique medium for the preservation of organic deposits, as permafrost soils exhibit very little leaching between layers and minimal soil cycling (ARNOLD *et al.*, 2011; HAILE *et al.*, 2007). Thus, the existing stratigraphic record allows for the examination of relatively discrete time points from tens to hundreds of thousands of years old. In this case, we were uniquely interested in the transition from the cold, tundra-steppe environment of the late Pleistocene to the warm, moist environment of the Holocene. Due to large variations in the climate before the Holocene, organisms that flourished before the transition were not necessarily the ones that survived or populated the landscape after the transition. We know from the fossil record that many large mammals died in the late Pleistocene and there is still debate on the cause of these extinctions (KOCH and BARNOSKY, 2006; TURVEY and FRITZ, 2011; GUTHRIE, 2001). From plant fossils, either as pollen or macrofossil remains, we know that at the time of the deposit of the Dawson tephra at 30k years before present, there was a robust community of grasses and sedges (ZAZULA *et al.*, 2006; VAN GEEL *et al.*, 2007). However, the fossil record representing plants and animals is subject to preservation and abundance biases, so the last fossil we find of an animal may not be the last member of that species in an area. Fortunately, modern sedimentary DNA has been used to demonstrate local diversity (STEVEN *et al.*, 2007; ANDERSEN *et al.*, 2012; YOCCOZ *et al.*, 2012), so we hypothesized the ability of ancient DNA to reflect the communities during the climate change of the PH transition and possibly extend our knowledge of extinct ecosystems beyond the fossil record.

We support our conclusions with two different, but complementary DNA analysis processes. In the first study, the amplicon data was generated in duplicate for each sample and combined to reduce the known ampliconic biases. We chose ampliconic loci that were previously used in modern contexts and also designed a new set of primers based on available reference sequences in GenBank. Theoretically, the greater the number of relevant reference sequences available, the more likely that we will be able to make a positive identification from our data. Of course, due to the differing ease of amplification from a variety of species, certain loci have a

bias towards specific classes of taxa. For example, cytochrome c oxidase subunit 1 (*cox1*) was chosen early on by ornithologists and is much more well-populated by bird taxa than other loci. For the metagenomic enrichment study, we designed oligonucleotide baits from two of the same genes as the amplicon study (12S rRNA and cytochrome b) as well as the barcoding gene *cox1*. For the enrichment, it was possible to cover the entire gene of the three loci, instead of only isolating a small (<150 base pair) amplicon as in the first study. Although we did not see the theoretical recovery predicted by *in silico* simulations, we did increase the eukaryotic composition of sample extracts by up to 50%.

We also considered the bias of sample extraction and paid close attention to sources of contamination. In both studies, there was strict adherence to established ancient DNA working guidelines and all samples were processed in a dedicated facility or clean room.

Both papers show that every set of sampled time points has a unique genetic fingerprint that is consistent between DNA analysis methods. Even though the studies worked with different genomic isolation methods, the overall picture of the Pleistocene-Holocene transition confirms and expands on the dominant changes in plant and animal communities predicted by the fossil record.

I References

ANDERSEN, K., K. L. BIRD, M. RASMUSSEN, J. HAILE, H. BREUNING-MADSEN, *et al.*, 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Molecular Ecology **21**: 1966–1979.

ARNOLD, L. J., R. G. ROBERTS, R. D. MACPHEE, J. S. HAILE, F. BROCK, *et al.*, 2011. Paper II–Dirt, dates and DNA: OSL and radiocarbon chronologies of perennially frozen sediments in Siberia, and their implications for sedimentary ancient DNA studies. Boreas **40**: 417–445.

GUTHRIE, R. D. 2001. Origin and causes of the mammoth steppe: a story of cloud cover, woolly mammal tooth pits, buckles, and inside-out Beringia. Quaternary Science Reviews **20**: 549 –574.

HAILE, J., R. HOLDAWAY, K. OLIVER, M. BUNCE, M. T. P. GILBERT, *et al.*, 2007. Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? Molecular Biology and Evolution **24**: 982–989.

KOCH, P. L. and A. D. BARNOSKY. 2006. Late Quaternary extinctions: state of the debate. Annual Review of Ecology, Evolution, and Systematics **37**: 215–250.

STEVEN, B., G. BRIGGS, C. P. MCKAY, W. H. POLLARD, C. W. GREER, and L. G. WHYTE. 2007. Characterization of the microbial diversity in a permafrost sample from the Canadian high Arctic using culture-dependent and culture-independent methods. FEMS Microbiology Ecology **59**: 513–523.

TURVEY, S. T. and S. A. FRITZ. 2011. The ghosts of mammals past: biological and geographical patterns of global mammalian extinction across the Holocene. Philosophical Transactions of the Royal Society B: Biological Sciences **366**: 2564–2576.

VAN GEEL, B., G. D. ZAZULA, and C. E. SCHWEGER. 2007. Spores of coprophilous fungi from under the Dawson tephra (25,300 ^{14}C years BP), Yukon Territory, northwestern Canada. Palaeogeography, Palaeoclimatology, Palaeoecology **252**: 481–485.

Yoccoz, N. G., K. A. Bråthen, L. Gielly, J. Haile, M. E. Edwards, *et al.*, 2012. DNA from soil mirrors plant taxonomic and growth form diversity. Molecular Ecology **21**: 3647–3655.

Zazula, G. D., D. G. Froese, S. A. Elias, S. Kuzmina, C. L. Farge, *et al.*, 2006. Vegetation buried under Dawson tephra (25,300 ^{14}C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada. Palaeogeography, Palaeoclimatology, Palaeoecology **242**: 253–286.

Chapter 1

Ecological Change Across the Pleistocene-Holocene Transition

TARA SADOWAY
McMaster University
tsadoway@gmail.com

Abstract

Reconstruction of the flora and fauna of ancient environments has typically been achieved with macroscopic evidence, including pollen grains and fossils. Targeted DNA has been used in specific cases to highlight the ecological transition from forest to grassland or to establish the presence of particular species, but it has rarely been used to support a more detailed picture of the transition of an ancient ecosystem to a modern one. We will describe the use of targeted DNA amplifications of multiple genetic loci from Yukon permafrost cores spanning twelve time points from 30,000 years before present until the modern era. Due to the degraded nature of ancient DNA (aDNA), we used primer sets to amplify short DNA fragments (49-154 base pairs) from high copy number cellular elements such as chloroplasts and mitochondria. Plants were targeted with a section of the barcoding gene rbcL and part of the trnL UAA intron. Mammalian DNA was amplified from the mitochondrial loci of 16S rRNA, 12S rRNA, and cytochrome b. The cytochrome b primers were designed to target species native to the area before and after the Holocene epoch. This study demonstrates the change in macroscopic organisms over the Holocene-Pleistocene transition as the area underwent an extreme environmental change and a complete reshuffling of the dominant taxa.

I Introduction

In the 10,000 years preceding the Holocene epoch (11,700 calibrated years before present (cal yrs BP), STEFFENSEN *et al.*, 2008), the North American landscape experienced a series of climatic shifts that caused a redefinition in the predominant groups of flora and fauna. At the end of the Pleistocene era, the North American continent was partially covered by the Cordilleran ice sheet in the North West and the Laurentide ice sheet in the Northern Central and Eastern areas. Through a series of short warming and cooling periods, the Pleistocene glaciers retreated, causing a major redefinition of the landscape and a rise in sea levels by up to 20 meters (WEAVER *et al.*, 2003). These dramatic climate shifts in geologically brief time spans caused or exacerbated the extinction of more than 200 species of large mammals "megafauna" (KOCH and BARNOSKY, 2006; TURVEY and FRITZ, 2011) and completely changed the functional composition of the plant community (BEATTY and PROVAN, 2011; BREEN *et al.*, 2012). In the Yukon (Canada), evidence supports 1,400 years of warming starting at 14.4k cal yrs BP, with a small cooling period of roughly 1,000 years around 12.2k cal yrs BP (IRVINE *et al.*, 2012). The warmest period for this area occurred at 10.8k cal yrs BP, for about 1000 years, beginning just after the start of the Holocene period (IRVINE *et al.*, 2012). As the Earth enters its newest extinction event, it is of particular interest and relevance to examine the Pleistocene to Holocene (PH) transition, which is a well demarcated shift from a cold, dry environment rich in megafauna to a wet, semi-boreal biome replete with a different congregation of vertebrate species. In this study, we use permafrost soil cores to reconstruct ancient ecosystems via preserved genetic material.

Permafrost preserves a very complete record of ancient ecosystems, comprising macrofossil, microfossil, DNA, and climate data (WILLERSLEV *et al.*, 2003; HAILE *et al.*, 2009). Although there exists a robust mammalian fossil record from the Yukon, all fossils records are inherently biased (TURVEY and BLACKBURN, 2011). Macrofossils may over-represent species with large mass and under-represent predators; as well, it is unlikely to discover the last surviving member of an extinct species in the fossil record. In terms of fossil flora, plant matter is subject to rapid cellular degradation (DINTER and BIRKS, 1996) and much of the knowledge of these organisms comes from the palynological record (BEATTY and PROVAN, 2011) and accumulated plant biomatter (BIRKS and BIRKS, 2000; ANDERSON-CARPENTER *et al.*, 2011). Pollen datasets are skewed toward high-pollen producers such as Asteraceae (the daisy family) and away from plants that spread more readily through shoots and rhizomes than seed dispersal. In the past few years, the genetic record has been used to expand and substantiate these ecological remnants (WILLERSLEV *et al.*, 2003; WILLERSLEV *et al.*, 2007; SØNSTEBØ *et al.*, 2010). DNA records can confirm the presence of species

(HAILE *et al.*, 2009) and describe diversity (LYDOLPH *et al.*, 2005; DE BRUYN *et al.*, 2011). The ampliconic loci for the hypothesized plant communities of the PH transition were chosen for a well-supported, modern reference genomic database (SAARELA *et al.*, 2013; SØNSTEBØ *et al.*, 2010). For the vertebrates, the loci were chosen to reflect the varied and diverse set of barcoding or mitochondrial sequences currently available in GenBank (for example, EPP *et al.*, 2012; ANDERSEN *et al.*, 2012; FICETOLA *et al.*, 2008).

Permafrost is an ideal medium for the preservation of biological material. In addition to the low, invariant temperature, permafrost soils typically lack the soil leaching and cycling that occur in active, temperate soils (HAILE *et al.*, 2007). In many areas of permafrost, the soil is frozen during or shortly after the deposition of organic material and sediment (ARNOLD *et al.*, 2011), thus preserving a snapshot of a particular time. Studies have demonstrated the preservation and recovery of soil DNA from permafrost in excess of 500,000 years old (D'COSTA *et al.*, 2011). As well, modern sedimentary DNA has been shown to reflect the diversity of the local macroflora (YOCCOZ *et al.*, 2012) and vertebrate communities (ANDERSEN *et al.*, 2012; HEBSGAARD *et al.*, 2009; DE BRUYN *et al.*, 2011).

Previous research has established the type and variability of plant diversity in Beringia. In the Yukon, research on plant macrofossils, fungal spores, and pollen grains has shown that the steppe-tundra environment was rich in graminoids, such as *Carex*, 30k years ago at the deposition of the Dawson tephra, but sparse in more arboreal species (VAN GEEL *et al.*, 2007; ZAZULA *et al.*, 2006). As well, their examination of ground squirrel nests showed a variety of dry and mesic forbs such as *Draba* and *Ranunculus* (ZAZULA *et al.*, 2006).

This research spans twelve time points from across the PH transition to elucidate the metagenetic record. In this research, steps have been taken to reduce contamination, sampling bias, and stochastic effects along the entire sampling pipeline. Our research challenges the graminoid (grasses, rushes, sedges) dominated view of the Arctic mammoth steppe in the very late Pleistocene (GUTHRIE, 2001) and suggests that a change in the forbs (herbaceous flowering plants) was instead the limiting factor for the megafauna.

II Methods

i Sample Extraction and Amplification

In this study, samples were obtained from placer gold mines near Dawson city (64°03′36"N 139°24′39"W) in the Yukon Territory of Canada. Exposed mining surfaces allowed for easy access to a well-defined stratigraphic record with highly identifiable volcanic tephra (FROESE *et al.*, 2006; DEMURO *et al.*, 2008). The sample sites were cleared back to the frozen layer (+30 cm) and cored with custom drilling equipment. Permafrost cores were sprayed with a green fluorescent protein (GFP) DNA construct to allow the detection of external contamination after DNA extraction. Cores were transported frozen to the University of Alberta (Edmonton, Alberta, Canada). Some cores were subsampled and homogenized in a clean hood at the University of Alberta. Other cores were subsampled with coring as per D'COSTA *et al.*, 2011 at the McMaster Ancient DNA facility (Supplementary Materials (SM), Table S1). All samples were manipulated in the presence of air and TE (0.1X Tris-EDTA) blanks (SM Table S2). Samples were extracted in duplicate with guanidinium protocols (D'COSTA *et al.*, 2011; BOOM *et al.*, 1990) from 250mg of soil, purified with silica (HÖSS and PÄÄBO, 1993; D'COSTA *et al.*, 2011), and eluted twice from the pellet (HANDT *et al.*, 1996). The samples LL2D and LL3D were extracted with MoBio PowerSoil Kits (Carlsbad, CA, USA). Due to the presence of extract inhibition, as determined using the method developed by KING *et al.*, 2009, the DNA extracts were diluted tenfold and amplified in duplicate for each primer set (SM Table S7), except for LL2D and LL3D which were amplified without further dilution. All samples were screened for GFP by quantitative polymerase chain reaction (qPCR). While most samples showed no GFP amplicon, a few samples (A2, S2) showed fewer than ten copies). The primers targeted the plant loci in a section of the large subunit of the ribulose-bisphosphate carboxylase (rbcL) gene (WILLERSLEV *et al.*, 2003; CBOL PLANT WORKING GROUP, 2009; HOLLINGSWORTH *et al.*, 2011) and part *gh* of the *trn*L UUA intron (TABERLET *et al.*, 2007) and followed established protocols. The mammalian primers targeted 49-154 base pair portions of three mitochondrial loci: 16S rRNA (HÖSS *et al.*, 1996; RASMUSSEN *et al.*, 2009), 12S rRNA (KUCH *et al.*, 2002), and cytochrome *b* (*cyt b*). Due to the number of relevant vertebrate reference sequences in GenBank, the *cytb* locus (80 taxa) was chosen over the barcoding locus cytochrome c oxidase 1 (60 taxa). For the *cytb* locus, a set of degenerate primers were designed with FastPCR (KALENDAR *et al.*, 2011) to amplify sequences from expected vertebrates. Of these, one forward (cytbF1 - ACACTACACNTCAGACACAACAACAGC) and two reverse primers (cytbR2 - ANCCGTAGTTTACGTCTCGGCAG, cytbR3 - CANCCGTAGTTTACGTCTCGGCA) were suf-

ficient to amplify from the majority of samples. The *cytb* amplification was most efficient in 20µL reactions (AmpliTaq Gold (0.05U/µL), 1X PCR Buffer II, 2.5mM $MgCl_2$, 0.25mM dNTPs, 0.5X Evagreen, 250nM forward/reverse primers) when cycled with a 3 minute denaturation at 95°C and 45 cycles of 95°C for 30 seconds and 60°C for 30 seconds. QPCR products were purified with 10K AcroPrep Pall plates (Pall Canada Direct Ltd., Mississauga, ON, Canada) using a vacuum manifold. QPCR quantitation assays were used to pool each amplicon set in equimolar concentrations for each extract.

ii Library Preparation, Indexing, and Data Pipeline

The amplicon pools were library prepped (MEYER and KIRCHER, 2010), quantified, and dual indexed (KIRCHER *et al.*, 2012) according to standard Illumina protocols. Each sample extract (*i.e.* GB1A) had its own unique combination of forward and reverse indices. Purification steps were done with MinElute PCR purification kits (Qiagen, Toronto, ON) with two buffer PE washes. The samples were sequenced on a HiSeq 2500 Rapid Run (2x100bp) at approximately 100,000 reads each (Illumina Cambridge Ltd, Essex, UK). DNA sequence reads were demultiplexed by Illumina software and then additionally with custom Perl scripts because low levels of mis-binning were identified in the datasets. Mis-indexed sequences are a noted amplicon problem (CARLSEN *et al.*, 2012) that would not have been apparent in our dataset if all of our reads were longer than the total sequencing run length. Although we removed the detectable mis-binned reads (<1%), we theorize that an equivalent amount of reads were undetectably mis-binned, but we do not believe this would significantly alter any of our results due to the total volume of reads. Cutadapt V1.2.1 (MARTIN, 2011) was used to remove Illumina adapters and quality trim and SeqPrep (ST JOHN, 2011) was used to merge the forward and reverse reads. Reads were sorted, chimera checked, and clustered with Usearch V6.0.307 (EDGAR, 2010) (please refer to SM Table S4 for Cutadapt, SeqPrep, and Usearch parameters). The reads were stripped of PCR adapters and sorted into bins with custom Perl scripts. All remaining reads were aligned with BLASTN (ALTSCHUL *et al.*, 1990) at default parameters (Evalue = 10^{-10}) against constrained reference databases of Arctic plants (SAARELA *et al.*, 2013; SØNSTEBØ *et al.*, 2010) and vertebrates. To create a constrained vertebrate reference list, up to three representative sequences for either the full mitochondrial genome or the individual loci (*cytb*, 12S rRNA, 16S rRNA) were compiled as available (SM Tables S5, S6). The aligned sequences were assigned to taxonomic ranks with MEGAN4 V4.70.4 (HUSON *et al.*, 2011) (Min Support 1, Min Score 50, Top Percent 10, Min Complexity 0.05) and mapped back to the original sequence counts. Singletons were removed after time point replicates

were merged to reduce spurious clusters and sequencing artefacts (DICKIE, 2010; HUSE *et al.*, 2010; TEDERSOO *et al.*, 2010).

To analyze the relative ecological dissimilarity between time points, the samples were grouped by core location and compared in R (R CORE TEAM, 2013) using a non-metric multidimensional scaling (NMDS) analysis with the Bray-Curtis algorithm; the Vegan package was used to transform and normalize the data (OKSANEN *et al.*, 2013).

III Results

i The Floral Record

The data are more informative for the plant loci than the mammalian loci, in relation to the total diversity and the relative shift in functional communities. There is a sharp decrease in the abundance of grasses, Poaceae, from the late Pleistocene into the Holocene (Table 1). Many of the forbs also saw a decrease towards the end of the Pleistocene, including the families Asteraceae (flowers), Plantaginaceae (plantain), and Boraginaceae (borage). Other families such as Cyperaceae (sedges) appear to be more opportunistic, as evidenced by their peak at the very end of the Pleistocene, during the least stable climatic period. By the beginning of the Holocene, the flowers and herbaceous communities were replaced by dwarf shrubs, shrubs, and small trees, particularly of the families Salicaceae (willows), Betulaceae (birch), and Rosaceae (roses).

We grouped the plant data into functional communities to get a better sense of the landscape during the PH transition (Figure 1, raw data available in SM Table S3). Classification into these groups was based primarily on WILLERSLEV *et al.*, 2014, but also included information from CODY, 2000. Families with more than three examples of the same functional representation at lower taxonomic ranks were grouped into those categories. One of the most striking results is the high relative abundance of forbs before the PH transition and their subsequent replacement with various shrubs and trees. Although there appears to be a discrepancy at 13150 cal yrs BP, it should be noted that this sample's carbon-14 error range overlaps with the range of the 13105 cal yrs BP and the 13205 cal yrs BP samples; thus, the 13150 cal yrs BP sample could be slighter older than the 13205 cal yrs BP sample. The amount of graminoid taxa, particularly grasses, varies slightly, but not as widely as the forbs; in fact, we see comparatively stable relative levels of these plants before and after the transition.

Table 1: Relative abundance of family level taxonomic assignments for plant loci. The number of DNA extracts per core is show as n=x.

Cal Yrs BP	30000	21135	15865	15405	13870	13205	13150	13105	10340	10220	9820	Modern
Family	n=3	n=3	n=3	n=4	n=1	n=4	n=3	n=3	n=2	n=3	n=3	n=7
Adoxaceae				<1%					4.7%	<1%	<1%	
Amaryllidaceae		<1%										
Apiaceae	2.8%	6.6%	2.7%	10.1%	<1%	<1%	<1%	<1%	<1%			
Asteraceae	19.5%	56.1%	67.1%	74.3%	92.6%	4.3%	79.4%	6.9%	9.6%	<1%	5.6%	7.0%
Betulaceae	<1%					6.4%	<1%	<1%	13.0%	7.6%	6.1%	1.0%
Boraginaceae	17.7%	6.5%	<1%	<1%		<1%			<1%			<1%
Brassicaceae	2.2%	<1%		<1%		<1%	<1%	1.0%	<1%			<1%
Caprifoliaceae									1.1%		3.9%	
Caryophyllaceae	<1%	<1%						<1%				
Celastraceae								<1%				
Cornaceae	<1%											
Crassulaceae	<1%	<1%										
Cyperaceae	4.9%	2.6%	<1%	<1%		27.0%	13.2%	15.9%	2.4%	3.7%	<1%	29.7%
Diapensiaceae		<1%										
Elaeagnaceae									<1%			
Equisetaceae									<1%			<1%
Ericaceae	<1%	<1%	<1%			39.6%		<1%	4.3%	<1%	6.9%	<1%
Fabaceae	<1%	<1%		<1%	<1%	1.9%	2.2%	<1%	<1%	<1%	<1%	
Grossulariaceae									7.8%		<1%	
Haloragaceae												<1%
Lentibulariaceae	<1%						<1%					
Onagraceae		<1%							7.9%			
Ophioglossaceae									<1%			
Orobanchaceae	<1%	<1%		<1%				<1%				
Papaveraceae	4.6%	2.3%					<1%					
Pinaceae		2.0%	<1%						<1%		20.0%	1.8%
Plantaginaceae	2.1%	2.8%	3.8%	<1%	5.7%			<1%				<1%
Plumbaginaceae												<1%
Poaceae	34.2%	11.6%	26.1%	14.3%	1.2%	<1%	1.2%	4.8%	2.0%	3.5%	<1%	3.0%
Polemoniaceae	<1%	<1%			<1%		<1%	<1%	<1%			
Polygonaceae	<1%	<1%		<1%								
Potamogetonaceae	<1%		<1%									
Primulaceae	<1%	<1%										
Ranunculaceae	<1%	5.5%						<1%	<1%		<1%	
Rosaceae	2.8%	<1%	<1%	<1%		<1%	2.0%	<1%	2.2%	79.3%	<1%	53.1%
Salicaceae	7.3%	1.1%	<1%	<1%		17.0%	1.2%	68.0%	41.9%	<1%	55.8%	3.2%
Saxifragaceae	<1%								2.6%		<1%	<1%
Typhaceae			<1%	1.1%		3.5%	<1%			5.6%	<1%	<1%

The line denotes the Pleistocene-Holocene transition.

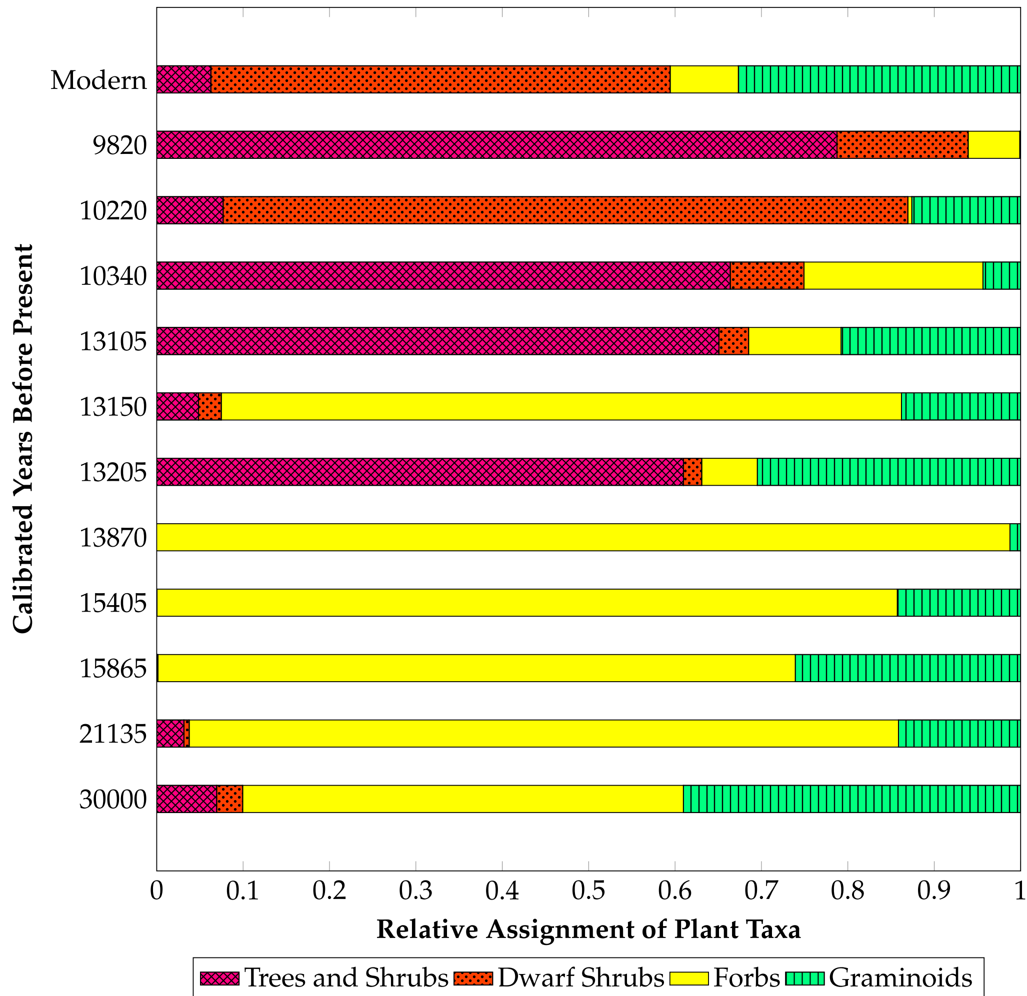


Figure 1: Relative assignment of plant taxa up to family rank identifications by major functional group.

We examined pollen data from the study site in relation to the ancient DNA data (Figure 2). The trends apparent in the functional group changes of plants are supported by the overall distributions of both the pollen and the DNA datasets. Certain taxonomic affiliations, such as Asteraceae, show a complementary relationship between the datasets both before and after the PH transition. However, some families, such as Salicaceae and Rosaceae, are highly represented in the DNA dataset, but poorly in the pollen set. Conversely, Equisetaceae, a horsetail family indicative of wet soil, is only present in the Holocene, and barely registers in the DNA data (0.005%). This may be due to normal pollen distributions for these plants (low versus high producers), DNA extraction biases (putatively skewed towards unbound molecules in the soil), or annealing primer bias in the DNA amplification. Additionally, the loci used to amplify the plant DNA is plastid-specific, so we expect that pollen DNA of angiosperms (but not gymnosperms) will not amplify and most of the angiosperm DNA will derive from macrofossil remains (MCCAULEY *et al.*, 2007), such as roots, shoots, or leaves.

The NMDS analysis shows that the oldest four samples cluster very closely together, indicating a similar set of taxa (Figure 3). The samples from around 13k cal yrs BP do not cluster within themselves as strongly, but they do show a median position between the Holocene samples and oldest samples (pre-Pleistocene warming). A preponderance of forbs, shown in Figure 3 as ✺, cluster with the pre-Holocene samples, while the Holocene samples are most closely linked to trees, shrubs, and dwarf shrubs. As in Figure 1, the graminoids (×) are evenly distributed amongst all the samples and plant groups.

ii The Faunal Community

Initially, the data for the animal loci were aligned to the nucleotide database for species identification. However, the 16S loci returned very few BLAST hits and many of the 12S sequences showed less than 90% similarity to their putative matches. To decrease the noise and reassess the animal sequences, the loci were matched against a curated list of sequences from expected taxa (SM Tables S5, S6). This resulted in a higher frequency of top identity matches, even in the very short (<50 base pairs) cytochrome *b* fragments. There are still rogue sequences that are not identifiable at lower than family rank, especially in the samples leading up to the Holocene, but they are most likely extinct species that do not have a reference sequence in GenBank. Unfortunately, the lack of a complete set of genetic sequences for Pleistocene vertebrates reduces our ability to completely quantify the total diversity of the vertebrate loci. While 35 species have a complete mitochondrial genome available, 30 taxa have no loci at all. An additional 28 species have all three

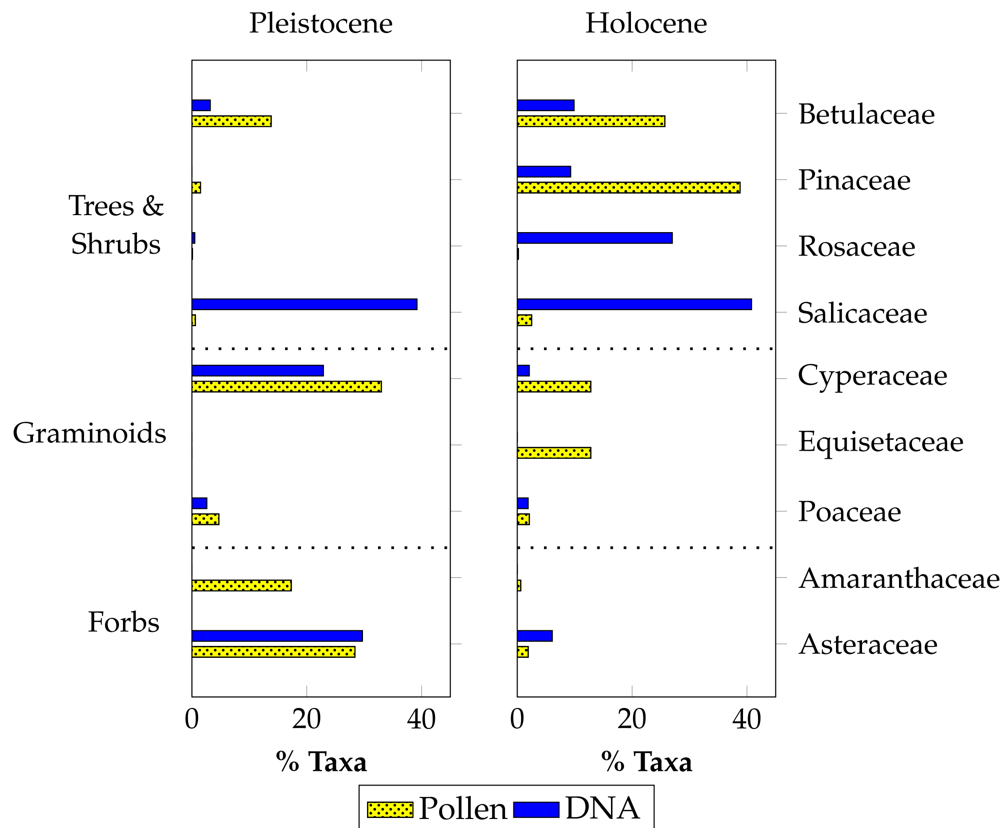


Figure 2: Family level assignments of taxa are shown for the pollen and DNA datasets. The pollen set was less diverse, so the DNA data was constrained to matching family level assignments. Both the Pleistocene (11,700-13,900 cal yrs BP) and Holocene (9,500-11,700 cal yrs BP) groups of both types of data are depicted as relative levels within each set. A few taxa contained only very low (<1%) levels of representation, so only families with at least one constituent higher than 2% are shown.

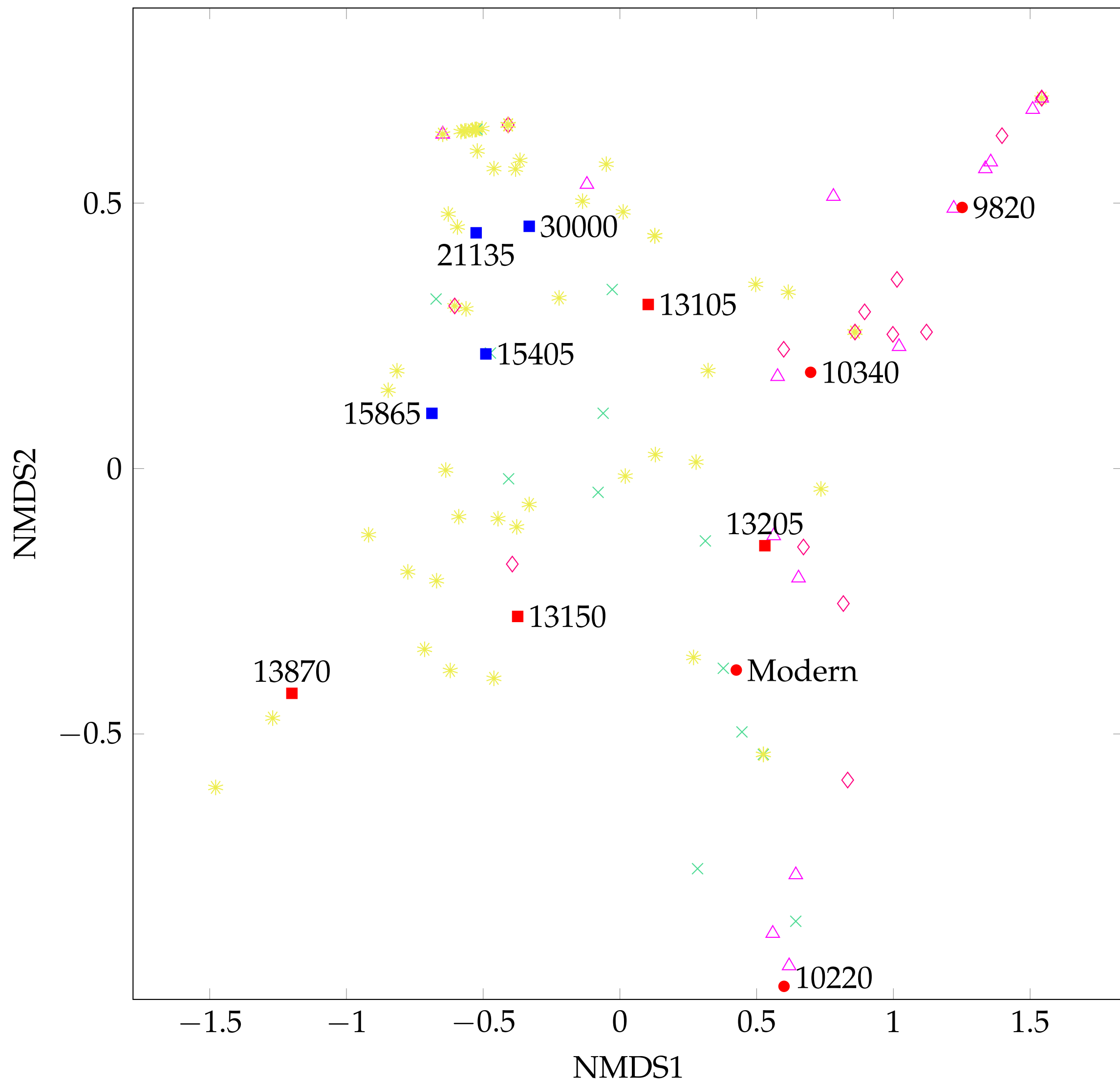


Figure 3: Bray-Curtis NMDS of genus level assignments of plants. Holocene samples are represented by closed circles, while Pleistocene samples are shown as closed squares. Plant species are marked according to function (forbs (✳), graminoids(×), dwarf shrubs (△), and trees and shrubs (◇)).

loci, either partially or in full, in the reference dataset, while 19 species have 2 loci, and 36 species have only 1, typically cytochrome *b*. Although cytochrome *b* had a better set of representative genetic sequences to compare with, our *cytb* dataset had the highest proportion of unidentifiable sequences. This may be due to the varying lengths of the amplified target and the small size of the amplicon.

Overall, the vertebrate loci had one tenth (~190k) of taxonomically identifiable reads compared to the plant loci (~1.8 million); however, this was expected as the qPCR counts for the vertebrate loci (data not shown) were much lower than for plants and the vertebrate reference database is less complete. Thus, it is unsurprising that the clustering would show a slightly different pattern in the NMDS compared to the plant data (Figure 4). Although the two oldest samples cluster, there is a looser correlation between the Holocene set and the very late Pleistocene set (~13k year old samples). The NMDS plot is consistent with our predictions for the distribution of vertebrate genera. The Holocene samples cluster with Homo (humans), Lagopus (ptarmigan), and Canis (wolves), while the oldest samples relate more strongly to the megafauna Mammuthus (mammoths), Ovis (sheep), and Bison. Ptarmigan do well when they have access to high quality proteins, which is supported by the concurrent increase in Fabaceae, a pea/legume based family (ALLEN and CLARKE, 2005), and the willow family, Salicaceae (WILSON and MARTIN, 2008). Ancient bison in the Northern latitudes typical ate a diet rich in graminoids (WILLERSLEV *et al.*, 2014) and are able to exclusively survive on sedges (Cyperaceae) (FORTIN *et al.*, 2003), but we do not see an expansion or even continuation in the bison species in the few thousand years before the Holocene, despite an increase in Carex at this time.

iii Metagenetic Profile

To build a metagenetic profile and focus the ecological resolution, the animal and plant loci datasets were amalgamated (ZHOU *et al.*, 2011). Figure 5 shows the Bray-Curtis dissimilarity between the twelve time points. A NMDS analysis of the difference between metagenetic constrained list and the total nucleotide assignments (from a general BLAST against the nucleotide database) can be found in SM Figure S1, but the relative positions did not change very much for the majority of the samples. In Figure 5, both the Holocene samples and the four oldest late Pleistocene samples cluster closely amongst themselves. The metagenetic dissimilarity between the 13k year old samples is quite different both within and without the set, reflecting the great ecological variability in the time directly preceding the Holocene epoch. The large dissimilarity of 13.8k cal yrs BP is probably due to the sample's low diversity and low replicate number (n=1); we were not able to overcome the

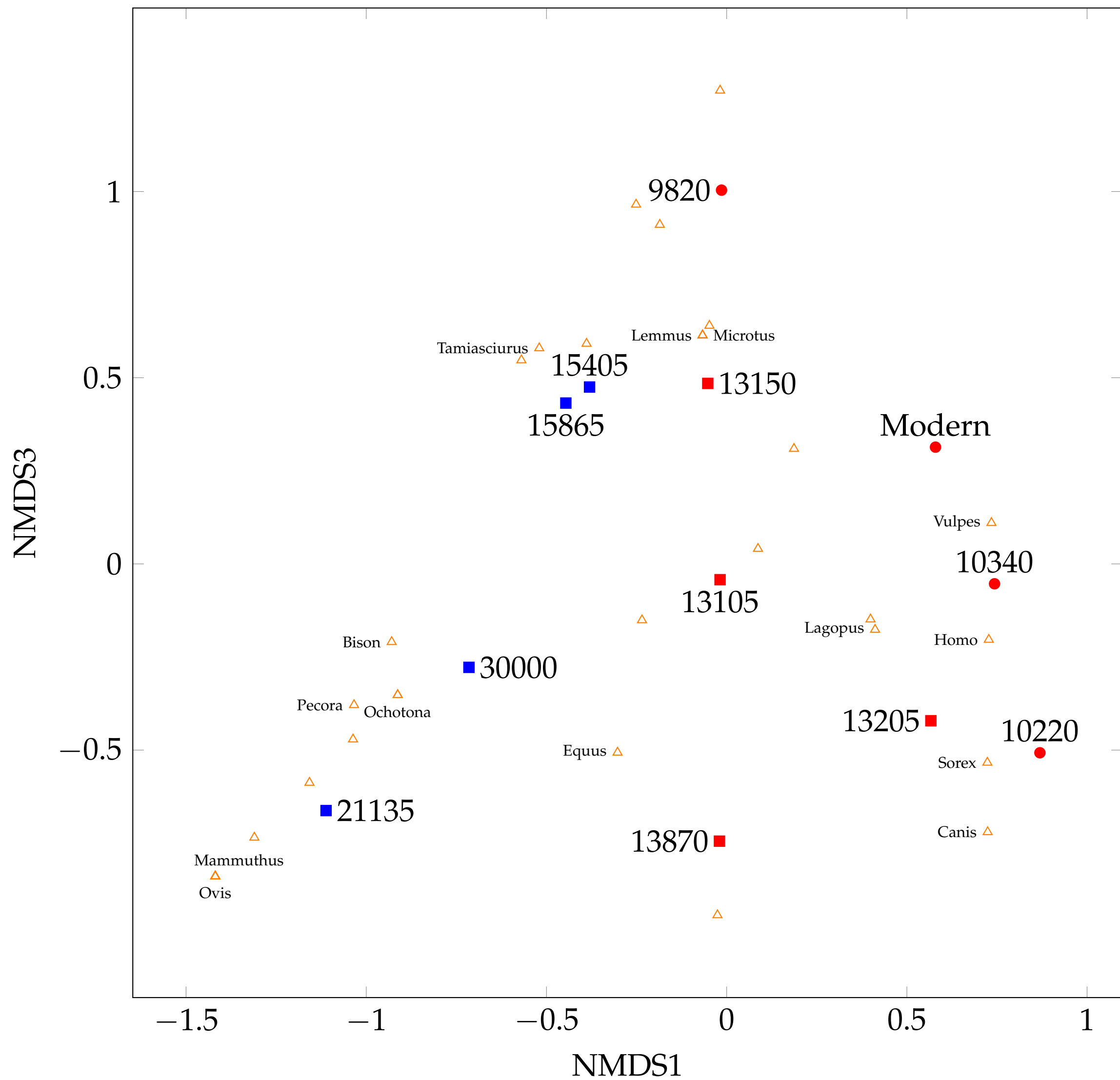


Figure 4: Bray-Curtis NMDS vertebrate data at the genus level and higher. Holocene samples are represented by closed circles, Pleistocene samples are shown as closed squares, and vertebrates are shown as open triangles. Vertebrates identified at the genus level are labelled.

duplicate extracts' inhibition through dilution or purification. It is also possible that the DNA in the 13.8 cal yrs BP sample is too degraded and the low signal quality is a preservational bias.

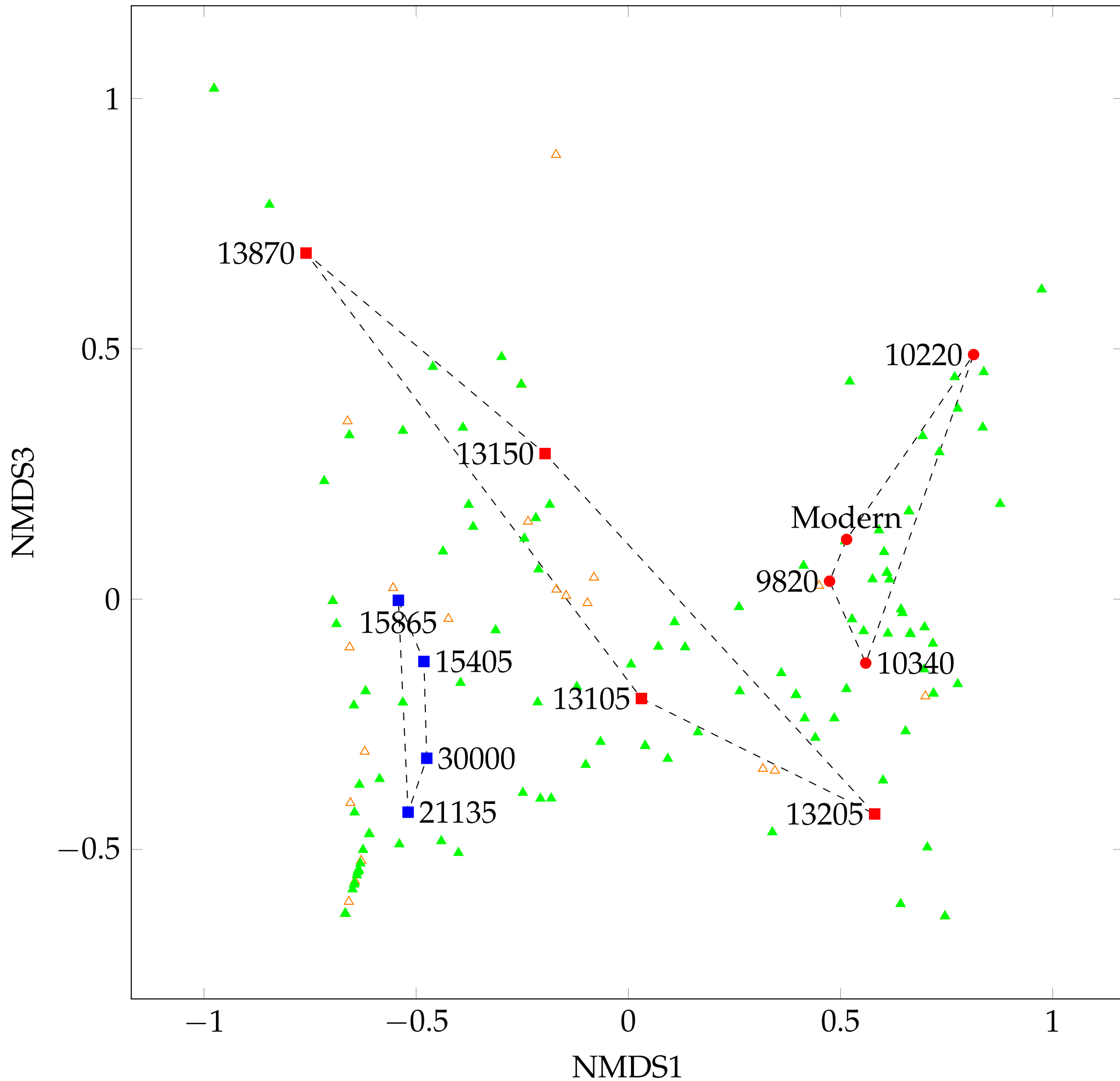


Figure 5: Bray-Curtis NMDS plant and vertebrate data combined at the genus level or higher. Holocene samples are represented by closed circles, Pleistocene samples are shown as closed squares, vertebrates are shown as open triangles, and plants are shown as closed triangles. The oldest samples from before 15k cal yrs BP cluster strongly with a subset of plants and animals tolerant of colder and drier climes. Samples from just before the PH transition (13k cal yrs BP) occupy a relative position between the older samples and the Holocene samples.

IV Discussion

Although other research has been quick to identify the decline of megafauna as human-driven or climate-driven, our research indicates that a more complex feedback loop caused the permanent community changes through the Pleistocene-Holocene transition. There is a definite shift in both the abundance and identity of plants that lived 15k years ago versus the plants that proliferated at the very end of the Pleistocene, which is different again from those in the Holocene. As many of the forbs were replaced by trees and shrubs, the diets of megafaunal grazers would have been negatively impacted. As the megafaunal taxa declined, so would their output of nitrogen-rich manure, their environment-shaping wallowing behaviours (COPPEDGE *et al.*, 1999), and their specific grazing patterns (WILLERSLEV *et al.*, 2014). As well, there is ample previous evidence to show that many species were in genetic decline even before the last glacial maximum (SHAPIRO *et al.*, 2004; BRACE *et al.*, 2012; STILLER *et al.*, 2013), so the combination of an unstable food source with the introduction of predation pressures from humans, lends credence to the blended theory of megafaunal extinction (PRESCOTT *et al.*, 2012).

The data suggest that the plant community dynamics shifted gradually as the warming trend began up to about 15k years ago and then there was a punctuated fluctuation leading up to the Holocene epoch and plant communities were not entirely established even 2000 years after this point. As trees and shrubs filled habitats previously filled by grasses, combined with a dearth of megafaunal grazers, the pre-Holocene environment was unable to shift back to a forb dominated community. At this point, our data suggests that as species were lost, there was a surge in the abundance of smaller mammals and other animals such as grouse, which were better suited to the new set of vegetation. This is interesting because we know that many of the extinct species survived a number of previous climate oscillations, but many species were not able to recover from this particular shift (HEWITT, 2004).

By 10k years BP, the ungulates sequence records are completely devoid of mammoths and sheep, as predicted by the fossil record, but we still see a small amount of bison and horse. This supports a landscape predominated by shrubs instead of grasslands, as living in a landscape filled with Betulaceae (birch) and Salicaceae (willow) of this period would have been slightly more tenable for bison than mammoth. As many of the largest ungulates and pachyderms went extinct, the soils would have quickly lost quality, and we do see an increase in taxa that do well in poor soils (for example, Ericaceae or Caprifoliaceae). The wetter climate at this point allowed new species to flourish, such as the marsh herb family of Typhaceae and the horsetails of the family Equisetaceae.

Our results show that environmental change in the late Pleistocene led to the destabilization of established functional groups of organisms. Climate change ignited a feedback loop of ecological revision, resulting in the extinction of the many vertebrates that were dependent on these Pre-Holocene biological communities. The ability to understand and putatively predict the effects of climate change may help us protect at-risk species currently in the Arctic (DAWSON *et al.*, 2011; GROSSE *et al.*, 2011; BROOK and BARNOSKY, 2012.).

V References

ALLEN, T. and J. A. CLARKE. 2005. Social learning of food preferences by white-tailed ptarmigan chicks. Animal Behaviour **70**: 305–310.

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. Journal of Molecular Biology **215**: 403–410.

ANDERSEN, K., K. L. BIRD, M. RASMUSSEN, J. HAILE, H. BREUNING-MADSEN, *et al.*, 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Molecular Ecology **21**: 1966–1979.

ANDERSON-CARPENTER, L. L., J. S. MCLACHLAN, S. T. JACKSON, M. KUCH, C. Y. LUMIBAO, and H. N. POINAR. 2011. Ancient DNA from lake sediments: Bridging the gap between paleoecology and genetics. BMC Evolutionary Biology **11**: 30.

ARNOLD, L. J., R. G. ROBERTS, R. D. MACPHEE, J. S. HAILE, F. BROCK, *et al.*, 2011. Paper II–Dirt, dates and DNA: OSL and radiocarbon chronologies of perennially frozen sediments in Siberia, and their implications for sedimentary ancient DNA studies. Boreas **40**: 417–445.

BEATTY, G. E. and J. PROVAN. 2011. Phylogeographic analysis of North American populations of the parasitic herbaceous plant *Monotropa hypopitys* L. reveals a complex history of range expansion from multiple late glacial refugia. Journal of Biogeography **38**: 1585–1599.

BIRKS, H. H. and H. J. B. BIRKS. 2000. Future uses of pollen analysis must include plant macrofossils. Journal of Biogeography **27**: 31–35.

BOOM, R, C. J. SOL, M. M. SALIMANS, C. L. JANSEN, P. M. WERTHEIM-VAN DILLEN, and J. VAN DER NOORDAA. 1990. Rapid and simple method for purification of nucleic acids. Journal of Clinical Microbiology **28**: 495–503.

BRACE, S., E. PALKOPOULOU, L. DALÉN, A. M. LISTER, R. MILLER, *et al.*, 2012. Serial population extinctions in a small mammal indicate Late Pleistocene ecosystem instability. Proceedings of the National Academy of Sciences, doi:10.1073/pnas.1213322109.

BREEN, A. L., D. F. MURRAY, and M. S. OLSON. 2012. Genetic consequences of glacial survival: the late Quaternary history of balsam poplar (*Populus balsamifera* L.) in North America. Journal of Biogeography **39**: 918–928.

BROOK, B. W. and A. D. BARNOSKY. 2012. Quaternary extinctions and their link to climate change. Springer, p. 179–198.

CBOL PLANT WORKING GROUP. 2009. A DNA barcode for land plants. Proceedings of the National Academy of Sciences USA **106**: 12794–12797.

CARLSEN, T., A. B. AAS, D. LINDNER, T. VRÅLSTAD, T. SCHUMACHER, and H. KAUSERUD. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecology **5**: 747–749.

CODY, W. J. 2000. Flora of the Yukon Territory. 2nd. NRC Research Press, 669 p.

COPPEDGE, B. R., S. D. FUHLENDORF, D. M. ENGLE, B. J. CARTER, and J. H. SHAW. 1999. Grassland soil depressions: Relict bison wallows or inherent landscape heterogeneity? American Midland Naturalist **142**: 382–392.

D'COSTA, V. M., C. E. KING, L. KALAN, M. MORAR, W. W. SUNG, *et al.*, 2011. Antibiotic resistance is ancient. Nature **477**: 457–461.

DAWSON, T. P., S. T. JACKSON, J. I. HOUSE, I. C. PRENTICE, and G. M. MACE. 2011. Beyond predictions: biodiversity conservation in a changing climate. Science **332**: 53–58.

DE BRUYN, M., A. R. HOELZEL, G. R. CARVALHO, and M. HOFREITER. 2011. Faunal histories from Holocene ancient DNA. Trends in Ecology & Evolution **26**: 405–413.

DEMURO, M., R. ROBERTS, D. FROESE, L. ARNOLD, F. BROCK, and C. RAMSEY. 2008. Optically stimulated luminescence dating of single and multiple grains of quartz from perennially frozen loess in western Yukon Territory, Canada: Comparison with radiocarbon chronologies for the late Pleistocene Dawson tephra. Quaternary Geochronology **3**: 346–364.

DICKIE, I. A. 2010. Insidious effects of sequencing errors on perceived diversity in molecular surveys. New Phytologist **188**: 916–918.

DINTER, M. and H. BIRKS. 1996. Distinguishing fossil *Betula nana* and *B. pubescens* using their wingless fruits: implications for the late-glacial vegetational history of western Norway. Vegetation History and Archaeobotany **5**: 229–240.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**: 2460–2461.

EPP, L. S., S. BOESSENKOOL, E. P. BELLEMAIN, J. HAILE, A. ESPOSITO, *et al.*, 2012. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. Molecular Ecology **21**: 1821–1833.

FICETOLA, G. F., C. MIAUD, F. POMPANON, and P. TABERLET. 2008. Species detection using environmental DNA from water samples. Biology Letters **4**: 423–425.

FORTIN, D., J. M. FRYXELL, L. O'BRODOVICH, and D. FRANDSEN. 2003. Foraging ecology of bison at the landscape and plant community levels: the applicability of energy maximization principles. Oecologia **134**: 219–227.

FROESE, D. G., G. D. ZAZULA, and A. V. REYES. 2006. Seasonality of the late Pleistocene Dawson tephra and exceptional preservation of a buried riparian surface in central Yukon Territory, Canada. Quaternary Science Reviews **25**: 1542–1551.

GROSSE, G., V. ROMANOVSKY, T. JORGENSON, K. W. ANTHONY, J. BROWN, and P. P. OVERDUIN. 2011. Vulnerability and feedbacks of permafrost to climate change. Eos, Transactions American Geophysical Union **92**: 73–74.

GUTHRIE, R. D. 2001. Origin and causes of the mammoth steppe: a story of cloud cover, woolly mammal tooth pits, buckles, and inside-out Beringia. Quaternary Science Reviews **20**: 549 –574.

HAILE, J., R. HOLDAWAY, K. OLIVER, M. BUNCE, M. T. P. GILBERT, *et al.*, 2007. Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? Molecular Biology and Evolution **24**: 982–989.

HAILE, J., D. G. FROESE, R. D. E. MACPHEE, R. G. ROBERTS, L. J. ARNOLD, *et al.*, 2009. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proceedings of the National Academy of Sciences **106**: 22352–22357.

HANDT, O., M. KRINGS, R. H. WARD, and S PÄÄBO. 1996. The retrieval of ancient human DNA sequences. American Journal of Human Genetics **59**: 368–376.

HEBSGAARD, M. B., M. T. P. GILBERT, J. ARNEBORG, P. HEYN, M. E. ALLENTOFT, *et al.*, 2009. 'The Farm Beneath the Sand' - an archaeological case study on ancient 'dirt' DNA. Antiquity **83**: 430–444.

HEWITT, G. M. 2004. Genetic consequences of climatic oscillations in the Quaternary. Philosophical Transactions of the Royal Society B: Biological Sciences **359**: 183–195.

HOLLINGSWORTH, P. M., S. W. GRAHAM, and D. P. LITTLE. 2011. Choosing and using a plant DNA barcode. PLoS One **6**: e19254.

HÖSS, M, A DILLING, A CURRANT, and S PÄÄBO. 1996. Molecular phylogeny of the extinct ground sloth *Mylodon darwinii*. Proceedings of the National Academy of Sciences **93**: 181–185.

HÖSS, M. and S. PÄÄBO. 1993. DNA extraction from Pleistocene bones by a silica-based purification method. Nucleic Acids Research **21**: 3913–3914.

HUSE, S. M., D. M. WELCH, H. G. MORRISON, and M. L. SOGIN. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental Microbiology **12**: 1889–1898.

HUSON, D. H., S. MITRA, H.-J. RUSCHEWEYH, N. WEBER, and S. C. SCHUSTER. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Research **21**: 1552–1560.

IRVINE, F., L. CWYNAR, J. VERMAIRE, and A. REES. 2012. Midge-inferred temperature reconstructions and vegetation change over the last ~15,000 years from Trout Lake, northern Yukon Territory, eastern Beringia. Journal of Paleolimnology **48**: 133–146.

KALENDAR, R., D. LEE, and A. H. SCHULMAN. 2011. Java web tools for PCR, *in silico* PCR, and oligonucleotide assembly and analysis. Genomics **98**: 137–144.

KING, C, R. DEBRUYNE, M. KUCH, C. SCHWARZ, and H POINAR. 2009. A quantitative approach to detect and overcome PCR inhibition in ancient DNA extracts. Biotechniques **47**: 941–949.

KIRCHER, M., S. SAWYER, and M. MEYER. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research **40**: e3.

KOCH, P. L. and A. D. BARNOSKY. 2006. Late Quaternary extinctions: state of the debate. Annual Review of Ecology, Evolution, and Systematics **37**: 215–250.

KUCH, M., N. ROHLAND, J. L. BETANCOURT, C. LATORRE, S. STEPPAN, and H. N. POINAR. 2002. Molecular analysis of a 11,700-year-old rodent midden from the Atacama Desert, Chile. Molecular Ecology **11**: 913–924.

LYDOLPH, M. C., J. JACOBSEN, P. ARCTANDER, M. T. P. GILBERT, D. A. GILICHINSKY, *et al.*, 2005. Beringian paleoecology inferred from permafrost-preserved fungal DNA. Applied and Environmental Microbiology **71**: 1012–1017.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10–12.

MCCAULEY, D. E., A. K. SUNDBY, M. F. BAILEY, and M. E. WELCH. 2007. Inheritance of chloroplast DNA is not strictly maternal in *Silene vulgaris* (Caryophyllaceae): evidence from experimental crosses and natural populations. American Journal of Botany **94**: 1333–1337.

MEYER, M. and M. KIRCHER. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols **6**: pdb.prot5448, doi: 10.1016/j.bbr.2011.03.031.

OKSANEN, J., F. G. BLANCHET, R. KINDT, P. LEGENDRE, P. R. MINCHIN, *et al.*, 2013 vegan: Community Ecology Package. R package version 2.0-10.

PRESCOTT, G. W., D. R. WILLIAMS, A. BALMFORD, R. E. GREEN, and A. MANICA. 2012. Quantitative global analysis of the role of climate and people in explaining late Quaternary megafaunal extinctions. Proceedings of the National Academy of Sciences **109**: 4527–4531.

R CORE TEAM. 2013 R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

RASMUSSEN, M., L. S. CUMMINGS, M. T. P. GILBERT, V. BRYANT, C. SMITH, *et al.*, 2009. Response to Comment by Goldberg *et al.* on "DNA from Pre-Clovis Human Coprolites in Oregon, North America". Science **325**: 148.

SAARELA, J. M., P. C. SOKOLOFF, L. J. GILLESPIE, L. L. CONSAUL, and R. D. BULL. 2013. DNA Barcoding the Canadian Arctic flora: Core plastid barcodes (rbcL + matK) for 490 vascular plant species. PLoS One **8**: e77982.

SHAPIRO, B., A. J. DRUMMOND, A. RAMBAUT, M. C. WILSON, P. E. MATHEUS, *et al.*, 2004. Rise and fall of the Beringian steppe bison. Science **306**: 1561–1565.

SØNSTEBØ, J., L GIELLY, A. BRYSTING, R ELVEN, M EDWARDS, *et al.*, 2010. Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Molecular Ecology Resources **10**: 1009–1018.

ST JOHN, J. 2011 SeqPrep. "https://github.com/jstjohn/SeqPrep".

STEFFENSEN, J. P., K. K. ANDERSEN, M. BIGLER, H. B. CLAUSEN, D. DAHL-JENSEN, *et al.*, 2008. High-Resolution Greenland ice core data show abrupt climate change happens in few years. Science **321**: 680–684.

STILLER, M., M. MOLAK, S. PROST, G. RABEDER, G. BARYSHNIKOV, *et al.*, 2013. Mitochondrial DNA diversity and evolution of the Pleistocene cave bear complex. Quaternary International, doi: 10.1016/j.quaint.2013.09.023.

TABERLET, P., E. COISSAC, F. POMPANON, L. GIELLY, C. MIQUEL, *et al.*, 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Research **35**: e14.

TEDERSOO, L., R. H. NILSSON, K. ABARENKOV, T. JAIRUS, A. SADAM, *et al.*, 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytologist **188**: 291–301.

TURVEY, S. T. and T. M. BLACKBURN. 2011. Determinants of species abundance in the Quaternary vertebrate fossil record. Paleobiology **37**: 537–546.

TURVEY, S. T. and S. A. FRITZ. 2011. The ghosts of mammals past: biological and geographical patterns of global mammalian extinction across the Holocene. Philosophical Transactions of the Royal Society B: Biological Sciences **366**: 2564–2576.

VAN GEEL, B., G. D. ZAZULA, and C. E. SCHWEGER. 2007. Spores of coprophilous fungi from under the Dawson tephra (25,300 ^{14}C years BP), Yukon Territory,

northwestern Canada. Palaeogeography, Palaeoclimatology, Palaeoecology **252**: 481–485.

WEAVER, A. J., O. A. SAENKO, P. U. CLARK, and J. X. MITROVICA. 2003. Meltwater Pulse 1A from Antarctica as a trigger of the Bølling-Allerød warm interval. Science **299**: 1709–1713.

WILLERSLEV, E., A. J. HANSEN, J. BINLADEN, T. B. BRAND, M. T. P. GILBERT, *et al.*, 2003. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science **300**: 791–795.

WILLERSLEV, E., E. CAPPELLINI, W. BOOMSMA, R. NIELSEN, M. B. HEBSGAARD, *et al.*, 2007. Ancient biomolecules from deep ice cores reveal a forested southern Greenland. Science **317**: 111–114.

WILLERSLEV, E., J. DAVISON, M. MOORA, M. ZOBEL, E. COISSAC, *et al.*, 2014. Fifty thousand years of Arctic vegetation and megafaunal diet. Nature **506**: 47–51.

WILSON, S. and K. MARTIN. 2008. Breeding habitat selection of sympatric White-tailed, Rock and Willow Ptarmigan in the southern Yukon Territory, Canada. Journal of Ornithology **149**: 629–637.

YOCCOZ, N. G., K. A. BRÅTHEN, L. GIELLY, J. HAILE, M. E. EDWARDS, *et al.*, 2012. DNA from soil mirrors plant taxonomic and growth form diversity. Molecular Ecology **21**: 3647–3655.

ZAZULA, G. D., D. G. FROESE, S. A. ELIAS, S. KUZMINA, C. L. FARGE, *et al.*, 2006. Vegetation buried under Dawson tephra (25,300 ^{14}C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada. Palaeogeography, Palaeoclimatology, Palaeoecology **242**: 253–286.

ZHOU, J., L. WU, Y. DENG, X. ZHI, Y.-H. JIANG, *et al.*, 2011. Reproducibility and quantitation of amplicon sequencing-based detection. The ISME Journal **5**: 1303–1313.

VI Supplementary Materials

Table S1: DNA Extractions with associated core data and dating measurements

Core Data	Years Before Present Carbon 14	Calibrated	Extract Type	ID
LL2S-60-F?	Modern PT	Modern	Cored	V1 V2
RD2-2011-185-G1	1955 PT	Modern	Homogenized	G1
RD2-2011-133-I1	1990 PT	Modern	Cored	Y1 Y2
RD2-2011-97-J1	Active layer	Modern	Cored	Z1 Z2
MM12-118b	8800	9820	Cored	GB1A GB1B GB1C
LL2S-189-E	9050	10220	Homogenized	E19A E19B E2
LL2S-253-D1	9195±20	10340	Cored	T1 T2
LL2C-118-C	11250±20	13105	Homogenized	C1 C2 C29
LL2C-205-B	11290±160	13150	Homogenized Cored	B1 S1 S2
LLII-12-84-3	11360±40	13205	Cored	LL3A LL3B LL3C LL3D
LL2C-243-A2	12000	13870	Homogenized	A2
LLII-12-170-6	12900	15405	Cored	LL2A LL2B LL2C LL2D
LLII-12-217-8	13200	15865	Cored	LL1A LL1B LL1C
MM12-116b	17500	21135	Cored	GB2A GB2B GB2C
BC 4-2B	25300	30000	Cored	BC1 BC2 BC3

PT refers to the permafrost table.

Table S2: Blanks and controls relevant to the amplified samples used in this study.

Blank ID	Type	Sample Reference	Source or Major Constituent	Test Result qPCR	Test Result Sequencing
ACR	Coring	A	Hood air and H2O	+	+
BCR	Coring	B	Hood air and H2O	-	na
CCR	Coring	C	Hood air and H2O	+	+
ECR	Coring	E	Hood air and H2O	+	+
GC	Coring	G	Hood air and H2O	-	na
KC	Coring	A-G	Closed Control	-	na
MC	Coring	A-G	Closed Dish	+	+
OC	Coring	A-G	Hood Lab Control	+	+
PC	Coring	A-G	Lab Control	-	na
QC	Coring	A-G	Drill Lab Control	-	na
BB	Amplification	Any	Reagents	+	+
NC1	Extraction	A-G	H2O	+	-
NC2	Extraction	A-G	H2O	+	-
NC3	Extraction	A-C	H2O	-	na
NC4	Extraction	A-C	H2O	-	na
CA5	Coring	A-C	Hood Air	-	na
CH5	Coring	A-C	Hood H2O	-	na
NC7	Extraction	A-E	H2O	+	-
NC8	Extraction	A-E	H2O	+	-
NC9	Extraction	G	H2O	-	na
NC10	Extraction	G	H2O	-	na
NC11	Coring	G	H2O	-	na
NC13	Extraction	S, T	H2O	+	-
NC14	Coring	S, T	Air	+	-
NC15	Extraction	V, Y, Z	H2O	-	na
NC16	Coring	V, Y, Z	Air	+	-
NC17	Extraction	V, Y, Z	H2O	+	-
NC20	Extraction	BC	H2O	-	na
NC21	Coring	BC, LL, GB	Air	+	+
NC22	Coring	BC, LL, GB	Air	-	na
NC23	Extraction	LL1, LL2, LL3	H2O	-	na
NC24	Extraction	GB1, GB2	H2O	+	+
NC25	Coring	LL1, LL2, LL3	Air	-	na
NC26	Coring	GB1, GB2	Air	+	+
LPB7	Library Prep	GB2, Y, Z, G1, V	H2O	-	-
LPB9	Library Prep	NC, CR Blanks, BC	H2O	-	-
LPB10	Library Prep	GB1, E, T, C, B1, S, MCR	H2O	-	-
LPB11	Library Prep	LL, OCR	H2O	-	-

Test results for qPCR and sequencing are positive (+), negative (-), and not applicable (na).

Table S3: Sequence data from Figure 1 showing total counts of plant loci in major functional community per calibrated year before present.

Category	30000	21135	15865	15405	13870	13205	13150	13105	10340	10220	9820	Modern
Forb	39120	87128	28199	82415	56153	15678	77930	23350	14413	286	5390	27295
Graminoid	29940	15012	9972	13735	705	74318	13656	45400	3040	8042	98	113126
Trees and shrubs	5321	3338	28	16	0	148639	4789	142097	46208	4904	71170	21690
Dwarf shrubs	2310	654	25	18	0	5165	2608	7511	5945	50571	13716	184063

Table S4: Parameters and sample script for the bioinformatic pipeline.

Software	Parameters
Cutadapt V1.2.1	Five consecutive rounds of Illumina adapter removal from both ends followed by one additional round with an error value of 0.2 `python cutadapt -b ACACGTCTGAACTCCAGTCACCAATTACATCTCGTATGCCGTCTTCTGCTTG RawFile.fastq > RawFile2.fastq && python cutadapt -b ATCTCGTATGCCGTCTTCTGCTTG RawFile2.fastq > RawFile3.fastq && python cutadapt -b AGATCGGAAGAGC RawFile3.fastq > RawFile4.fastq && python cutadapt -b AGATCGAAGAGC RawFile5.fastq > RawFile6.fastq && python cutadapt -e 0.2 -b AGATCGAAGAGC RawFile6.fastq > RawTrimmedForward.fastq`
SeqPrep V1.1	One merging alignment with a "minimum fraction of matching bases for primer/adapter overlap" (N) of 0.8, "maximum fraction of good quality mismatching bases to overlap reads" (m) of 0.1, "maximum fraction of good quality mismatching bases for primer/adapter overlap" (M) of 0.1 (ST JOHN, 2011) `SeqPrep -f RawTrimmedForward.fastq -r RawTrimmedReverse.fastq -1 Merged.output1.fastq.gz -2 Merged.output2.fastq.gz -3 Merged.discard1.fastq.gz -4 Merged.discard2.fastq.gz -N 0.8 -s MergedReads.fastq.gz -m 0.1 -M 0.1 -E MergedAlignments.fastq.gz -x 100000 -A ACACGTCTGAACTCCAGTCAC -B AAGAGCGTCGTGTAGGGAAAGAGTGT`
Cutadapt V1.2.1	Three additional consecutive rounds of adapter removal from one end followed by one additional round that trimmed bases with quality (q) scores less than 20 `python cutadapt -a AGATCGGAAGAG MergedReads.fastq > MergedReads2.fastq && python cutadapt -a AGATCGAAGAG MergedReads2.fastq > MergedReads3.fastq && python cutadapt -q 20 MergedReads3.fastq > MergedReadsTrimmed.fastq`
Usearch V6.0.307	Dereplicated the reads, recorded the original count in every line, and sorted the reads by size before performing a *de novo* chimera search and removal step. `Usearch6.0.307_i86linux32 -derep_fulllength MergedReadsTrimmed.fasta -sizeout -strand both -minseqlength 1 -output MergedReadsDereplicated.fasta Usearch6.0.307_i86linux32 -uchime_denovo MergedReadsDereplicated.fasta -nonchimeras MergedReadsDereplicated.notchimera.fasta -chimeras MergedReadsDereplicated.chimera.fasta`

Table S5: List of Species in Curated list -
Part 1: Full mitochondria and no mitochondria

Animals with a Full Mitochondrial Sequence		Animals with No Representation	
Common Name	*Scientific Name*	Common Name	*Scientific Name*
American beaver	*Castor canadensis*	Alaskan bison	*Bison alaskensis*
American marten	*Martes americana*	American scimitar cat	*Homotherium serum*
American mastodon	*Mammut americanum*	Deceit vole	*Microtus deceitensis*
American mink	*Mustela vison*	Deering vole	*Phenacomys deeringensis*
American moose	*Alces alces*	Eskimo curlew	*Numenius borealis*
American pika	*Ochotona princeps*	Flat-headed peccary	*Platygonus compressus*
Badger	*Taxidea taxus*	Giant beaver	*Castoroides ohioensis*
Black bear	*Ursus americanus*	Giant moose	*Alces latifrons*
Brown bear	*Ursus arctos*	Giant pika	*Ochotona whartoni*
Caribou	*Rangifer tarandus*	Helmeted muskox	*Bootherium bombifrons*
Collared pika	*Ochotona collaris*	Hopkins' lemming	*Predicrostonyx hopkinsi*
Common teal	*Anas crecca*	Jackson's weasel	*Mustela jacksoni*
Coyote	*Canis latrans*	Matthews' vole	*Guildayomys matthewsi*
Fisher	*Martes pennanti*	Morlan's vole	*Microtus morlani*
Giant short-faced bear	*Arctodus simus*	Noble marten	*Martes nobilis*
House mouse	*Mus musculus*	Northern collared lemming	*Dicrostonyx kilangmiutak*
Least weasel	*Mustela nivalis*	Ogilvie mountain collared lemming	*Dicrostonyx nunatakensis*
Lion	*Panthera leo*		
Mallard	*Anas platyrhynchos*	Plains shrew	*Planisorex dixonensis*
Mountain lion	*Felis concolor*	Primitive vole	*Allophaiomys deceitensis*
Mountain sheep	*Ovis canadensis*	Scott's horse	*Equus scotti*
Mule deer	*Odocoileus hemionus*	Short-faced skunk	*Mephitis obtusata*
Polar bear	*Ursus maritimus*	Soergel's muskox	*Soergelia mayfieldi*
Red deer	*Cervus elaphus*	Southern mammoth	*Mammuthus meridionalis*
Red fox	*Vulpes vulpes*	Staudinger's muskox	*Praeovibos priscus*
Red-throated loon	*Gavia stellata*	Steppe bison	*Bison priscus*
Saiga	*Saiga tatarica*	Steppe mammoth	*Mammuthus trogontherii*
Striped skunk	*Mephitis mephitis*	Western camel	*Camelops hesternus*
Tundra muskox	*Ovibos moschatus*	Xenocyon	*Xenocyon lycaonoides*
Tundra swan	*Cygnus columbianus*	Yukon horse	*Equus lambei*
Walrus	*Odobenus rosmarus*		
Western bison	*Bison bison occidentalis*		
Wolf	*Canis lupus*		
Wolverine	*Gulo gulo*		
Woolly mammoth	*Mammuthus primigenius*		

Table S6: List of Species in Curated List -
Part 2: Partial mitochondria listed by gene presence (present = y, absent = n)

Animals with Limited or Selected Mitochondrial Representation				
Common Name	Scientific Name	cyt b	12S	16S
American golden plover	*Pluvialis dominica*	n	y	y
American water shrew	*Sorex palustris*	y	y	y
American widgeon	*Anas americana*	y	y	n
Arctic fox	*Alopex lagopus*	y	y	y
Arctic ground squirrel	*Spermophilus parryii*	y	n	n
Arctic hare	*Lepus arcticus*	y	n	y
Arctic lemming	*Dicrostonyx torquatus*	y	n	n
Arctic loon	*Gavia arctica*	y	y	n
Arctic shrew	*Sorex arcticus*	y	n	n
Barrenground shrew	*Sorex ugyunak*	y	n	n
Big brown bat	*Eptesicus fuscus*	y	y	y
Black -bellied plover	*Pluvialis squatarola*	y	y	n
Black-footed ferret	*Mustela nigripes*	y	n	n
Blue-winged teal	*Anas discors*	y	y	n
Boreal chorus frog	*Pseudacris maculata*	y	y	y
Brown lemming	*Lemmus sibiricus*	y	n	n
Bushy-tailed woodrat	*Neotoma cinerea*	y	y	y
Columbia spotted frog	*Rana luteiventris*	y	y	y
Common eider	*Somateria mollissima*	y	n	n
Common loon	*Gavia immer*	y	y	y
Common raven	*Corvus corax*	y	n	y
Common scoter	*Melanitta nigra*	y	n	n
Dall sheep	*Ovis dalli*	y	y	y
Deer mouse	*Peromyscus maniculatus*	y	y	n
Dusky shrew	*Sorex monticolus*	y	n	n
Ermine	*Mustela erminea*	y	y	n
Harlequin duck	*Histrionicus histrionicus*	y	y	n
Heather vole	*Phenacomys intermedius*	y	n	n
Hermit thrush	*Catharus guttatus*	y	y	y
Hoary marmot	*Marmota caligata*	y	n	n
Jefferson's ground sloth	*Megalonyx jeffersonii*	y	n	n
Keen's mouse	*Peromyscus keeni*	y	y	n
Least chipmunk	*Tamias minimus*	y	n	y
Lesser scaup	*Aythya affinis*	y	n	n
Lincoln's sparrow	*Melospiza lincolnii*	y	n	n

Continued on next page

Table S6 – continued from previous page

Animals with Limited or Selected Mitochrondrial Representation				
Common Name	Scientific Name	cyt b	12S	16S
Long-tailed vole	*Microtus longicaudus*	y	n	n
Long-toed salamander	*Ambystoma macrodactylum*	y	n	n
Lynx	*Felis canadensis*	y	y	y
Masked shrew	*Sorex cinereus*	y	y	y
Meadow jumping mouse	*Zapus hudsonius*	y	n	n
Meadow vole	*Microtus pennsylvanicus*	y	y	n
Muskrat	*Ondatra zibethicus*	y	y	y
North American porcupine	*Erethizon dorsatum*	y	y	y
Northern bog lemming	*Synaptomys borealis*	y	n	n
Northern collared lemming	*Dicrostonyx groenlandicus*	y	n	y
Northern flying squirrel	*Glaucomys sabrinus*	y	n	n
Northern long-eared bat	*Myotis septentrionalis*	y	y	y
Northwestern salamander	*Ambystoma gracile*	y	n	n
Oldsquaw	*Clangula hyemalis*	y	n	n
Olive-sided flycatcher	*Contopus cooperi*	y	n	n
Otter	*Lontra canadensis*	y	y	y
Pintail	*Anas acuta*	y	y	n
Pygmy shrew	*Sorex hoyi*	y	n	n
Red squirrel	*Tamiasciurus hudsonicus*	y	y	y
Red-backed vole	*Clethrionomys rutilus*	y	n	n
Red-tailed hawk	*Buteo jamaicensis*	y	y	y
Richardson's collared lemming	*Dicrostonyx richardsoni*	y	n	n
Robin	*Turdus migratorius*	y	y	y
Rock ptarmigan	*Lagopus muta*	y	y	y
Say's phoebe	*Sayornis saya*	y	n	n
Singing vole	*Microtus miurus*	y	n	n
Snow bunting	*Plectrophenax nivalis*	y	y	n
Snow goose	*Chen caerulescens*	y	n	n
Snowshoe hare	*Lepus americanus*	y	y	y
Solitary sandpiper	*Tringa solitaria*	y	y	y
Southern red-backed vole	*Myodes gapperi*	y	n	n
Steppe lion	*Panthera leo spelaea*	y	n	n
Surf scoter	*Melanitta perspicillata*	y	n	n
Tree or chipping sparrow	*Spizella arborea*	y	n	n
Trumpeter swan	*Cygnus buccinator*	n	y	n
Tundra shrew	*Sorex tundrensis*	y	n	n

Continued on next page

Table S6 – continued from previous page

Animals with Limited or Selected Mitochrondrial Representation				
Common Name	Scientific Name	cyt b	12S	16S
Tundra vole	*Microtus oeconomus*	y	y	n
Wapiti	*Cervus canadensis*	y	y	n
Western toad	*Anaxyrus boreas*	y	y	y
White-footed mouse	*Peromyscus leucopus*	y	y	y
White-tailed ptarmigan	*Lagopus leucura*	y	y	y
Willow ptarmigan	*Lagopus lagopus*	y	y	y
Wood frog	*Lithobates sylvaticus*	y	y	y
Woodchuck	*Marmota monax*	y	y	y
Yellow-billed loon	*Gavia adamsii*	y	y	n
Yellow-cheeked vole	*Microtus xanthognathus*	y	n	n

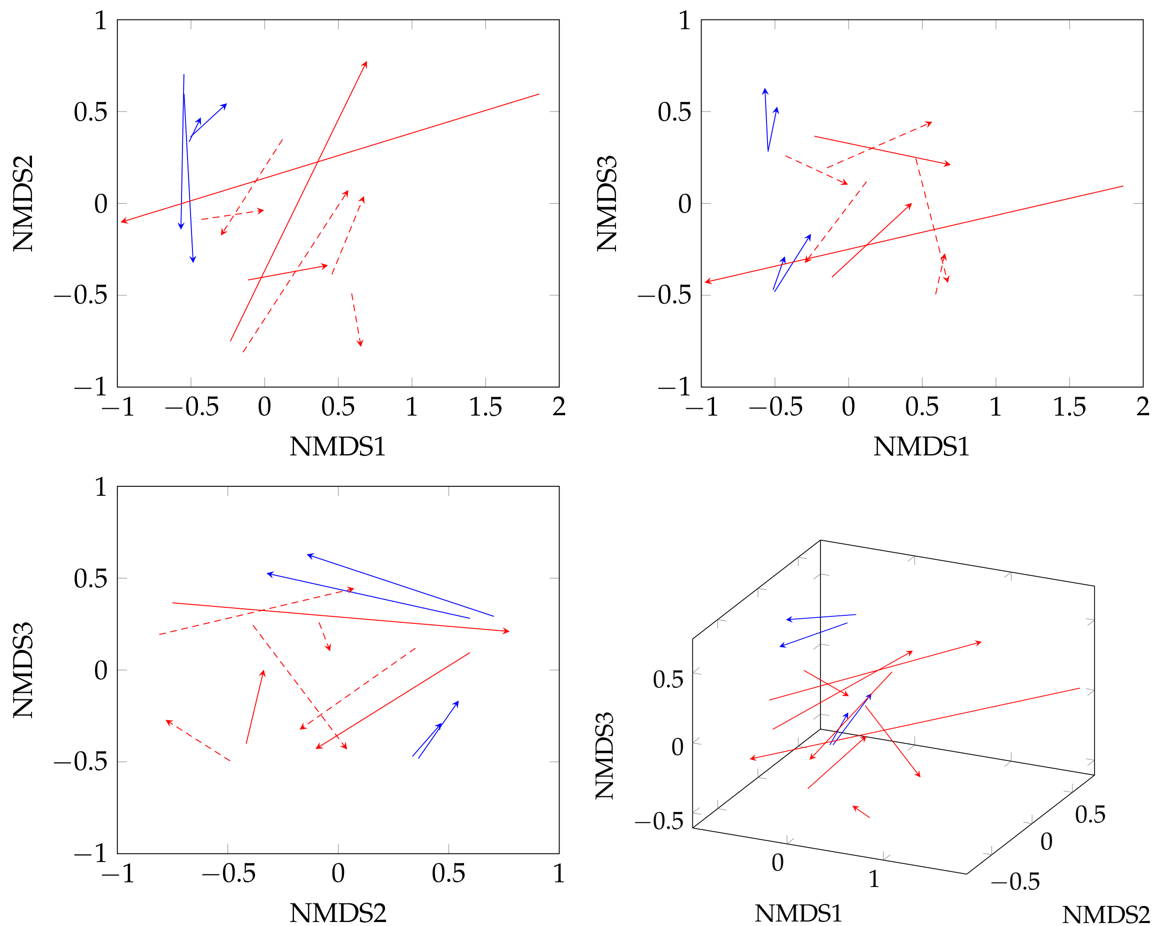
Table S7: Primers for ampliconic loci amplification

Gene	Primer Name	Sequence	Citation
GFP	GFP_F	`TGATGCAACATACGGAAAACTTACCCTTA`	(D'COSTA *et al.*, 2011)
	GFP_R	`CACGTGTCTTGTAGTTCCCGTCATCT`	(D'COSTA *et al.*, 2011)
rbcL	H1a_F	`GGCAGCATTCCGAGTAACTCCTC`	(WILLERSLEV *et al.*, 2003)
	H2a_R	`CGTCCTTTGTAACGATCAAG`	(WILLERSLEV *et al.*, 2003)
trnL	Primer g 49425	`GGGCAATCCTGAGCCAA`	(TABERLET *et al.*, 2007)
	Primer h 49466	`CCATTGAGTCTCTGCACCTATC`	(TABERLET *et al.*, 2007)
16S	16S6 L02828	`TTTCGGTTGGGGCGACCTCGGAG`	(HÖSS *et al.*, 1996)
	16S7 H02922	`TTGCGCTGTTATCCCTAGGGTAACT`	(HÖSS *et al.*, 1996)
	16SA&M FV2	`TCACTATTTTGCNACATAGA`	(RASMUSSEN *et al.*, 2009)
	16SA&M RV2	`CCCCGAAACCAGACGAGCTA`	(RASMUSSEN *et al.*, 2009)
12S	12Sa'	`CTGGGATTAGATACCCCACTAT`	(KUCH *et al.*, 2002)
	12So	`GTCGATTATAGGACAGGTTCCTCTA`	(KUCH *et al.*, 2002)
cyt b	cytb F1	`ACACTACACNTCAGACACAACAACAGC`	This paper
	cytb R2	`ANCCGTAGTTTACGTCTCGGCAG`	This paper
	cytb R3	`CANCCGTAGTTTACGTCTCGGCA`	This paper

The sequencing results were initially aligned with BLAST against the entire nucleotide database, but this yielded false negatives and reduced the ability to assign taxa to the most specific rank. The false negatives were determined through manual inspection of the BLAST output and included the assignment of certain sequences to taxa not present in the Arctic region. Thus, the data were re-identified with a constrained list of Arctic plants and animals. Figure S1 shows the before and after

coordinates, in 2-dimensions for the first three plots and in 3-dimensions for the fourth plot. The abnormally long arrow represents a low extract number sample (n=1) at 13.87k cal yrs BP.

Figure S1: Change in relative ecological position after constraining the datasets. The extremely long arrow in the first two NMDS plots is the 13870 cal yrs BP sample (n=1)



Chapter 2

Catching a Break: Targeted DNA Enrichment of the Holocene Epoch

Tara Sadoway
McMaster University
tsadoway@gmail.com

Abstract

Sedimentary DNA can be used to describe organismal diversity and distributions in modern ecological contexts, but has only been selectively applied to ancient ecosystems, typically through single gene, PCR-based methods. Here we propose a system to enrich for specific organisms to overcome limitations and biases inherent in a single gene PCR-based screening of sedimentary DNA extracts through the use of oligonucleotide baits. In particular, DNA extracts from 30k cal yrs BP to the present were enriched for the mammalian target loci of 12S rRNA, cytochrome oxidase 1, and cytochrome b. The results were not the expected sequences in the theoretically predicted abundances, but were still relatively informative for each sample.

I Introduction

Ancient DNA (aDNA) has been instrumental in elucidating the identity of macrofossil remains and evolutionary relationships between extinct and extant species. DNA can confirm the identity of a fossil or organic relict through targeted amplification of informative regions of the genome (KUCH *et al.*, 2002) or can be deeply sequenced and bioinformatically assembled to recreate extinct mitochondrial genomes (SHAPIRO *et al.*, 2004). While many fossils contain DNA, only some fossils are well preserved enough to yield useful quantities and qualities of DNA for analysis. As well, a fossil may contain primarily exogenous DNA and be sparse in endogenous DNA content. However, these limitations can be overcome with oligonucleotide bait enrichment, a technique that uses complementary RNA sequences as 'baits' to catch DNA regions of interest. In this study, a specially designed bait set was used to enrich the mammalian DNA content from permafrost cores extracts that are rich in bacterial DNA (BARTRAM *et al.*, 2011; YERGEAU *et al.*, 2010; STEVEN *et al.*, 2007; MACKELPRANG *et al.*, 2011). This putatively reduces the bias towards bacterial amplicons, while simultaneously allowing the study of a low copy number target (MAMANOVA *et al.*, 2010; HORN, 2012).

The enrichment process has been very successful for single organism targets (CHOU *et al.*, 2010; CARPENTER *et al.*, 2013). The causative agent of the Black Death was confirmed using this method (BOS *et al.*, 2011) and the endogenous constituent of mammoth mitochondrial genomes has been enriched by an order of magnitude from bone samples (ENK *et al.*, 2013). Microchip arrays, such as the GeoChip (HE *et al.*, 2007; HE *et al.*, 2010), can identify 10,000 different gene or functional targets from metagenomic samples. Thus, it seems likely that targeting a specific and rare group of organisms in a metagenomic sample should be possible, even for the small, degraded DNA molecules typically found in aDNA extracts (LYNCH *et al.*, 2012). Sample extracts that were previously characterized and quantified using ampliconic data were reprocessed and enriched for mammalian target loci. Three genetic loci were chosen for this study, to minimize both the bias from a single locus (WAUGH, 2007; PIGANEAU *et al.*, 2011) and to better represent the current reference sequences in GenBank. The loci are all mitochondrial, which are expected to be in higher copy number than nuclear loci. Two of the genetic loci were previously amplified from the samples, including 12S rRNA and cytochrome b (*cyt b*), while the third was chosen for barcoding purposes, cytochrome c oxidase subunit 1 (*cox1*)(HEBERT *et al.*, 2003). The reference sequences for these loci were downloaded from GenBank to account for more than 100 vertebrates relevant to the study site. Each sequence was tiled (divided up) at 10 base pairs (bp) (CHOU *et al.*, 2010) and clustered to remove duplicates and highly similar sequences (99% or higher). In theory, these RNA

baits have a tolerance of up to 15% mismatch, but work better above a mismatch threshold of 95% (He *et al.*, 2005). To prevent low-energy hybridization and increase binding stability (Herwig *et al.*, 2000), bait sets were screened to remove all baits with less than 40% GC content (Mamanova *et al.*, 2010) or repetitive, weak binding regions (ATATATAT) (Wernersson *et al.*, 2007). In solution baits were created by MYcroarray and processed using the standard work flow. Two modern samples, one amphibian (*Xenopus borealis*) and one mammal (*Loxodonta* sp.), were enriched to assess the ability of the baits to hybridize in a mono-vertebrate sample.

In the first round of enrichment, baits were enriched in two steps, the first at 50°C and the second at 55°C. At the time they were performed, these enrichments appeared to fail due to a streptavidin, magnetic-bead-binding issue. Some of these samples were sequenced as a comparison. Two separate and complete rounds of enrichments were conducted at 60°C and 50°C with new reagents. As in PCR-based methods, a higher temperature for annealing should create a more stringent hybridization (Tiquia *et al.*, 2004). We expect to capture a larger fraction of vertebrates at 60°C than 50°C and we predict a larger number of total sequences in the 50°C hybridization reaction. All samples were quantified by qPCR (quantitative polymerase chain reaction) and then pooled. Pools were quantified by an additional round of qPCR and a high sensitivity Bioanalyzer run before sequencing.

We saw the expected large reduction in the bacterial DNA constituent of the permafrost core DNA extracts, however, most of the identifiable DNA in the metagenomic samples was still bacterial. We also expected a much larger vertebrate fraction of sequences, reflecting the composition of the bait set, than was found. Due to the naturally low ratio of vertebrate to bacterial DNA in these extracts, it is possible that the vertebrate hybridization was as efficient as it could be, and the results reflect the best possible output of the enrichment. Interestingly, the GC content of the samples does reflect the constructed parameters of the *in silico* baits, putatively supporting the hypothesis that the oligonucleotide baits worked, but the samples were simply too sparse in vertebrate DNA to enrich any further. We will discuss the ways in which the baits worked and possible reasons why they might have failed.

II Methods

i Sample Extraction and Library Preparation

DNA was extracted from sedimentary permafrost cores from the Yukon Territory, Canada. The raw soil cores were obtained from exposed stratigraphies on gold

mining sites near the city of Dawson (64°03’36"N, 139°24’39"W). All sample sites were identified in relation to volcanic tephra (FROESE *et al.*, 2006; DEMURO *et al.*, 2008) and cored with attention to contamination protocols (COOPER and POINAR, 2000). Frozen cores were subsampled either at the University of Alberta (homogenized samples) or at the McMaster Ancient DNA facility (as per D’COSTA *et al.*, 2011)(Supplementary Materials (SM), Table S1). All samples were manipulated in the presence of air and TE (0.1X Tris-EDTA) blanks. Guanidinium protocols (BOOM *et al.*, 1990; D’COSTA *et al.*, 2011) were used to extract the DNA from 250mg of core soil, then the samples were silica purified (HÖSS and PÄÄBO, 1993; D’COSTA *et al.*, 2011), and twice eluted from the silica pellet (HANDT *et al.*, 1996). The sample LL3D was extracted with a PowerSoil Kit (MoBio, Carlsbad, CA, USA). The frog and elephant samples were extracted with the DNeasy Blood & Tissue kit (Qiagen, Toronto, ON) and the DNA extracts were sheared with a Covaris sonicator to 150bp (175W, duty factor 10%, cycles per burst 200, time 430 seconds)(Covaris, Inc, MA, USA). External contamination, detected by a topically applied green fluorescent protein (GFP) DNA construct on the outside of the permafrost cores, was minimal (sample A2, <10 copies) or non-existent (all other samples). Although the undiluted DNA extracts were too inhibited for PCR, this did not affect library preparation. We determined that the library preparation of undiluted samples was not a limiting factor by processing an aliquot of BC1 on the GeoChip microarray; all data (not shown) were within normal parameters.

Library preparations were made from the straight DNA extracts (MEYER and KIRCHER, 2010); however, some samples were not UDG (uracil-DNA glycosylase) repaired as to maintain the damage pattern unique to ancient DNA (BRIGGS *et al.*, 2010). Most samples were UDG-treated because the oligonucleotide baits were based on undamaged reference sequences. The expected DNA damage patterns are qualified by an increase in damage at the ends of the sequence and an increased frequency of cytosine deamination (BRIGGS *et al.*, 2010).

All samples were uniquely dual indexed with standard Illumina protocols (KIRCHER *et al.*, 2012) and amplified with Herculase II Fusion DNA polymerase (Agilent Technologies Inc, CA, USA) for 12 cycles. The amplifications were silica column purified with MinElute PCR Purification columns (Qiagen, Toronto, ON) and eluted in 15µL of elution buffer (EB).

ii Oligonucleotide Bait Design

The RNA based oligonucleotide baits were designed *in silico*. Three sets of mitochondrial loci sequences (12S rRNA, *cytb*, and *cox1*) were retrieved from GenBank

for more than 100 North American vertebrates. The sequences were at least 500 base pairs long and did not include any degeneracy. The sequences were divided into 100 kmer lengths at 10 base pair intervals and then combined into sets. Each locus set was filtered of any baits with less than 40% GC content and cleared of any baits with AT runs longer than 8 base pairs to increase binding strength. Each bait set was tested bioinformatically by sampling baits (without replacement) and using BLAST (ALTSCHUL *et al.*, 1990) and MEGAN (HUSON *et al.*, 2011) to assess the number of original target species detected; the number of species detected of the original number is represented as the putative efficacy. While clustering the samples to remove redundancy, it was determined that further clustering (beyond 99%) did not significantly change the theoretical efficacy of the bait sets. Thus, all bait sets were clustered at 99% sequence similarity with Usearch 6.0.307 (EDGAR, 2010). Table 1 shows the composition and theoretical performance of each final bait set.

Table 1: The oligonucleotide bait sets for vertebrate enrichment were tested *in silico* (without replacement) for the theoretical recovery of target species.

	Number of			Theoretical
Locus	Genera	GenBank sequences	Baits	Recovery
12S rRNA	95	125	570	95%
Cytochrome b	105	163	1265	99%
Cytochrome oxidase 1	112	172	975	96%

iii Enrichment and Sequencing

All three desired loci for the enrichment bait set were created by MYcroarray in a single aliquot. All of the enrichments were performed in duplicate (to reduce stochastic effects) as per standard MYcroarray protocols, except as noted. There were three different hybridization temperature experiments. The first set of enrichments (labelled as 50/55) were enriched at 50°C hybridization temperature for 40 hours, reamplified (as before, 12 cycles), and enriched again at 55°C; the increased hybridization temperature should have theoretically increased specificity in the second round. The initial batch of enrichments appeared to have failed due to defective streptavidin beads, but may have failed due to the sample type or other factors that will be discussed later. Due to the low copy number and low change in the frequency of expected ampliconic products (as determined by qPCR of both total quantification and *cytb*), two further, but separate rounds of enrichment were carried out. Starting again from the indexed, library-prepared extract, the first

additional enrichment was a single round at 50°C, but it didn't appear to increase the desired product (based on qPCR quantification). A separate enrichment using modern, sonicated elephant (*Loxodonta* sp.) and frog (*X. borealis*) DNA was conducted at four temperatures (50°C, 55°C, 60°C, and 65°C) to examine the trade-off between enrichment specificity and total molecules recovered with this particular bait set. Since the 60°C enrichment of these modern specimens seemed to be the best compromise, an additional set of ancient extracts were enriched at this temperature. All samples and treatments that were sequenced are shown in Table 2.

The enrichments were reamplified for 12 cycles with Herculase II Taq Polymerase, as before, and re-quantified with a total molecule qPCR count. The samples were then combined in five sets and gel size selected with MinElute Gel Extraction kits (Qiagen, Toronto, ON) for a final volume of 20μL. These extract pools were quantified with KAPA DNA quantification and pooled in equimolar concentrations after an Agilent High Sensitivity DNA analysis with a Bioanalyzer (Agilent Technologies Inc, CA, USA). The sequences were run on a HiSeq 2500 Rapid Run (2x80bp) at approximately 100,000 reads per each sample (Illumina, San Diego, CA, USA).

iv Data Analysis

The data were demultiplexed both by Illumina's proprietary software and custom Perl scripts that detect the imprecise binning of short sequences. Since each sample has a unique combination of forward and reverse reads, the Perl scripts detect and remove any sequences that are so short as to still show the index in the read (alternately, reads that have two index adapters attached). Table 3 shows the forward and reverse indexes for each sample and the number of mis-binned reads that were detected in read 1 or read 2. In the modern samples, this should be less frequent because most reads are longer than the upper limit of the sequencing run; modern sequences are sheared to specific lengths. The improper index may also be an artifact of the library preparation process, where two adapters may have ligated onto the same end of a molecule or it may be an copying error during PCR or cluster generation (FAIRCLOTH and GLENN, 2012). We do not see a clear pattern between index sequence parameters (GC content, etc.) and the number of mis-binned sequences.

Sequencing adapters with trimmed with CutAdapt V1.2.1(MARTIN, 2011). Forward and reverse reads were merged with SeqPrep (ST JOHN, 2011) and merged reads were dereplicated and chimera-checked with Usearch V6.0.307 (EDGAR, 2010). Chimeras were only detected in the frog sample. Reads were compared to the nucleotide database with BLAST (BLASTN) using the default parameters and

Table 2: Enrichment samples are listed by name and calibrated date in years before present for UDG treatment, number of enrichment rounds, and hybridization temperature.

Sample Name	Calibrated Date	UDG	Rounds	Temperature(°C)
Elephant	Modern	Yes	1	60
Frog	Modern	Yes	1	60
LPB1	Modern	Yes	1	60
LPB3	Modern	Yes	1	60
LPB5	Modern	Yes	1	60
LPB12	Modern	Yes	1	60
V2	Modern	Yes	1	60
GB1C	9820	Yes	1	60
E19A	10220	Yes	1	60
T2	10340	Yes	1	60
A2	13105	Yes	1	60
S1	13150	Yes	1	50
S2	13150	Yes	1	60
LL3A	13205	Yes	1	60
LL3B	13205	Yes	1	50
LL3D	13205	Yes	1	60
LL2A	15405	Yes	1	60
LL2C	15405	Yes	1	50
LL1A	15865	No	1	60
LL1A	15865	No	2	50/55
LL1C	15865	Yes	2	50/55
LL1C	15865	No	2	50/55
GB2A	21135	Yes	1	60
GB2B	21135	Yes	2	50/55
BC1	30000	Yes	2	50/55
BC7	30000	Yes	1	60

LPB = (Library Preparation Blanks)

Table 3: Number of mis-binned sequences per each sequencing direction.

Sample	Temp (°C)	Index Forward	Reverse	Removed Sequences Forward	Reverse	Percent Mis-binned[1]
A2	60	CTGCGCG	ACGCAAC	7	43	0.3
BC1	50/55	CTGCGCG	AGAACCG	87	542	0.2
BC7	60	CTGCGCG	AACTAGA	11	29	0.1
E19A	60	CTGCGCG	GGATCAA	2	22	0.3
GB1C	60	CTGCGCG	CCTAGGT	2	11	0.5
GB2A	60	CTGCGCG	CAGTACT	151	21	1.8
GB2B	50/55	CTGCGCG	CATCCGG	10	212	2.0
LL1A*	60	CTGCGCG	CAATTAC	7	18	0.4
LL1A*	50/55	CTTGACC	GCAAGAT	1	52	1.2
LL1C*	50/55	CTGCGCG	ATGCCGC	2	2	0.3
LL1C	50/55	AGACCTT	GCTCGAA	0	1	0.0
LL2A	60	CTGCGCG	GCATTGG	1	4	0.3
LL2C	50	CTGCGCG	TGACGTC	26	100	0.2
LL3A	60	CTGCGCG	AGGTACC	2	19	1.9
LL3B	50	CTGCGCG	TGCGTCC	25	170	0.2
LL3D	60	CTGCGCG	CATGCTC	4	16	2.4
LPB1	60	CTGCGCG	GCAGTCC	2	0	2.8
LPB3	60	CTGCGCG	AACCAAG	1	0	1.4
LPB5	60	CTGCGCG	CTATGGC	0	0	3.1
LPB12	60	CTGCGCG	CGACGGT	1	0	0.0
Elephant	60	ACTTCAA	AATTCAA	11	32	9.0
Frog	60	ACTTCAA	CGCGCAG	63	4083	2.4
S1	50	CTGCGCG	CGACCTG	81	321	0.5
S2	60	CTGCGCG	AGCAGGT	3	12	0.4
T2	60	ACCATGA	AATAGTA	6	9	0.3
V2	60	CATATTG	TTGGATC	14	9	0.1
					Average	1.2%

* Non UDG treated samples
1. Percent mis-binned = number of mis-binned reads / number of raw reads x 100.

taxonomically assigned with MEGAN4 V4.70.4 (Min Support 1, Min Score 50, Top Percent 10, Min Complexity 0.05) (HUSON *et al.*, 2011). Reads were mapped back to their original sequence counts before analysis.

III Results

i Initial Quantitative Observations

Despite extensive molecular quantitation before sequencing, fewer sequences resulted than expected except in two samples (Table 4). Approximately 100,000 reads were expected for each sample (fewer for the library prep blanks and the modern elephant and frog); however, only BC1 (Bear Creek, 30k cal yrs BP) had more than 100,000, while the next highest ancient sample, LL3B (13205 cal yrs BP), had 52,000 reads, just over half of what was expected. Samples from the first set of enrichments at 50°C/55°C had the most reads because of the BC1 sample, but theoretically, the 50/55°C should have slightly fewer reads than the solely 50°C enrichment because a small loss is expected in the purification steps between the two rounds and 55°C is a more stringent hybridization. Overall, 9.6% of expected read counts were obtained at 60°C, 35.2% at 50°C/55°C (or 0.7% without BC1), and 41.1% at 50°C.

Most of the ancient samples have only 10-15% of reads assigned at the broadest taxonomic rank of kingdom. The modern, non-metagenomic samples including the elephant and the frog, have 20% and 30% assigned, respectively; however, it should be noted that the elephant sample has only 65 complete reads. Table 5 shows the low proportion of reads that could be taxonomically identified.

As expected, the greater the number of reads that were sequenced, the greater the number of reads that could be assigned at the most inclusive rank 'kingdom' (Figure 1). This pattern is expected, but our sampling curve was not saturated due to the low number of reads sequenced, so we do not see a plateau in the data.

To determine if the taxonomically assigned sequences were large clusters of replicated amplicons or unique sequences, we plotted the total number of each, as shown in Figure 2. Regardless whether a sample is metagenomic or not, the more assigned sequences in a sample, the more of those were unique. This result rejects the hypothesis that there may be large clusters of amplicon bias from the small amplification steps before and after enrichment.

Table 4: Enrichment sequence read counts per data sorting step. Reads were merged with SeqPrep (St John, 2011) and dereplicated (shown as unique) with Usearch (Edgar, 2010).

Sample	Temp (°C)	Expected	Initial	Merged	Unique	% Unique[1]
A2	60	100,000	8,972	7,700	4,839	62.8
BC1	50/55	100,000	165,534	133,772	125,717	94.0
BC7	60	100,000	16,467	11,234	8,874	79.0
E19A	60	100,000	3,740	3,218	2,056	63.9
GB1C	60	100,000	1,397	1,056	722	68.4
GB2A	60	100,000	4,427	3,472	2,823	81.3
GB2B	50/55	100,000	5,642	4,442	3,849	86.7
LL1A*	60	100,000	2,140	1,557	1,465	94.1
LL1A*	50/55	100,000	2,818	2,232	1,856	83.2
LL1C*	50/55	100,000	1,463	1,076	1,008	93.7
LL1C	50/55	100,000	606	429	406	94.6
LL2A	60	100,000	825	710	587	82.7
LL2C	50	100,000	31,973	24,844	21,958	88.4
LL3A	60	100,000	517	378	327	86.5
LL3B	50	100,000	52,022	40,820	36,992	90.6
LL3D	60	100,000	411	294	250	85.0
LPB1	60	41,667	36	19	19	100.0
LPB3	60	41,667	16	12	10	83.3
LPB5	60	41,667	7	1	1	100.0
LPB12	60	41,667	36	12	11	91.7
Elephant	60	41,667	238	65	60	92.3
Frog	60	41,667	84,223	61,437	49,663	80.8
S1	50	100,000	39,393	29,938	27,020	90.3
S2	60	100,000	1,757	1,366	1,094	80.1
T2	60	100,000	2,005	1,607	802	49.9
V2	60	100,000	11,925	8,510	4,475	52.6

* Non UDG treated samples

1. Percent unique = number of unique reads / number of merged reads x 100.

Table 5: Number of assigned reads at each taxonomic rank per sample and treatment.

Sample	Temp (°C)	Species	Genus	Family	Order	Class	Phylum	Kingdom	% Assigned[1]
A2	60	292	354	499	557	592	624	688	8.94
BC1	50/55	3683	4410	5081	6228	6764	7220	8122	6.07
BC7	60	604	709	800	1045	1150	1262	1487	13.24
E19A	60	165	191	230	252	273	288	315	9.79
Elephant	60	7	8	9	11	13	13	13	20
Frog	60	14614	18147	18153	18162	18202	18515	18540	30.18
GB1C	60	96	110	121	129	136	140	150	14.2
GB2A	60	210	254	337	376	400	441	466	13.42
GB2B	50/55	369	404	493	546	592	628	640	14.41
LL1A*	60	110	132	198	226	231	240	256	16.44
LL1A*	50/55	130	159	237	275	286	296	313	14.02
LL1C*	50/55	70	87	97	126	130	137	149	13.85
LL1C	50/55	25	29	41	53	55	58	65	15.15
LL2A	60	30	36	47	59	67	78	86	12.11
LL2C	50	1604	1958	2599	2997	3174	3307	3647	14.68
LL3A	60	25	28	34	39	42	49	55	14.55
LL3B	50	2574	3007	3463	4236	4857	5423	5839	14.3
LL3D	60	18	20	24	27	29	34	37	12.59
LPB1	60	1	1	2	4	4	4	4	21.05
LPB3	60	1	1	1	1	1	3	3	25
LPB5	60	0	0	0	0	0	0	0	0
LPB12	60	0	0	0	0	0	0	0	0
S1	50	1633	1895	2230	2902	3099	3257	3654	12.21
S2	60	90	111	126	161	175	186	202	14.79
T2	60	97	111	122	158	172	190	210	13.07
V2	60	248	279	357	422	454	502	590	6.93

* Non UDG treated samples

1. Percent assigned = number of assigned reads at kingdom rank / number of merged reads x 100.

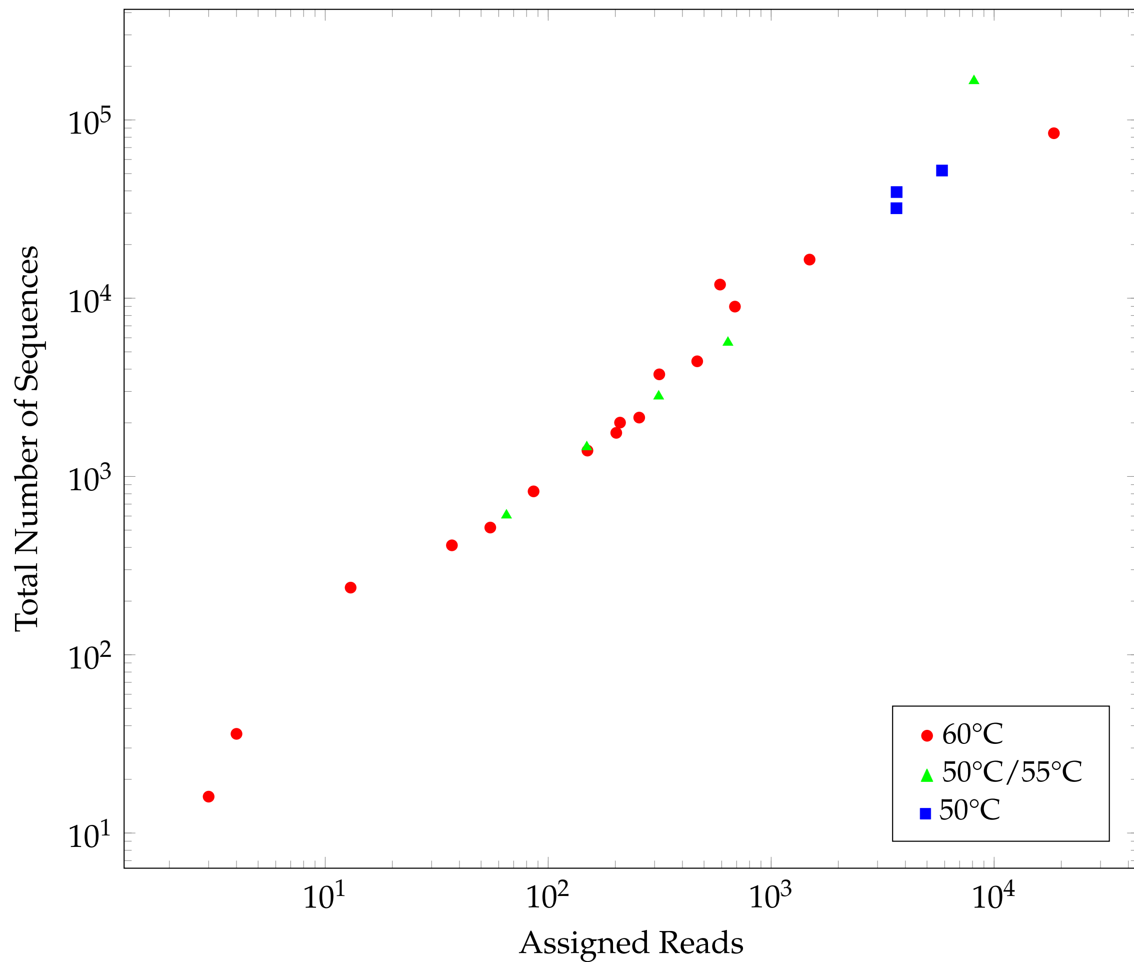


Figure 1: The total number of sequencing reads increases proportionally to the total number of assigned reads at the kingdom rank. This trend occurs independently of temperature. Data are shown on a log-log axis.

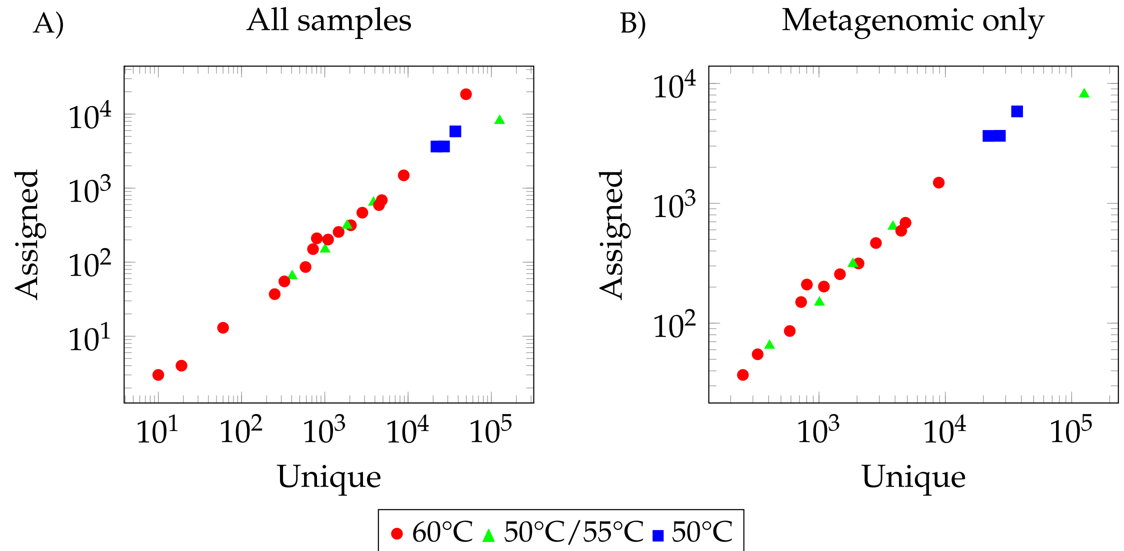


Figure 2: Assigned reads correlated to the number of unique sequences for a) all samples b) only non-blank, metagenomic samples. Data are shown on a log-log axis.

ii Hybridization Temperature Patterns

To assess the effect of hybridization temperature on the specificity of the enrichment, Figure 3 (Part A) shows the total number of raw reads obtained at each temperature experiment. As expected, the samples hybridized at the lowest temperature (50°C) have the highest number of reads, however, both other temperature experiments have samples in the same range. The four lowest samples at 60°C (two have 36 initial reads) are the library prep blanks; the actual soil samples have an order of magnitude more reads. In Part B of Figure 3, the average GC content for each sample is shown by hybridization temperature. The GC content for all three hybridization temperatures appears to be in the same range, except for three samples at 60°C. These three samples were extracted with kits, instead of with the homebrew extraction protocols used for the rest of the metagenomic samples. Because sample LL3D, but not its core replicates, LL3A and LL3B, is in this low GC grouping, it appears that the extraction procedure is important to the recovered GC composition in a yet-unknown mechanism. In reference to previous amplicon work, the LL3D sample also showed a different plant and vertebrate metagenetic profile than LL3A/B, indicating that ancient DNA work must be cognizant of this bias.

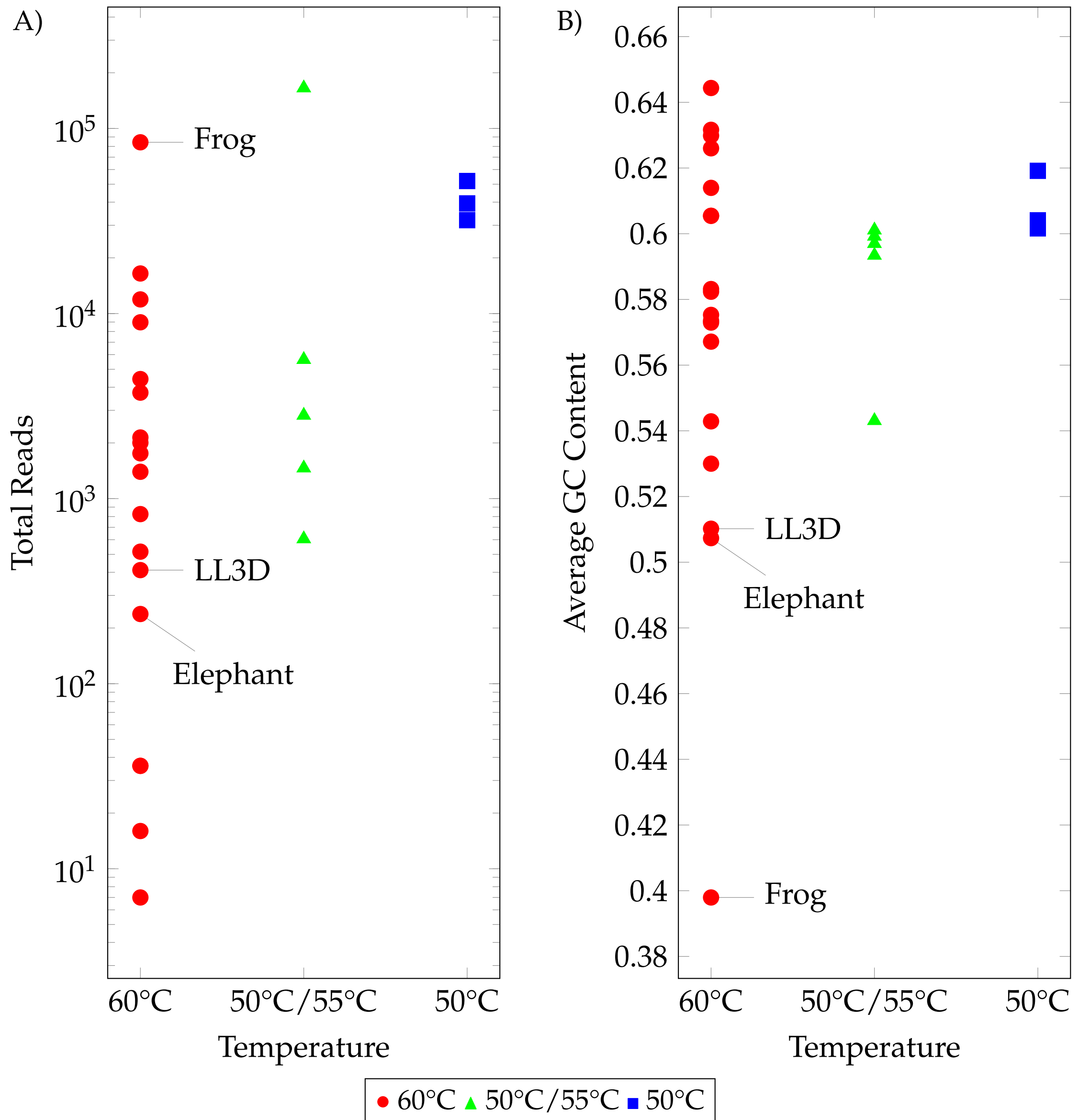


Figure 3: For each hybridization experiment, the temperature of hybridization was compared to A) the total sequencing reads B) the average GC content.

iii Community Analysis

A non-metric multidimensional scaling (NMDS) analysis is used to elucidate the nature of relationships between environmental samples with the Bray-Curtis dissimilarity index. Typically, samples that cluster together are more similar to each other. In the NMDS, we expect from previous amplicon work that samples 15k years and older will form a group, Holocene samples younger than 10k years will cluster, and samples from between those times appear between these two clusters. Overall, we see this pattern in the metagenomic NMDS (Figure 4), but without the tight, ordered structure of expected groupings that we saw in a plant and animal metagenetic NMDS plot. There are a few anomalies, such as the two oldest samples, BC1 and BC7 (30k cal yrs BP), cluster close to, and thus appear taxonomically similar to, S1 (13150 cal yrs BP) and LL3B (13205 cal yrs BP). Sample LL2C (15405 cal yrs BP) is expected to group with the other BC samples, but its core replicate, LL2A, is more similar to LL1A and C (15865 cal yrs BP). Other than T2 (10.3k cal yrs BP), the Holocene extracts (E19A, V2, GB1C) cluster on the lower right side of the NMDS plot.

Another pattern that is evident is that samples from the exact same core (such as GB2A and GB2B or LL1A and LL1C) do cluster more closely amongst themselves than between the other samples. Interestingly, the hybridization temperature difference between samples appears to have less of an effect on the ratio of assigned taxa than the UDG treatment. For example, LL1A samples have different temperatures, but both lack UDG repair and they show related communities, whereas LL1C samples share the same hybridization temperature, but differ in UDG treatment and are not as similar as the LL1A's. We also see the temperature relationship in BC, GB2, and LL1A/B; this trend may indicate that although higher temperatures are more restrictive in terms of DNA binding, the baits are catching as much of the vertebrate sequences as are actually present. Both LL1C samples, non UDG and UDG, both had a similar proportion of unique sequences, but the non UDG samples had more total sequences than its UDG-treated counterpart, although this relationship is not statistically significant. This trend is not what we would expect as a lack of UDG treatment should make it less likely for damaged sequences to bind baits that were designed without a damage pattern.

The oligonucleotide baits were supposed to catch vertebrate loci, but the majority of taxonomic assignments were to bacteria (Table 6), except for the *X. borealis* sample, which was primarily composed of the expected *Xenopus* species. It would be erroneous to assign significance to the *Loxodonta* sample, as the nearly 50-50 division of reads between bacteria and eukaryota comprise only 13 total assigned sequences. The general eukaryotic composition of these enriched extracts is higher than the

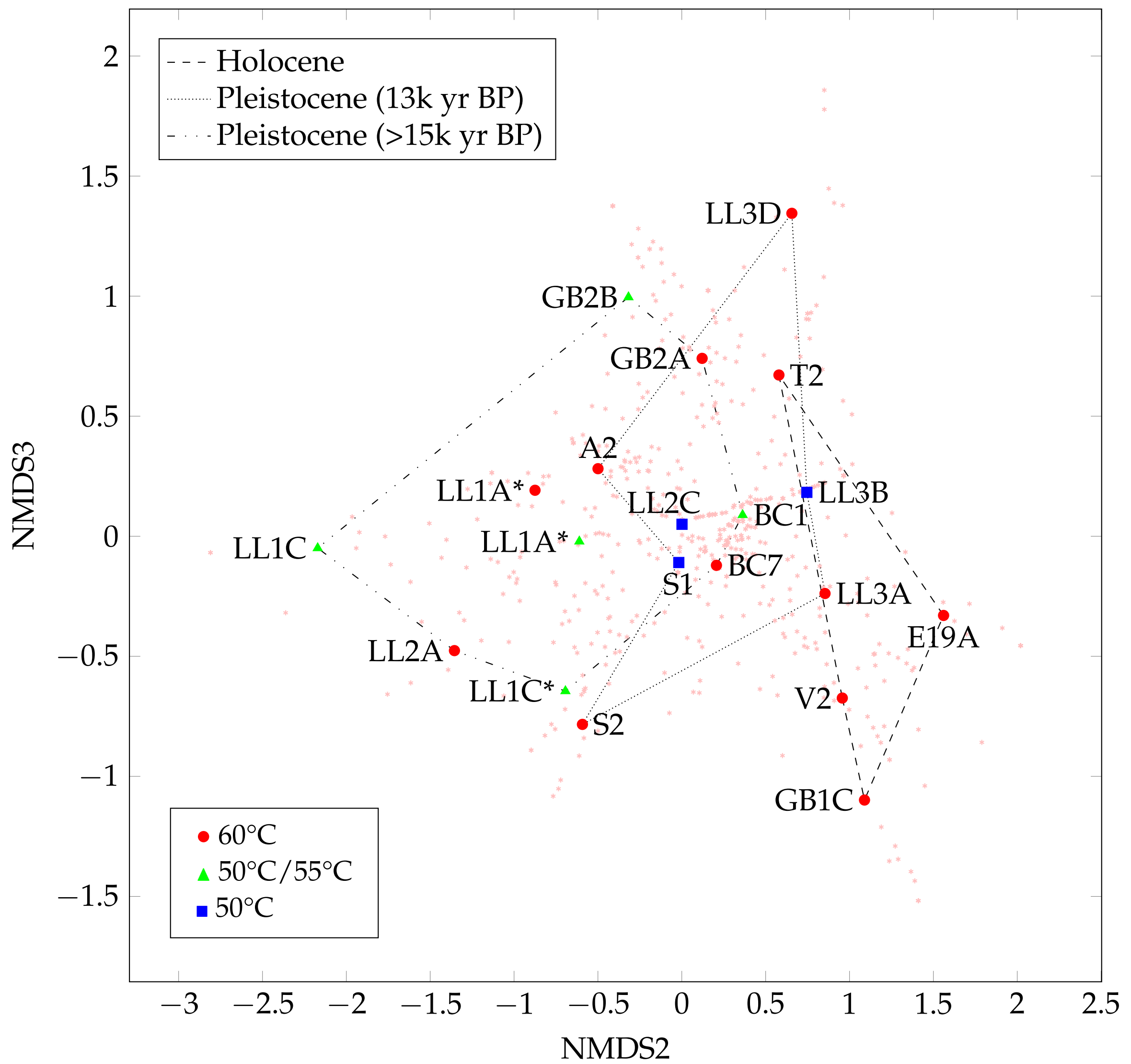


Figure 4: Bray-Curtis NMDS plot of genus level rank and higher in the metagenomic samples. Stress was 0.08 with three dimensions. Non UDG treated samples are marked with an asterisk. Taxonomic identifications are shown as asterisks.

standard background noise, a signal primarily composed of bacterial sequences. Eukaryota make up an average of 25% of the assigned reads of the metagenomic samples. Of that, Archaeplastida (plants) outnumber the Opisthokonta (animals and fungi), and the other eukaryotic kingdoms are represented by a few scattered sequences.

Table 6: Taxonomic identification by kingdom and domain for each sample. Non UDG treated samples are marked with an asterisk. Samples are listed by descending order of age (modern to 30k years).

		Bacteria	Archaea	Eukaryota						Viruses
Sample	Temp (°C)			Amoebozoa	Archaeplastida	Chromalveolata	Excavata	Opisthokonta	Rhizaria	
Elephant	60	7						6		
Frog	60	1			8	2	1	18520		4
LPB1	60	3			1					
LPB3	60	3								
V2	60	579	1		5			5		
GB1C	60	74	5		68	1		2		
E19A	60	287	18		5			5		
T2	60	149			52			9		
A2	60	378	5		293			12		
S1	50	2815	21		766	2		50		
S2	60	160	1		41					
LL3A	60	48			3		1	3		
LL3B	50	5298	12	2	430	9	43	43		2
LL3D	60	29			7			1		
LL2A	60	72			13			1		
LL2C	50	2265	16	1	1334	1	1	28		1
LL1A*	60	145			108			3		
LL1A*	50/55	199		1	112			1		
LL1C*	50/55	121			27			1		
LL1C	50/55	43			22					
GB2A	60	245	8		191			22		
GB2B	50/55	232	4		168		1	235		
BC1	50/55	7130	73	2	852	6		56	1	1
BC7	60	1442	5		34	2		4		

iv GC Content

Although the baits caught many undesired sequences, we wanted to determine if the GC content of the baits was reflected in the GC content of the sequenced reads. Figure 5 demonstrates in Part A that most of the samples had fewer than 10% of reads with less than 0.40 GC content, reflecting our expectations from the bait set. Two samples, LL3D and the elephant, had fewer than 20% of sequences with a GC content less than 0.40, while the frog sample had over 40% with a low GC. These samples were extracted differently, as previously discussed, which appears to bias both their average GC content and the number of extremely unlikely GC matches, as based on the bait GC composition. This may either indicate that low GC samples will bind the RNA baits, even in the absence of suitable matches, or it may indicate that carry over from the sample extract is a larger factor than expected. Overall, the mismatch tolerance seems to be higher than 15% and extremely dependent on the extraction method used for the sample.

v The Blanks and Controls

The library preparation blanks (LPBs) are blanks that were carried through from the various rounds of library preparation of the DNA extracts. All four were sequenced, regardless of putative content, to measure the type and quantity of contamination associated with the library preparation reagents. While the LPBs were nearly empty, they did quantify using the KAPA total molecule qPCR before sequencing, indicating that there is a large amount of empty adapter sequences in these negative controls. Each of the LPBs has less than 20 raw sequences and in LPB5 and LPB12, none of the sequences can be taxonomically assigned, despite having normal GC profiles. LPB3 has all three assigned sequences match to bacteria, while LPB1 has three bacteria and one plant sequence. There were no animal sequences detected in any of the blanks, suggesting that bait carry over into sequencing is not detected in the cleaned sequence reads.

IV Discussion

Although the enrichment did not improve the vertebrate DNA fraction, it would be remiss to dismiss the enrichment results entirely. A normal metagenomic sample will be highly biased towards bacteria (BARTRAM *et al.*, 2011; YERGEAU *et al.*, 2010), but we saw an improvement in eukaryotic DNA of up to fifty percent in several extracts. The obvious issue is that the designed baits for 12S rRNA, *cytb*, and *cox1*

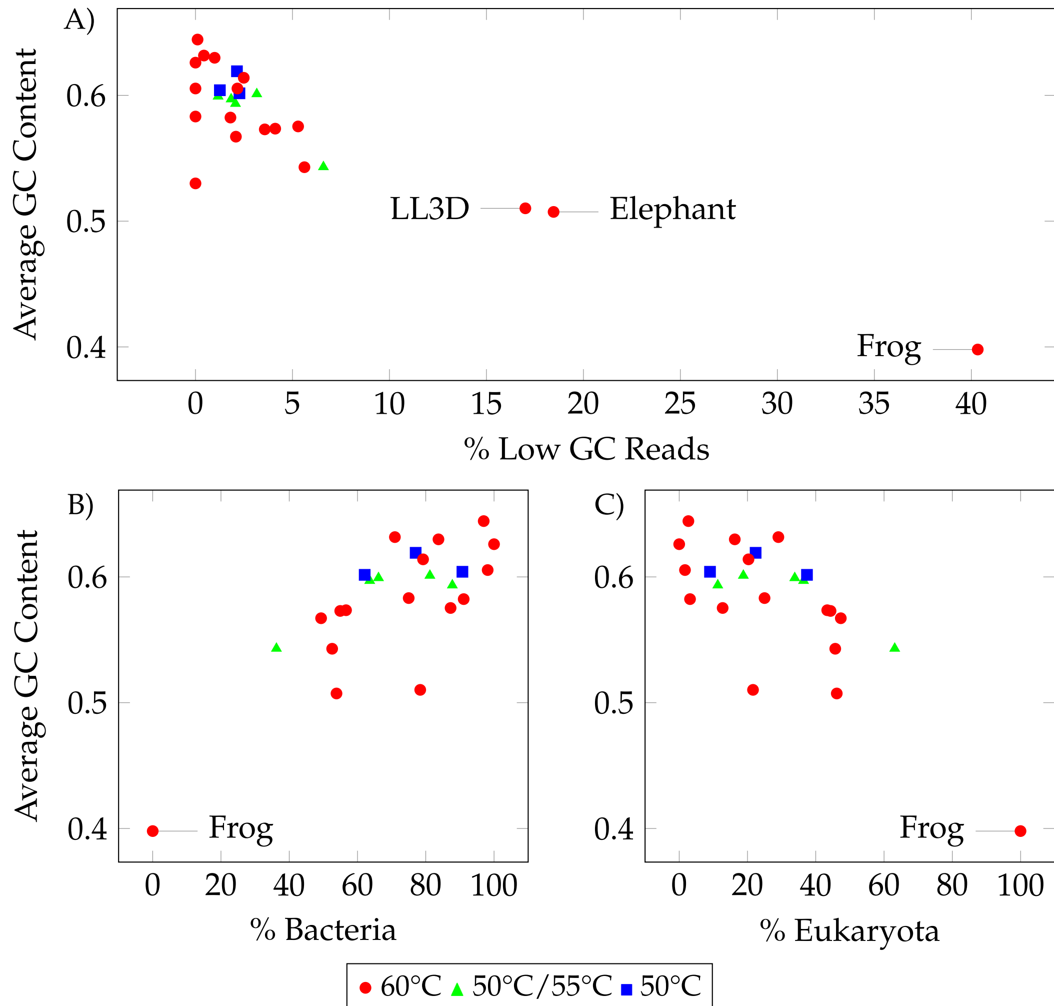


Figure 5: The average GC content was plotted against A) the percent of GC reads that were less than 40% B) the percentage Bacteria in each sample. C) the percentage of Eukaryota in each sample. Most samples had fewer than 10% of reads with a GC content less than 0.4. There is a general trend that a higher GC content in the enriched reads corresponds to a higher bacterial percentage and a lower eukaryotic composition.

did not catch almost any of these sequences. Instead, we saw a large amount of plant sequences and a low amount of Opisthokonta, most of which was not assigned to the right loci. This may be due to a very low frequency of the expected DNA in the samples, as other studies suggest there is a lower limit for microarray detection (KANE *et al.*, 2000; MOKRY *et al.*, 2010). Although the baits were not loci specific, we can examine the sequence data in other ways to assess what may have worked.

A very low total number of reads were obtained through the sequencing of these samples. A higher number of total sequences correlated with a higher frequency of unique reads and taxonomically assignable reads. However, in some samples such as BC1, only six percent of over 130,000 reads could be assigned to a kingdom level rank. Despite the large number of reads for some samples, most samples are probably below sequencing saturation for total DNA molecules ENK *et al.*, 2013. Since the purpose of enrichment is to reduce the total amount of sequencing that needs to be done, it would be pointless to sequence these samples to a higher depth to get the resolution we need. However, most of these samples quantified wrong, possibly due to adapter-dimers, so it may be possible to re-sequence these samples at higher coverage depth if the quantification problems could be resolved.

The GC content of the baits reflects the GC content of the recovered sequences. The GC content of the baits should be more than 0.4 and most samples had fewer than 10% of sequences with less than that cut-off. Since there is a GC bias in the soil samples that correlates with bacterial composition and negatively with eukaryotic composition, it may be that the baits need to be redesigned with an upper limit on GC instead of a lower limit. The lower limit was designed to increase the specific binding of the DNA to the baits by allowing more high-strength GC bonds than low-strength AT bonds (WERNERSSON *et al.*, 2007). Because the baits were based on vertebrate sequences and were *in silico* restricted to a GC content higher than 0.40, it was not hypothesized that the GC skew would cause an over-representation of bacteria. This supports the hypothesis that non-loci DNA sequences were attached to the baits, but also indicates that the recovered sequences are enriched and are not just background noise from raw extract carry-over. We could confirm the background DNA signal through the shot-gun sequencing of a few soil samples (TABERLET *et al.*, 2012), but this is cost prohibitive to do for all samples.

Fewer samples are represented in the 50/55°C and 50°C hybridization experiments than the 60°C degree dataset, however, there is still a trend that the lower temperatures are less specific in binding DNA and result in a larger number of recovered sequences. Interestingly, there is a definite relationship between the extraction replicates and the general taxonomic composition, as seen in the NMDS analysis. Despite orders of magnitude difference between sequence counts and assignments,

samples from the same time point are still more similar to each other than to other time points. As well, the enriched samples represent their ampliconic counterparts, despite radically different loci abundance. This suggests that only a few genetic markers are needed to define the difference between soil samples and elucidate a meaningful ecological relationship (BARTRAM *et al.*, 2011; PIGANEAU *et al.*, 2011).

One soil and two cellular samples were extracted with a different extraction process than the rest of the samples. Although three samples is a very small sample size, there was definitely a difference in the GC content and assignable taxa for these samples. In our amplicon work, the samples extracted with a kit comprised larger DNA fragments, were less inhibited, and amplified more easily from the genes we targeted. Future work would include a comparison of these kits and protocols with enrichment processes, as specific extraction protocols may help increase specific DNA constituents in the DNA enrichment process.

The purpose of using three genetic loci to design oligonucleotide baits was to 'catch' a wider variety of sequences than a PCR-based approach to soil metagenetics. We demonstrate that enrichment does reduce the total sequencing required to delimit ecological differences between time points and sites. However, additional testing is needed to work out some of the issues described in this paper.

V References

ALTSCHUL, S. F., W. GISH, W. MILLER, E. W. MYERS, and D. J. LIPMAN. 1990. Basic local alignment search tool. Journal of Molecular Biology **215**: 403–410.

BARTRAM, A. K., M. D. J. LYNCH, J. C. STEARNS, G. MORENO-HAGELSIEB, and J. D. NEUFELD. 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Applied and Environmental Microbiology **77**: 3846–3852.

BOOM, R, C. J. SOL, M. M. SALIMANS, C. L. JANSEN, P. M. WERTHEIM-VAN DILLEN, and J. VAN DER NOORDAA. 1990. Rapid and simple method for purification of nucleic acids. Journal of Clinical Microbiology **28**: 495–503.

BOS, K. I., V. J. SCHUENEMANN, G. B. GOLDING, H. A. BURBANO, N. WAGLECHNER, *et al.*, 2011. A draft genome of *Yersinia pestis* from victims of the Black Death. Nature **478**: 506–510.

BRIGGS, A. W., U. STENZEL, M. MEYER, J. KRAUSE, M. KIRCHER, and S. PÄÄBO. 2010. Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. Nucleic Acids Research **38**: e87.

CARPENTER, M. L., J. D. BUENROSTRO, C. VALDIOSERA, H. SCHROEDER, M. E. ALLENTOFT, *et al.*, 2013. Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries. American Journal of Human Genetics **93**: 852–864.

CHOU, L.-S., C.-S. J. LIU, B. BOESE, X. ZHANG, and R. MAO. 2010. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. Clinical Chemistry **56**: 62–72.

COOPER, A. and H. N. POINAR. 2000. Ancient DNA: do it right or not at all. Science **289**: 1139.

D'COSTA, V. M., C. E. KING, L. KALAN, M. MORAR, W. W. SUNG, *et al.*, 2011. Antibiotic resistance is ancient. Nature **477**: 457–461.

DEMURO, M., R. ROBERTS, D. FROESE, L. ARNOLD, F. BROCK, and C. RAMSEY. 2008. Optically stimulated luminescence dating of single and multiple grains of quartz

from perennially frozen loess in western Yukon Territory, Canada: Comparison with radiocarbon chronologies for the late Pleistocene Dawson tephra. Quaternary Geochronology **3**: 346–364.

EDGAR, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**: 2460–2461.

ENK, J., J.-M. ROUILLARD, and H. POINAR. 2013. Quantitative PCR as a predictor of aligned ancient DNA read counts following targeted enrichment. BioTechniques **55**: 300–309.

FAIRCLOTH, B. C. and T. C. GLENN. 2012. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. PLoS One **7**: e42543.

FROESE, D. G., G. D. ZAZULA, and A. V. REYES. 2006. Seasonality of the late Pleistocene Dawson tephra and exceptional preservation of a buried riparian surface in central Yukon Territory, Canada. Quaternary Science Reviews **25**: 1542–1551.

HANDT, O., M. KRINGS, R. H. WARD, and S PÄÄBO. 1996. The retrieval of ancient human DNA sequences. American Journal of Human Genetics **59**: 368–376.

HE, Z., L. WU, X. LI, M. W. FIELDS, and J. ZHOU. 2005. Empirical establishment of oligonucleotide probe design criteria. Applied and Environmental Microbiology **71**: 3753–3760.

HE, Z., T. J. GENTRY, C. W. SCHADT, L. WU, J. LIEBICH, *et al.*, 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. The ISME Journal **1**: 67–77.

HE, Z., Y. DENG, J. D. VAN NOSTRAND, Q. TU, M. XU, *et al.*, 2010. GeoChip 3.0 as a high-throughput tool for analyzing microbial community composition, structure and functional activity. The ISME Journal **4**: 1167–1179.

HEBERT, P. D. N., S. RATNASINGHAM, and J. R. DEWAARD. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Philosophical Transactions of the Royal Society B: Biological Sciences **270 Suppl 1**: S96–S99.

HERWIG, R., A. O. SCHMITT, M. STEINFATH, J. O'BRIEN, H. SEIDEL, *et al.*, 2000. Information theoretical probe selection for hybridisation experiments. Bioinformatics **16**: 890–898.

HORN, S. 2012. Target enrichment via DNA hybridization capture. Methods in Molecular Biology **840**: 177–188.

HÖSS, M. and S. PÄÄBO. 1993. DNA extraction from Pleistocene bones by a silica-based purification method. Nucleic Acids Research **21**: 3913–3914.

HUSON, D. H., S. MITRA, H.-J. RUSCHEWEYH, N. WEBER, and S. C. SCHUSTER. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Research **21**: 1552–1560.

KANE, M. D., T. A. JATKOE, C. R. STUMPF, J. LU, J. D. THOMAS, and S. J. MADORE. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. Nucleic Acids Research **28**: 4552–4557.

KIRCHER, M., S. SAWYER, and M. MEYER. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research **40**: e3.

KUCH, M., N. ROHLAND, J. L. BETANCOURT, C. LATORRE, S. STEPPAN, and H. N. POINAR. 2002. Molecular analysis of a 11,700-year-old rodent midden from the Atacama Desert, Chile. Molecular Ecology **11**: 913–924.

LYNCH, M. D. J., A. K. BARTRAM, and J. D. NEUFELD. 2012. Targeted recovery of novel phylogenetic diversity from next-generation sequence data. The ISME Journal **6**: 2067–2077.

MACKELPRANG, R., M. P. WALDROP, K. M. DEANGELIS, M. M. DAVID, K. L. CHAVARRIA, *et al.*, 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature **480**: 368–371.

MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER, *et al.*, 2010. Target-enrichment strategies for next-generation sequencing. Nature Methods **7**: 111–118.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10–12.

MEYER, M. and M. KIRCHER. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols **6**: pdb.prot5448, doi: 10.1016/j.bbr.2011.03.031.

MOKRY, M., H. FEITSMA, I. J. NIJMAN, E. DE BRUIJN, P. J. VAN DER ZAAG, *et al.*, 2010. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Research **38**: e116.

PIGANEAU, G., A. EYRE-WALKER, S. JANCEK, N. GRIMSLEY, and H. MOREAU. 2011. How and why DNA barcodes underestimate the diversity of microbial eukaryotes. PLoS One **6**: e16342.

SHAPIRO, B., A. J. DRUMMOND, A. RAMBAUT, M. C. WILSON, P. E. MATHEUS, *et al.*, 2004. Rise and fall of the Beringian steppe bison. Science **306**: 1561–1565.

ST JOHN, J. 2011 SeqPrep. "https://github.com/jstjohn/SeqPrep".

STEVEN, B., G. BRIGGS, C. P. MCKAY, W. H. POLLARD, C. W. GREER, and L. G. WHYTE. 2007. Characterization of the microbial diversity in a permafrost sample from the Canadian high Arctic using culture-dependent and culture-independent methods. FEMS Microbiology Ecology **59**: 513–523.

TABERLET, P., E. COISSAC, M. HAJIBABAEI, and L. H. RIESEBERG. 2012. Environmental DNA. Molecular Ecology **21**: 1789–1793.

TIQUIA, S. M., L. WU, S. C. CHONG, S. PASSOVETS, D. XU, *et al.*, 2004. Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. Biotechniques **36**: 664–70, 672, 674–5.

WAUGH, J. 2007. DNA barcoding in animal species: progress, potential and pitfalls. Bioessays **29**: 188–197.

WERNERSSON, R., A. S. JUNCKER, and H. B. NIELSEN. 2007. Probe selection for DNA microarrays using OligoWiz. Nature Protocols **2**: 2677–2691.

YERGEAU, E., H. HOGUES, L. G. WHYTE, and C. W. GREER. 2010. The functional potential of high Arctic permafrost revealed by metagenomic sequencing, qPCR and microarray analyses. The ISME journal **4**: 1206–1214.

VI Supplementary Materials

Table S1: Sample DNA extracts used for enrichment are described by original core designation, year, type of DNA extract, and their common designation.

Core/Sample	Years Before Present		Extract	
Reference	Carbon 14	Calibrated	Type	ID
Loxodonta sp	Modern	Modern	Cellular DNA	Elephant
Xenopus borealis	Modern	Modern	Cellular DNA	Frog
LL2S-60-F?	Modern PT	Modern	Cored	V2
MM12-118b	8800	9820	Cored	GB1C
LL2S-189-E	9050	10220	Homogenized	E19A
LL2S-253-D1	9195±20	10340	Cored	T2
LL2C-205-B	11290±160	13150	Cored	S1 S2
LLII-12-84-3	11360±40	13205	Cored	LL3A LL3B LL3D
LL2C-243-A2	12000	13870	Homogenized	A2
LLII-12-170-6	12900	15405	Cored	LL2A LL2C
LLII-12-217-8	13200	15865	Cored	LL1A LL1C
MM12-116b	17500	21135	Cored	GB2A GB2B
BC 4-2B	25300	30000	Cored	BC1

PT refers to the permafrost table.

Conclusion

Both metagenetic studies yielded similar conclusions about the diversity and ecological change over the Pleistocene-Holocene transition. These studies would have benefited from a series of additional time point extracts immediately preceding the Holocene. However, we were still able to demonstrate that the same analyses applied to modern soil communities (ex. LOMBARD *et al.*, 2011) are relevant and informative to ancient sedimentary permafrost collections (ANDERSEN *et al.*, 2012). We can clearly see that the three expected groups (pre-15k, 12k-15k, and Holocene) share similar organismal compositions within their groups, even in the presence of different genetic loci. These results reflect what we know about the climate change in Beringia and also demonstrate the unstable dynamics of communities right before the Holocene epoch.

The first study used amplicons from five, partial genetic loci to describe the communities and functional groups of organisms that lived around the time of the Holocene epoch. From this, we saw support for a change in the forb communities, more-so than the graminoids, and a transition from a landscape sparse in shrubs to a landscape rich in trees, shrubs, and dwarf shrubs. With these changes, the megafauna and mammoth species that were in genetic decline saw a further stress on their populations. Strong megafaunal genetic signals drop out of the DNA record just before the PH transition. For some species, this signifies their extinction in this area of the Yukon.

The second study built on the first, through the use of enrichment to target three, full genetic loci at multiple positions. Theoretically, this would reduce amplicon bias of single-gene PCR-based methods and would sample from a wider range of genetic material, without exponentially increasing high-throughput sequencing costs. Unfortunately, the oligonucleotide baits did not work as designed, possibly due to low numbers of target DNA in these samples. Assigned DNA sequences for the vertebrate loci were on the same magnitude in terms of unique species represented, putatively indicating that the samples were just very low in vertebrate DNA. Additionally, we did not see a huge increase in specificity of the baits as the temperature increased, as all samples were within the same range of total reads, assigned reads, and GC content. We would caution that future studies on permafrost vertebrate DNA use extraction techniques that increase the availability of vertebrate sequences, as opposed to traditional methods. We expect that if we had used a soil extraction kit on more samples, we would have seen the same skew towards vertebrates and away from bacteria, even at a cost of percentage of the recoverable DNA sequences.

Future work would include removing regions of the bait gene library that are not sufficiently discriminatory. As well, re-designed baits would be ideally tested on samples of known composition, both created artificially and from modern soil, to help identify biases that may have been overlooked and to optimize the process.

The Pleistocene-Holocene transition was a period of rapid, tumultuous climate change that destabilized the dominant flora and fauna. This era saw the decline and extinction of many iconic megafauna and a complete redefinition of the plant community. Although the Holocene biome does not comprise the same taxa as the late-Pleistocene, there is still a vibrant community that is a testament to the resourcefulness of life.

I References

ANDERSEN, K., K. L. BIRD, M. RASMUSSEN, J. HAILE, H. BREUNING-MADSEN, *et al.*, 2012. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. Molecular Ecology **21**: 1966–1979.

LOMBARD, N., E. PRESTAT, J. D. VAN ELSAS, and P. SIMONET. 2011. Soil-specific limitations for access and analysis of soil microbial communities by metagenomics. FEMS Microbiology Ecology **78**: 31–49.