UNIVERSITY OF ADELAIDE

DOCTORAL THESIS

Climate-Driven Ecological Changes Through The Last Glacial Period

Innovations in Plant Ancient DNA and Stable Isotope Palaeoecology

Author:

Mark Timothy
RABANUS-WALLACE

Supervisors:

Prof. Alan COOPER
Dr. James BREEN
Dr. Hugh CROSS
Dr. Bastien LLAMAS

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Australian Centre for Ancient DNA
Department of Biological Sciences

September 17, 2017

Declaration of Authorship

I, Mark Timothy RABANUS-WALLACE, declare that this thesis titled, "Climate-Driven Ecological Changes Through The Last Glacial Period: Innovations in Plant Ancient DNA and Stable Isotope Palaeoecology" and the work presented in it are my own. I confirm that:

- This work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.
- No part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.
- I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.
- Copyright of published works contained within this thesis resides with the copyright holder(s) of those works. I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: _____

Date: _____

UNIVERSITY OF ADELAIDE

Abstract

School of Genetics and Evolution

Department of Biological Sciences

Doctor of Philosophy

Climate-Driven Ecological Changes Through The Last Glacial Period

by Mark Timothy RABANUS-WALLACE

The impact of climate-driven ecological changes can be understood by reconstructing the effects of past climate variation on the flora and fauna. This thesis develops and applies new methods for inferring the history of the graminoid-dominated steppes of the northern Holarctic and Patagonia as they declined during the end of the Last Glacial Period (25,000–10,000 years ago). Stable nitrogen isotope data are used to argue for the pivotal role that landscape moisture played in the decline of the Pleistocene megafauna, and a new method for inferring relative changes in plant-available moisture from herbivore collagen isotopic measurements is developed. Experimental methods for working with botanical ancient DNA are presented, tested, and used to explore the taxonomy and evolutionary histories of three ancient plant species, ultimately yielding the two oldest known draft chloroplast genome sequences, dating to between 50,000 and 80,000 years ago. The results confirm the strongly-reticulated phylogenies characteristic of plants evolved to employ great plasticity as an adaptive ability, even with minimal sexual reproduction. All new genetic methods are tested with the aid of a newly designed program SimWreck, which simulates sequence data with the known characteristics of ancient DNA.

Contents

List of Figures

List of Tables

Supplementary Material

A digital Data Supplement (DS) is provided with this thesis, consisting of the following files:

DS_Readme.txt	Column-by-column descriptions of the contents of the data-files.
DS_Samples.csv	A list of herbarium specimens used in the study, with metadata on their collection location, source, etc. Referred to in text as DS:Samples.
DS_Data.csv	Metadata on libraries sequenced during the study, including read counts, quality summaries, etc. The summary stats are produced for libraries at different stages of the read-processing pipeline described in chapter 4 (raw, merged, trimmed, etc.) Referred to in text as DS:Data.
DS_Code.txt	Programs, scripts, and commands used in the analysis, with numbered sections. Referred to in text as DS:Code:Section Number.

Abbreviations

The following abbreviations are used in the thesis (commercial reagent names are omitted):

AB	Alberta
ACAD	Australian Centre for Ancient DNA
ACRF	Australian Cancer Research Foundation
AFLP	Amplified Fragment Length Polymorphism
AGRF	Australian Genome Research Facility
ATP	Andenosine TriPhosphate
BC	British Columbia
BLAST	Basic Local Alignment Search Tool
BWA	Burrows Wheeler Aligner
CDS	Coding DNA Sequences
CI	Confidence Interval
CTAB	Cetyl Trimethyl Ammonium Bromide
DNA	DeoxyriboNucleic Acid
DS	Data Supplement
EB	East Beringia
EDTA	EthyleneDiamineTetraacetic Acid
HTS	High Throughput Sequencing
IPCC	Intergovernmental Panel on Climate Change
IUPAC	International Union of Pure and Applied Chemistry
LGIT	Last Glacial Interglacial Transition
LGM	Last Glacial Maximum
LSC	Long Single Copy
NATO	North American Treaty Organisation
NCBI	National Centre for Biotechnology Information
NEB	New England Biolabs
NW	NorthWest
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PEG	PolyEthylene Glycol
PEU	Paired End Untrimmed
PGSB	Plant Genome and Systems Biology
PNK	PolyNucleotide Kinase
PVP	PolyVinylPyrrolidone
QBI	Queensland Biology Institute
QT	Quality Trimmed
RAD	Restriction site Associated DNA
RE	Repetitive Elament
RNA	RiboNucleic Acid
RO	Reverse Osmosis

RSA	Rabbit Serum Albumen
SBS	Sequencing By Synthesis
SE	SouthEast
SET	Single End Truncated
SM	Supplementary Materials
SNP	Single Nucleotide Polymorphism
TAIR	The Arabidopsis Information Resource
USA	United States of America
UV	UltraViolet
WA	Washington
WB	West Beringia
YPP	Yukon Palaeontology Program
YT	Yukon Territory

Conventions

All times given are in absolute years (c.f. radiocarbon years) unless otherwise stated. The units used are kiloyears (ky; 1,000 years) and kiloannum (ka; 1,000 years before present).

Unless otherwise specified, logs are given in base *e*.

References to material in the digital data supplement are explained in the relevant section of the front matter, and prefixed with "DS:".

All original scripts written by the author are prefixed with MTRW, and are available in DS:Code.

Where necessary, samples are identified by their partial or whole filenames, which all have the following format for parsing and globbing convenience:

```
Genus_{species|nest ID}_[mod_][enr_]libXXX###_extYYY###\
_MoreInformation.extension
```

Where the genus is followed by a species name and the identifier “mod” for herbarium specimens, and a nest ID code for permafrost specimens. The identifier “enr” is present for enriched samples. “XXX” and “YYY” are mnemonic 2-3 letter codes for libraries and extractions respectively, and "###" are unique number identifiers for samples in library prep or extraction. “MoreInformation” may contain barcodes, or details about the processing pipeline such as the reference genome if the file is a mapped reads file.

Each batch of extractions, batch of library preparations, and seqencing run is identified with a 2–3 letter identification code corresponding to the metadata available in DS:Data. The identity codes correspond to simple mnemonics: ’tan’ <=> "tangled", ’dar’ <=> "Darwin", ’tun’ <=> "tundra", ’nd’ <=> "second", ’pow’ <=> "powder", ’kmga’ <=> "Kendrick Marr/Geraldine Allen" (see acknowledgements), ’kw’ <=> "knotweed", ’fw’ <=> "fireweed", ’cm’ <=> "checkmate, ’bm’ <=> "Bistorta modern", ’mm’ <=> "modern mix", ’ms’ <=> "modern slam", ’mst’ <=> "modern slam two", ’lon’ <=> "longer", ’cmp’ <=> "compare", ’ste’ <=> "steppe", ’tc’ <=> "tube crack", ’Dhiseq’ <=> "Draba HiSeq", ’cmb’ <=> "checkmate/bears", ’cet’ <=> "chloroplast enrichment test", ’fe’ <=> "fireweed enriched", ’lf’ <=> "longer/fireweed". Identity codes are occasionally reused for, say, both an extraction batch and a library preparation batch.

Acknowledgements

The work presented here would not have been possible without many collaborators, colleagues, and friends. I am hugely grateful to the following people:

- For their mentorship in all things Quaternary: Matthew Wooller, Grant Zazula, Elizabeth Hall, and Sue Hewitson.
- For their hospitality: The Zazula family.
- For their enthusiastic support in the field: Families Johnson, Schmidt, and Christie, and all the Yukon goldmining community.
- Variously for brainstorming sessions, trading ideas, training, sharing protocols, offering explanations, collaboration, laboratory services, proof reading, professional opportunities, and occasional debugging: All the staff at ACAD, in particular Maria Lekis, Corinne Callegari, Steve Richards, Julian Soubrier, Kieren Mitchell, Oli Wooley, Graham Gower, and Pere Bover Arbos. From elsewhere, Eric DeChaine, Ingrid Jordon-Thaden, Chris Turney, Chris Helgen, Tyler Faith, Fred Longstaffe, Natalie Betts, Geoff Fincher, Iain Searle, John Conran, Rosalie Kenyon, Joel Geoghegan, Andreas Schreiber, Janette Edson, Marc Hew-Jones, Adam Croxford, Jono Tuke, Ben Rohrlach, Alice Gorman, Cathy Miller, Steve Pedersen, Thomas Windram, Denis Sjostrom, and Jake Parker. I also owe many thanks to the staff of Cibo Kurralta Park (long black with a dash of skim) for keeping me awake, and to Lauren White for keeping me together.
- The samples used in this study are listed in the data supplement. DNA samples for modern *Bistorta vivipara* were kindly provided by Geraldine. A. Mulligan and Kendrick L. Marr from the University of Victoria, B.C., Canada. Bruce Bennett of the Yukon Department of Environment, Y.T., Canada generously provided access to his personal herbarium of Yukon plants. John Conran at the University of Adelaide helpfully provided a sample of *Ranunculus repens*.

Many authors and scientists who were pivotal in developing and communicating the ideas that underlie this work. Those to whom I am most indebted include R. Dale Guthrie, Evelyn Pielou, Richard Harington, Robert Whittaker, Richard Dawkins, David Attenborough, Les Cwynar, Grant Zazula, Duane Froese, Scott Elias, Daniel Mann, Warren Evans, Gregory Grant, Larry Wall, and many more.

Finally, I am most grateful to my supervisors for their invaluable support. In particular I thank Jimmy Breen and Bastien Llamas, who helped me most closely day-to-day throughout the entire project. I greatly appreciate the freedom they gave me to think creatively and try new things, and it is a privilege to have been given the opportunity to follow the examples set by such talented scientists and supportive friends.

To my parents, Angela Rabanus and Allan Wallace—for your love and support, and for encouraging me to be a curious person…

Chapter 1

Introduction

1.1 Studying Anthropogenic Climate Change

In 1856, the author and women's rights advocate Eunice Newton Foote used two glass flasks, several thermometers, and an air pump, to demonstrate the effects of compression, moisture, and carbonic acid gas on the absorption of radiant energy by air (FIG. 1.1) [21]. Professor Joseph Henry read her findings before the American Association for the Advancement of Science in August that year, three years before John Tyndall's famed 1859 announcement that "different gases are thus shown to intercept radiant heat in different degrees" [59]. Eighteen fifty-nine also marked the publication of the first edition of On the Origin of Species whose author, Charles Darwin, stirred public consciousness of the depth of time and changeable nature of the planet over millions of years[18]. Contemporary geologists including Louis Agassiz, William Buckland, and—after a brief period of resistance—Charles Lyell, were instrumental in convincing the scientific community of an early *ice-age theory*. This revolutionary school of thought held that various geological features of modern Europe had been formed long ago by glacial processes acting at continental scales [8]. By the late eighteen hundreds, most scientists accepted the fact that drastic changes in climate occurred over geological timescales [42]. The realisation that alterations in atmospheric composition may exert significant control over Earth's climate systems persisted as a topic of some minor interest within the scientific community, kept afloat by new discoveries in palaeontology, such as the variations in the width of tree rings over long time scales. Svante Arennius, in 1897, specifically addressed the greenhouse gas link, focusing on water vapour and CO_2, and calculating that doubling atmospheric CO_2 would increase the mean surface temperature by around 5°C [5].

As the drivers influencing global climate became better studied, CO_2 concentration gained acceptance as a major driver, alongside alterations in Earth's orbit, and feedbacks involving Earth's albedo. In 1969, NATO became the organisers of the first international discussion on climate change [1].

Today, there is strong consensus among scientists that anthropogenic carbon inputs are increasingly responsible for the effects of current climate destabilisation. In some areas, these

In Common Air.		In Carbonic Acid Gas.	
In shade.	In sun.	In shade.	In sun.
80	90	80	90
81	94	84	100
80	99	84	110
81	100	85	120

FIGURE 1.1: A table excerpted from Foote, 1856 [21], showing the temperature (°F) increase in flasks containing air with and without added CO_2 in both sun and shade.

changes have been a windfall: In 2013, greater snowmelt and a longer growing season allowed a Greenland chef to surprise his royal guests with a handful of locally-grown strawberries: just one of the new plants growing in the region that would have presented a practical impossibility just one decade earlier (FIG. 1.2) [23]. During this same decade, however, the strengthening El Niño that flooded central eastern Africa displacing residents, destroying property, and leading to cholera and malaria epidemics, also exacerbated droughts in the south of the continent, plunging nearly 20 million people into a state of severe food insecurity [2].

FIGURE 1.2: **Left:** A farmer overlooks newly-cultivated potato crops in Greenland, 2006 [23]. **Right:** Droughts destroy livestock in Sudan, 2014 [2].

The challenge of building a world where unpredictable—sometimes drastic—changes in the habitability, production capacity and resource demands of many regions can be navigated without stressing internal and international relationships to the point of collapse, is especially daunting given the lack of historical precedents. Accurate foresight is arguably the most valuable piece of the human arsenal for adaptation. Historical trends, such as climatic changes in crop production regions (FIG. 1.3) can be used to infer the direction of future changes at decadal scales, but their predictive power wanes at longer timescales [38].

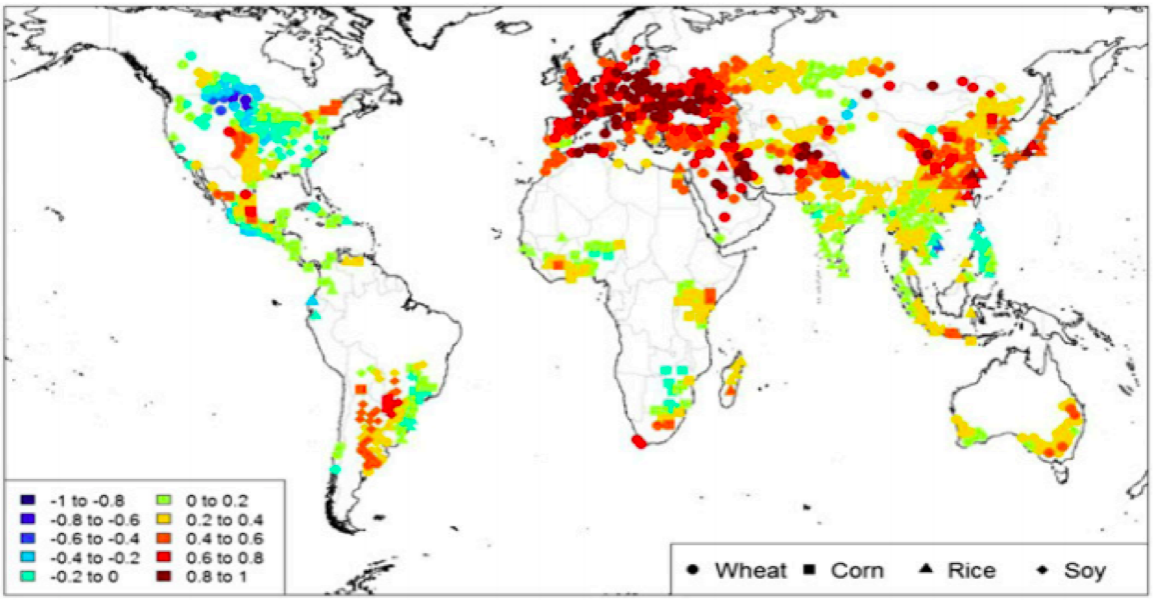

FIGURE 1.3: Change in average maximum growing season daily temperature (°C) in crop-producing regions since 1980 [41].

1.2 Climate Change and Ecology

Climate is intimately linked to ecology, and the distribution of biome types across the globe is controlled largely by climatic variables [61]. Our growing understanding of the climate-ecology relationship is reflected in climate models that explicitly model the effect of living organisms on the Earth's albedo, chemical cycles such as the water cycle, and atmospheric composition [4]. The juxtaposition of palaeontology against palaeoclimatology is essential to this understanding.

Nevertheless, understanding climate-driven ecology at a planetary scale presents many ready challenges. At face value, the pursuit is barely scientific, since it relies largely upon on a single system—the Earth—meaning there are no controls, no replicates, and limited ability to manipulate the experimental object. Furthermore, direct observation of the past is impossible, and must be inferred using proxies, each of which entails a suite of limitations, biases, and possible sources of inaccuracy. Despite such challenges, the field has made considerable headway. Makeshift controls and replicates can be approximated by studying multiple regions and multiple climate events, and technology constantly increases the number and accuracy of available proxies, allowing humans to combine the data from many proxies to yield coherent interpretations of past climatic and ecological changes.

It is on the basis of such interpretations that informed responses to ecological changes can be planned. The use of computers to simulate possible futures based upon climate models has enabled much more specific predictions about possible outcomes, complete with formal estimates of certainty [38, 51, 58]. However, climate simulations are only as accurate as the models they employ represent reality, and the empirical data they train on are abundant

[51]. The development of new methods for reconstructing the recent past is, therefore, highly beneficial in the global effort to adequately understand climate change.

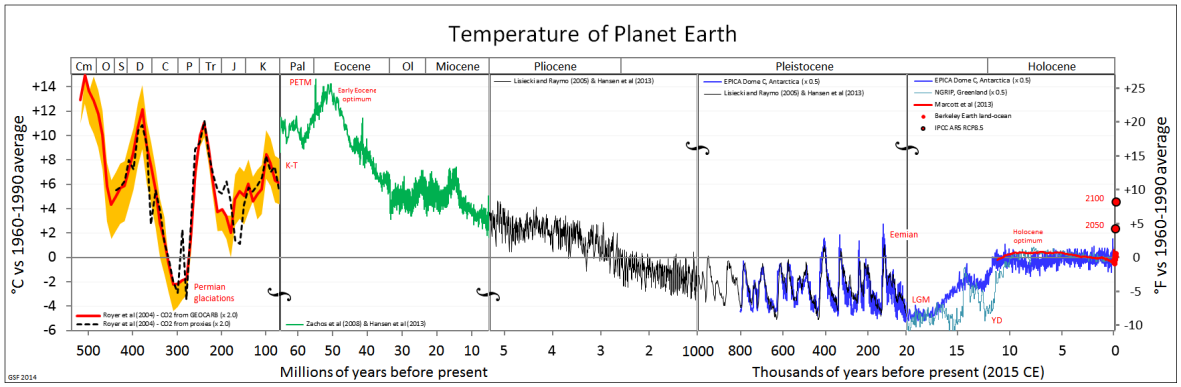

FIGURE 1.4: Complied temperature proxy data for the last 500 million years (*Image: Wikipedia Commons*). Note breaks in the scale along the time axis.

This thesis describes three approaches to extending existent proxies for environmental change during recent climate events. These projects all centre around the environmental changes of the Late Quaternary (~the last 500 ky), and focus in particular on an extinct Holarctic ecozone known as the mammoth steppe. Background on these topics therefore makes up the remainder of chapter 1.

1.3 A Natural Laboratory for Climate-Driven Environmental Change

1.3.1 The Late Quaternary

The last two million years has seen a series of climatic fluctuations (FIG. 1.4), yielding ample opportunity to investigate past warming and cooling events. Late Quaternary climate science was greatly accelerated by the chemical analysis of geological core samples [52], allowing climate to be reconstructed going back millions of years [40]. These cores yielded high-resolution temporal profiles for, among other things, temperature, airborne dust, and atmospheric carbon dioxide levels.

Ice cores from Greenland and Antarctica reveal that Late Quaternary climate fluctuated between glacial and interglacial periods, with recent glacials altering the mean sea surface temperature by around 10°C approximately every 100 ky [40]. Northern hemisphere proxies have recorded strong quasiperiodic 2–4°C temperature shifts known as Dansgaard-Oeschger (D-O) events, superimposed over the last glacial-interglacial cycle [17]. Nineteen D-O events occur in the period 20–80 ka (FIG. 1.4). These northern-hemispheric fluctuations of around may relate to a less pronounced series of fluctuations in Antarctic records (known as the Antarctic Isotope Maxima), and to spikes the deposition of iceberg-transported debris, known as Heinrich events [30, 36]. Heinrich events often occurred during cooler intervals in the Greenland record, and warmer intervals in the Antarctic record. It has therefore

been suggested that D-O events are caused by rapid melting in the north, which floods the north Atlantic with cold, fresh water, sometimes so catastrophically that a Heinrich event is recorded as calving icebergs transport massive amounts sediment into the oceans and seas [30]. This influx may then interrupt the Atlantic thermohaline current that links the two proxies, allowing events affecting one to be recorded in the other[10]. During glacial periods, Northern ice sheets expanded southwards reach latitudes around 38°N [13].

The current interglacial began around the beginning of the Holocene (10 ka), following the Last Glacial Maximum (LGM), during which time ice sheets extended over the greater part of the northern hemisphere (FIG. 1.5) [13]. The term "maximum" in this context refers to the spatial extent of glaciation, and as such, the LGM occurred at differing times across the globe, ranging from perhaps around 23 ka in Siberia to around 14 ka in Europe [14, 50]. In the southern hemisphere, ice sheets were limited to Antarctica and the southern Andes [37].

Ice sheets resting on land cause considerable eustatic sea level change, lowering the sea level by over 100 metres at the peak of the LGM [13]. The exposure of continental shelf during cool periods altered the biogeography of the globe allowing the passage of flora and fauna across land bridges [25, 27, 34], such as those joining Siberia to Alaska, and Australia to New Guinea [50]. The Last Glacial-Interglacial Transition (LGIT) refers to the events that occurred between the LGM and the modern comparatively stable climate regime that can be seen in recent times at the right of figure 1.4.

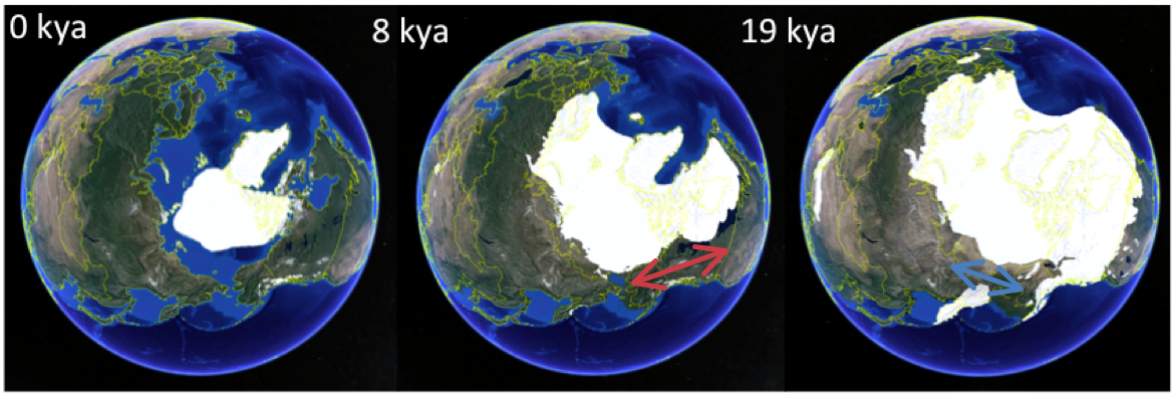

FIGURE 1.5: Extent of winter ice in the northern hemisphere at three times after the LGM (view from the North Pole.) **Red arrow:** McKenzie Corridor. **Blue arrow:** Bering Isthmus.

1.3.2 The Last Glacial-Interglacial Transition

Ice core proxies show an overall increase in global temperature throughout the LGIT (FIG. 1.4) was interrupted by temporary reversals that occur at different times in Greenland and Antarctica proxies[47, 53]. These reversals include the Younger Dryas, Older Dryas, and Oldest Dryas stadials, the Bølling and Allerød Warmings (Known as the Bølling-Allerød warming where the proxies do not record a separation between them), and the Antarctic Cold Reversal (see chapter 2, figure 2 for more detail). As global temperatures rise, retreating ice sheets release a great amount of water, which is complemented by increased precipitation that occurs in most global regions as the temperature of air rises, increasing its moisture holding capacity [27, 48]. Many terrestrial hydrological features were formed during the LGIT, including inland seas and river systems that drained huge amounts of fresh water into the northern oceans [20, 46]. Pollen cores record the expansion of temperate and mesic-adapted vegetation. Fossil records are reinforced by genetic, archaeological, and other indirect proxies (e.g. *Sporormiella* fungal spores), all of which show a global decline in the megafaunal biomass [6, 39], alongside various human migrations including into the Americas from Eurasia.

Causality is notoriously difficult to establish for the Quaternary extinctions. Changing environments and human hunting no doubt contribute to megafaunal decline, but climate also influences human migration, while the presence of megafauna influences the environment, and so forth [6, 7, 11, 39]. Investigating the causes of the Quaternary megafaunal extinctions, therefore, ideally focuses on how these factors interacted, rather than which particular factor is most to blame [6, 49]. This endeavor struggles severely from a lack of natural replicates, because the Quaternary extinctions differ between global regions (see chapter 2). Notably, while human migration immediately precedes extinction events in Australia and the Americas, humans lived side-by-side with megafauna for millennia in Europe, Southeast Asia, and Africa, without major extinction events [6, 39, 49] (see figure 1.6). A satisfactory explanation must predict such differences in timing, as well as other general trends such as the observation that grassland specialists were usually affected more than forest browsers. This is a major theme in chapter 2.

1.3.3 The Mammoth Steppe

The mammoth steppe was an ecozone that spanned two-thirds of the northern hemisphere for much of the Quaternary period (2.58 mya to present) [27, 33]. This expansive region became arguably the world's best natural laboratory for understanding the ecological consequences of climate change, having produced abundant fossil material that can be recovered in any rivers, tunnels, or mines that cut through permafrost 'muck' deposits [29, 60, 64] (see chapter 3). The perennially cold preservation conditions also allow unparalleled survival of nucleic acids for ancient DNA analysis [32]. The preservation of long lake sediment cores preserving pollen and plant macrofossils allow for longitudinal profiles to be constructed at many widespread sites [9], and the landscapes preserve many geological features formed

during this critical period of cyclical climate [36, 40], ecological change [3, 24, 48], human migration [31, 57], and extinction events [15, 39].

During glacial periods, the Bering Isthmus, which joins Alaska and Siberia, was exposed allowing the survival of ice age fauna and flora in the region now known as Beringia [34]. Hence, the mammoth steppe was a continuous band including Europe, Siberia, and north-eastern America (Alaska, USA, and the Yukon Territory, Canada)[27]. The easternmost boundary of the mammoth steppe was, during full glacials, bounded by the ice sheets of the North American ice sheet complex, comprising the Laurentide (east) and Cordilleran (west) ice sheets that retreat to open the Mackenzie Ice Corridor—an ice-free passage to regions south of the ice sheets—during interglacials (see figure 1.5, middle image) [27]. As a result, East Beringia functioned like an air lock, allowing the dispersal of organisms either across the Bering strait, or through the Mackenzie corridor at different times[25, 27]. The iconic megafauna of the mammoth steppe [25, 31, 33, 34, 39, 48] included, at various intervals and in various regions, smaller mammals such as voles (*Microtus, Phenacomys, Clethrionomys*), lemmings (*Dicrostonyx, Lemmus*), ground squirrels (*Spermophyllous*), and badgers (*Taxidea*), large herbivores like horses (*Equus*), mammoths (*Mammuthus*), mastodon (*Mammut*) bison (*Bison*), muskoxen (*Ovibos*), rhinoceros (*Coelodonta, Stephanorhinus*), Irish elk (*Megaloceros*), moose (*Alces*), reindeer (*Rangifer*), Saiga antelopes (*Saiga*), camels (*Camelops*), giant beavers (*Castroides*) and ground sloths (e.g. *Megalonyx*), and predators such as lions (*Panthera*), hyenas (*Crocuta*), saber-tooth cats (*Smilodon, Homotherium*), wolves (*Canis*), and bears (*Ursus, Arctodus*).

The mammoth steppe was characterised by cold, aridity, and seasonality [27, 46]. Mountain ranges and ice sheets probably shielded the interior from heavy precipitation. During Glacials, the tree line migrated southward, giving way to arid steppe-like biomes, variously interspersed with tundras, shrublands and even deserts. Aeolian processes transported and deposited large amounts of fine dust (or loess) created by glaciers [27]. The cold-arid-seasonal combination created unique conditions that sustained an environment often said to have no modern-day analogue, though the makeup of the landscape has generated intense debate: while pollen cores often suggest a predominance of tundra plants interspersed with graminoids (grasses and grass-like plants in the families Poaceae, Cyperaceae, and Juncaceae) [64–67], the fossil records a large biomass of megafaunal herbivores. These animals' modern analogues, digestive systems, dental morphology, stomach and dung contents, tooth wear, and tooth contents (plant macrofossils stuck in the teeth) suggest they were primarily grazers, dependent upon large swathes of grassland [16, 24, 25, 27, 60].

Tundra vegetation actively resists herbivory (for instance via toxic compounds), and grows slowly, meaning few nutrients would have been available to sustain grazing herds [24, 27]. Resolution of the so-called productivity paradox [24, 33] has been attempted in several ways: The mosaic hypothesis posits that a patchwork of forest, shrubland, grassland and tundra-like biomes existed. Other authors posit overall grassland dominance, arguing that faunal assemblages are the most reliable guide to past environments, and that the pollen profiles that inspired the mosaic hypothesis are taphonomically biased and cannot be accepted

prima facie as a good record of the vegetation present at the time [24]. New metagenomic methods have suggested megafauna survived on a grassy steppe-tundra by supplementing their heavy grass intake with nutritious non-graminoid herbs (or forbs) [62], though these studies no doubt suffer from many of the same potential sources of sampling and identification bias that affect the palaeontological and palynological findings.

1.3.3.1 A Moisture-Controlled Ecozone

R. Dale Guthrie has defended an interpretation of the mammoth steppe that represents mainstream thinking today [24, 27], and is central to chapter 2, in which evidence is presented suggesting that his moisture-centric model of the mammoth steppe is applicable to similar systems across the globe. Guthrie maintains that seasonality and aridity are of utmost importance. The cold climate of the north means that permafrost (perennially-frozen soil) underlay most of the mammoth steppe ecozone, leaving geological scars and imposing a unique suite of interactions and feedbacks with the biosphere.

Low precipitation means that snow cover is minimal, and in the long days of the arctic summer, solar radiation, unimpeded by heavy cloud cover, deeply thaws permafrost near the surface creating an active layer of soil with mobile moisture [24]. The depth of this thaw was increased further by the clearing of insulating vegetation from the surface by grazers [25].

The seasonal thaw creates ideal conditions for grasses, which compete by dominating the subterranean zone with large root systems that make up the largest part of the whole plant's biomass [61]. Resources stored in the roots can be mobilised during the growing season to produce copious, expendable foliage. The intercalary meristems of graminoid blades allow them to continue growth even after being clipped by grazers, and many grazers do indeed take advantage of the abundant nutrients offered by grass [25].

Subsistence on grass alone is most efficiently managed by a large digestive system that processes the large amount of low-energy-per-mass material required to meet the organism's energy requirements [25]. Alongside the thermal efficiency benefits conferred upon large-bodied organisms subject to the low temperatures of the Quaternary north, this digestion-energetic benefit of large body size when the primary source of nutrition is distributed across abundant low-quality forage explains why the "big three" most numerous herbivores in the Late Pleistocene mammoth steppe food web were large: horses, bison, and mammoths [19, 25, 54].

Many mammoth steppe grazers underwent (local or global) extinction during the LGIT, with both the fossil record and ancient DNA studies tracking the decline [6, 7, 11, 26, 28, 35, 39, 43–45, 49]. Most obligate grazers went extinct on most of the mammoth steppe, with some horse and bison surviving late in Europe and on the Eurasian steppes, and other Bison populations surviving south of the North American Ice Sheet Complex on the North American prairies [55, 56]. Pollen proxies reveal the northward LGIT migration of the treeline and

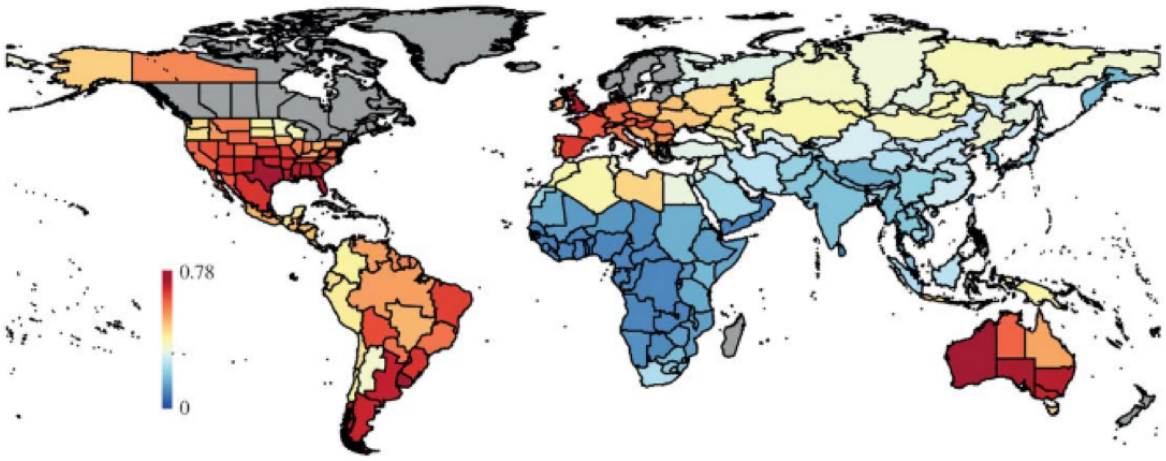

FIGURE 1.6: Proportion of large mammal species becoming extinct during the LGIT by region, from Sandom et al. (2014). Grey shading indicates excluded data.

the development of widespread tundra, taiga, and wetland environments in regions once occupied by mammoth steppe indicator vegetation [64–67], suggesting that climate-driven ecological changes probably played a major role in the extinction [15]. Browsers and mixed feeders such as reindeer, moose, and muskox now make up the bulk of the large mammals populations on the former mammoth steppe. There also exists extensive evidence for human hunting including stable isotopic dietary analysis, cut marks on butchered fossils, and the hunting tools found in archaeological sites [12, 63].

1.4 The Thesis in Context

The following chapters describe work aimed at developing new tools to enhance current understanding of the past climate change, focusing on the kinds of material available from the mammoth steppe. The approaches are somewhat disparate, with some projects being undertaken opportunistically, yet the palaeoecology of the mammoth steppe as an analogue for contemporary events remains the central theme.

The chapters are best read in order, since the findings and data from one chapter often forms the basis for the next. Chapter 2 describes the use of collagen nitrogen isotopes, often used to reconstruct the diets of fossilised herbivores, as a novel proxy for changes in moisture. Chapters 3–5 concern the nature of ancient DNA, which is highly fragmented and is present in very small quantities in the sample; a full description is given in the introduction to Chapter 4. Specifically, these middle chapters describe efforts in applying ancient DNA methods to fossilised flora preserved in permafrost, including field (Chapter 3), laboratory (Chapters 3 and 4), and computational (Chapter 4) methods for extracting useful information from the DNA these samples yield, even when no well-characterised reference genomes are available. Chapter 6 describes software developed for the project that has applications in ancient

DNA more broadly, allowing researchers to assess the impact of DNA degradation on their analyses more easily.

The marriage of a herbivore palaeoisotope study with work on ancient plants is fairly natural, since plants play such an important role connecting herbivores to the nitrogen cycle: With moisture levels exerting primary control over the plant community, which in turn exerts primary control over the herbivore community, the work described here focuses on a strongly interconnected suite of biological processes.

Much of the described work is also applicable to work on other material and environments, for instance, the stable nitrogen isotopic approaches can be applied wherever abundant dated fossils from a dry palaeoenvironment exist. The botanical ancient DNA investigations and methods may be informative on studies using similar material found in other long-frozen environments, such as the mid-Holocene plants that have been recovered from high-Andean glaciers since 2002 [22], or even similar samples found outside permafrost environments. The DNA degradation simulator is applicable to High Throughput Sequencing (HTS) studies in general.

Chapter 1 Bibliography

[1] Web Page. URL: http://archives.nato.int/committee-on-challenges-of-modern-society-ccms-2;isaar.

[2] *AFDB Approves $133 Million To Combat Drought In Eritrea, Ethiopia, Somalia, and Sudan*. Web Page. URL: http://www.raimoq.com/afdb-approves-133-million-to-combat-drought-in-eritrea-ethiopia-somalia-and-sudan/.

[3] Patricia M Anderson and Linda B Brubaker. "Vegetation history of northcentral Alaska: a mapped summary of late-Quaternary pollen data". In: *Quaternary Science Reviews* 13.1 (1994), pp. 71–92. ISSN: 0277-3791.

[4] O.A. Anisimov et al. "Polar regions (Arctic and Antarctic)". In: *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by M.L. Parry et al. Cambridge University Press, Cambridge, UK, 2007, Cambridge University Press, Cambridge, UK, 653–685.

[5] Svante Arrhenius and Edward S Holden. "On the Influence of Carbonic Acid in the Air upon the Temperature of the Earth". In: *Publications of the Astronomical Society of the Pacific* 9.54 (1897), pp. 14–24. ISSN: 0004-6280.

[6] Anthony D Barnosky et al. "Assessing the causes of Late Pleistocene extinctions on the continents". In: *Science* 306.5693 (2004), pp. 70–75. ISSN: 0036-8075.

[7] Anthony D Barnosky et al. "Variable impact of late-Quaternary megafaunal extinction in causing ecological state shifts in North and South America". In: *Proceedings of the National Academy of Sciences* 113.4 (2016), pp. 856–861. ISSN: 0027-8424.

[8] Patrick J Boylan. "Lyell and the dilemma of Quaternary glaciation". In: *Geological Society, London, Special Publications* 143.1 (1998), pp. 145–159.

[9] S Brewer, J Guiot, and D Barboni. "Pollen Methods and Studies". In: *Encyclopedia of Quaternary science* (2007), pp. 2497–2508.

[10] Wallace S Broecker. "Was the Younger Dryas triggered by a flood?" In: *Science* 312.5777 (2006), pp. 1146–1148.

[11] Barry W Brook and Anthony D Barnosky. "Quaternary extinctions and their link to climate change". In: *Saving a Million Species*. Springer, 2012, pp. 179–198. ISBN: 1610911822.

[12] David A Burney and Timothy F Flannery. "Fifty millennia of catastrophic extinctions after human contact". In: *Trends in Ecology & Evolution* 20.7 (2005), pp. 395–401.

[13] Peter U Clark and Alan C Mix. "Ice sheets and sea level of the Last Glacial Maximum". In: *Quaternary Science Reviews* 21.1 (2002), pp. 1–7. ISSN: 0277-3791.

[14] Project Members CLIMAP. "The surface of the ice-age Earth." In: *Science (New York, NY)* 191.4232 (1976), p. 1131.

[15] Alan Cooper et al. "Abrupt warming events drove Late Pleistocene Holarctic megafaunal turnover". In: *Science* 349.6248 (2015), pp. 602–606. ISSN: 0036-8075.

[16] L.C. Cwynar. "A late-Quaternary vegetation history from Hanging Lake, northern Yukon". In: *Ecological Monographs* (1982), pp. 2–24. ISSN: 0012-9615.

[17] W Dansgaard et al. “Evidence for general instability of past climate from a 250-kyr”. In: *Nature* 364 (1993), p. 15.

[18] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. 1st ed. London: John Murray, 1859.

[19] Robert S Feranec, Elizabeth A Hadly, and Adina Paytan. “Stable isotopes reveal seasonal competition for resources between late Pleistocene bison (< i> Bison</i>) and horse (< i> Equus</i>) from Rancho La Brea, southern California”. In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 271.1 (2009), pp. 153–160. ISSN: 0031-0182.

[20] Ken L Ferrier, Kimberly L Huppert, and J Taylor Perron. “Climatic control of bedrock river incision”. In: *Nature* 496.7444 (2013), pp. 206–209. ISSN: 0028-0836.

[21] Eunice Foote. “Circumstances Affecting the Heat of the Sun’s Rays”. In: *American Journal of Science* 22 (1856), pp. 382–383.

[22] Billie A Gould et al. “Evidence of a high-Andean, mid-Holocene plant community: an ancient DNA analysis of glacially preserved remains”. In: *American Journal of Botany* 97.9 (2010), pp. 1579–1584. ISSN: 0002-9122.

[23] *Greenland’s Agricultural Boom*. Web Page. URL: `http://www.spiegel.de/fotostrecke/photo-gallery-greenland-s-agricultural-boom-fotostrecke-15903-2.html`.

[24] R Dale Guthrie. “Frozen fauna of the mammoth steppe”. In: *The Story of Blue Babe Chicago University* (1990).

[25] R Dale Guthrie. “Mammals of the mammoth steppe as paleoenvironmental indicators”. In: *Paleoecology of Beringia* (1982), pp. 307–326.

[26] R Dale Guthrie. “New carbon dates link climatic change with human colonization and Pleistocene extinctions”. In: *Nature* 441.7090 (2006), pp. 207–209. ISSN: 0028-0836.

[27] R Dale Guthrie. “Origin and causes of the mammoth steppe: a story of cloud cover, woolly mammal tooth pits, buckles, and inside-out Beringia”. In: *Quaternary Science Reviews* 20.1 (2001), pp. 549–574. ISSN: 0277-3791.

[28] R Dale Guthrie. “Rapid body size decline in Alaskan Pleistocene horses before extinction”. In: *Nature* 426.6963 (2003), pp. 169–171. ISSN: 0028-0836.

[29] C Richard Harington. “Pleistocene vertebrate localities in the Yukon”. In: *Late Cenozoic history of the interior basins of Alaska and the Yukon. US Geological Survey Circular* 1026 (1989), pp. 93–98.

[30] Sidney R Hemming. “Heinrich events: Massive late Pleistocene detritus layers of the North Atlantic and their global climate imprint”. In: *Reviews of Geophysics* 42.1 (2004).

[31] John F Hoffecker et al. “Beringia and the global dispersal of modern humans”. In: *Evolutionary Anthropology: Issues, News, and Reviews* 25.2 (2016), pp. 64–78.

[32] Michael Hofreiter et al. “The future of ancient DNA: Technical advances and conceptual shifts”. In: *BioEssays* 37.3 (2015), pp. 284–293. ISSN: 1521-1878.

[33] David M Hopkins, John V Matthews, and Charles E Schweger. *Paleoecology of Beringia*. Elsevier, 2013.

[34] Eric Hulten. “Outline of the history of arctic and boreal biota during the Quaternary period”. In: (1972).

[35] Chris N Johnson et al. "Rapid megafaunal extinction following human arrival throughout the New World". In: *Quaternary International* 308 (2013), pp. 273–277. ISSN: 1040-6182.

[36] Jean Jouzel et al. "Orbital and millennial Antarctic climate variability over the past 800,000 years". In: *science* 317.5839 (2007), pp. 793–796.

[37] MR Kaplan et al. "Southern Patagonian glacial chronology for the Last Glacial period and implications for Southern Ocean climate". In: *Quaternary Science Reviews* 27.3 (2008), pp. 284–294. ISSN: 0277-3791.

[38] Z. Klimont. "Near-term climate change: Projections and predictability". In: *Climate Change 2013: The Physical Science Basis. IPCC Working Group I Contribution to AR5*. Cambridge University Press, Cambridge, 2013. Chap. 11.

[39] Paul L Koch and Anthony D Barnosky. "Late Quaternary extinctions: state of the debate". In: *Annual Review of Ecology, Evolution, and Systematics* (2006), pp. 215–250. ISSN: 1543-592X.

[40] Lorraine E Lisiecki and Maureen E Raymo. "A Pliocene-Pleistocene stack of 57 globally distributed benthic d18O records". In: *Paleoceanography* 20.1 (2005). ISSN: 1944-9186.

[41] David B Lobell and Sharon M Gourdji. "The influence of climate change on global crop productivity". In: *Plant Physiology* 160.4 (2012), pp. 1686–1697. ISSN: 1532-2548.

[42] Charles Lyell. *Principles of geology: being an attempt to explain the former changes of the earth's surface, by reference to causes now in operation*. Vol. 1. J. Murray, 1832.

[43] GM MacDonald et al. "Pattern of extinction of the woolly mammoth in Beringia". In: *Nature communications* 3 (2012), p. 893.

[44] Daniel H Mann et al. "Ice-age megafauna in Arctic Alaska: extinction, invasion, survival". In: *Quaternary Science Reviews* 70 (2013), pp. 91–108. ISSN: 0277-3791.

[45] Daniel H Mann et al. "Life and extinction of megafauna in the ice-age Arctic". In: *Proceedings of the National Academy of Sciences* 112.46 (2015), pp. 14301–14306. ISSN: 0027-8424.

[46] Daniel H Mann et al. "Responses of an arctic landscape to Lateglacial and early Holocene climatic changes: the importance of moisture". In: *Quaternary Science Reviews* 21.8 (2002), pp. 997–1021. ISSN: 0277-3791.

[47] Joel B Pedro et al. "The spatial extent and dynamics of the Antarctic Cold Reversal". In: *Nature Geoscience* (2015). ISSN: 1752-0894.

[48] Evelyn C Pielou. *After the ice age: the return of life to glaciated North America*. University of Chicago Press, 2008. ISBN: 0226668096.

[49] Graham W Prescott et al. "Quantitative global analysis of the role of climate and people in explaining late Quaternary megafaunal extinctions". In: *Proceedings of the National Academy of Sciences* 109.12 (2012), pp. 4527–4531. ISSN: 0027-8424.

[50] CLIMAP Project. *Seasonal reconstructions of the Earth's surface at the last glacial maximum*. Geological Society of America, 1981.

[51] David A Randall et al. "Climate models and their evaluation". In: *Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*. Cambridge University Press, 2007, pp. 589–662.

[52] Sune O Rasmussen et al. "Dating, synthesis, and interpretation of palaeoclimatic records of the Last Glacial cycle and model-data integration: advances by the INTIMATE (INTegration of Ice-core, MArine and TErrestrial records) COST Action ES0907". In: *Quaternary Science Reviews* 106 (2014), pp. 1–13.

[53] Hans Renssen et al. "Multiple causes of the Younger Dryas cold period". In: *Nature Geoscience* (2015). ISSN: 1752-0894.

[54] Christopher Sandom et al. "Global late Quaternary megafauna extinctions linked to humans, not climate change". In: 281.1787 (2014), p. 20133254.

[55] Beth Shapiro et al. "Rise and Fall of the Beringian Steppe Bison". In: *Science* 306.5701 (2004), pp. 1561–1565. DOI: 10.1126/science.1101074. URL: http://www.sciencemag.org/content/306/5701/1561.abstract.

[56] Julien Soubrier et al. "Early cave art and ancient DNA record the origin of European bison". In: *Nature Communications* 7 (2016), p. 13158. ISSN: 2041-1723.

[57] Chris Stringer et al. "Human migration: Climate and the peopling of the world". In: *Nature* 538.7623 (2016), pp. 49–50.

[58] Solomon Susan. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. Vol. 4. Cambridge University Press, 2007. ISBN: 0521705967.

[59] John Tyndall. "Note on the transmission of radiant heat through gaseous bodies". In: *Proceedings of the Royal Society of London* 10 (1859), pp. 37–39. ISSN: 0370-1662.

[60] Jesse C Vermaire and Les C Cwynar. "A revised late-Quaternary vegetation history of the unglaciated southwestern Yukon Territory, Canada, from Antifreeze and Eikland ponds". In: *Canadian Journal of Earth Sciences* 47.1 (2010), pp. 75–88. ISSN: 0008-4077.

[61] Robert H Whittaker. "Classification of natural communities". In: *The Botanical Review* 28.1 (1962), pp. 1–239. ISSN: 0006-8101.

[62] Eske Willerslev et al. "Fifty thousand years of Arctic vegetation and megafaunal diet". In: *Nature* 506.7486 (2014), pp. 47–51. ISSN: 0028-0836.

[63] Stephen Wroe et al. "Megafaunal extinction in the late Quaternary and the global overkill hypothesis". In: *Alcheringa* 28.1 (2004), pp. 291–331.

[64] G.D. Zazula et al. "Arctic ground squirrels of the mammoth-steppe: paleoecology of Late Pleistocene middens (~24000–29450 ^{14}C yr BP), Yukon Territory, Canada". In: *Quaternary Science Reviews* 26.7 (2007), pp. 979–1003. ISSN: 0277-3791.

[65] G.D. Zazula et al. "Vegetation buried under Dawson tephra (25,300 ^{14}C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 242.3 (2006), pp. 253–286. ISSN: 0031-0182.

[66] Grant D Zazula et al. "Early Wisconsinan (MIS 4) Arctic ground squirrel middens and a squirrel-eye-view of the mammoth-steppe". In: *Quaternary Science Reviews* 30.17 (2011), pp. 2220–2237. ISSN: 0277-3791.

[67] Grant D Zazula et al. “Palaeobotany: Ice-age steppe vegetation in east Beringia”. In: *Nature* 423.6940 (2003), pp. 603–603. ISSN: 0028-0836.

Statement of Authorship

Title of Paper	Megafaunal isotopes reveal role of increased moisture on rangeland during Late Pleistocene extinctions
Publication Status	☐ Published ☒ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted work written in manuscript style
Publication Details	M. Timothy Rabanus-Wallace, Matthew J. Wooller, Grant D. Zazula, Elen Shute, A. Hope Jahren, Pavel Kosintsev, James A. Burns' James Breen, Bastien Llamas, Alan Cooper. 2017. *Megafaunal isotopes reveal role of increased moisture on rangeland during Late Pleistocene extinctions* (in press, Nature Ecology and Evolution)

Principal Author

Name of Principal Author (Candidate)	Mark Timothy Rabanus-Wallace
Contribution to the Paper	Compiled the data, conceived and implemented the analysis, contributed to interpretation, led the writing, and made the figures.
Overall percentage (%)	50
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.
Signature	Date 24/02/2017

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Alan Cooper
Contribution to the Paper	Conceived project, collected samples, coordinated laboratory work, contributed to writing and interpretation
Signature	Date 24/02/2017

Name of Co-Author	Matther Wooler
Contribution to the Paper	Contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	Grant Zazula
Contribution to the Paper	Provided samples, contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	Elen Shute
Contribution to the Paper	Compiled data, contributed to analysis, writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	Hope Jahren
Contribution to the Paper	Performed laboratory analysis, contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	Pavel Kosintsev
Contribution to the Paper	Provided samples and contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	James A. Burns
Contribution to the Paper	Provided samples and contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	James Breen
Contribution to the Paper	Contributed to writing and interpretation.
Signature	Date 24/02/2017

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Contributed to writing and interpretation.		
Signature		Date	24/02/2017

Chapter 2

Nitrogen isotopes illuminate the influence of increased moisture on rangeland megafauna during Late Pleistocene extinctions

---

The following chapter presents a manuscript currently accepted for publication by Nature Ecology and Evolution. The version included is up to date as of the 1st of July, 2017. The manuscript and the Supplementary Materials are included. The complete dataset is available in the Data Supplement to this thesis. The paper discusses the implications of changes in the nitrogen isotopic composition of herbivore collagen over time, as revealed by a dataset compiled from the literature and including several yet-unpublished data collected in the course of previous studies at the Australian Centre for Ancient DNA (ACAD). A new smoothing method appropriate to the isotope data used is described in detail in the Supplementary Materials (SM). This method was designed specifically to identify potential trends in the isotopic data over time while giving a fair graphical representation of the uncertainty in the reconstruction at different points in time.

Chapter 3

Ancient Fruits In Permafrost-Preserved Squirrel Nests

Specimen Processing Methods In The Field And Laboratory

The following chapter describes the samples used in the studies conducted in chapters 3, 4 and 5, and all the non-molecular techniques used in the field and laboratory for handling and processing the samples. These techniques differ from standard field and laboratory procedures owing to the heightened need to prevent contamination and DNA degradation. Much of the information on field practices is based upon the advice of Dr. Grant Zazula, Elizabeth Hall, and Sue Hewittson with the Yukon Paleontology Program (YPP).

3.1 Introduction

3.1.1 Ancient Squirrel Nests In The Klondike Permafrost

The placer mines of the North American and Siberian permafrost zones usually comprise a stretch of exposed permafrost cut from the side of a north- or east-facing hillslope or narrow river valley. Once the centre of a major nineteenth century gold rush, miners in the Yukon Territory's Klondike region spray the exposures with high-pressure monitors throughout the warm season to release layers of ancient stream gravel (FIG. 3.1), from which gold can be extracted by sluicing. This also exposes the fossils of a diverse range of fauna and flora that inhabited the region over the last several million years [3, 7, 20]. Permafrost-preserved fossils can be discovered in situ, and much of the material is washed into the streams created by the monitor. Many of the fossils can be dated directly by radiocarbon dating, but the

FIGURE 3.1: Monitoring a permafrost exposure at Quartz Creek, Yukon Territory, Canada. Image courtesy of the Government of Yukon.

presence of dated volcanic ash layers (or tepthra) through the permafrost also allows constraints to be placed on fossils' ages even when they are older than the radiocarbon dating limit of ~50,000 years [3, 12, 13].

Along with the remains of many mammoth steppe megafauna, placer mines also yield the preserved nests and burrows of the Arctic ground squirrel (*Spermophilus parryii*). The males of this circumboreal species cache seeds and fruits in underground *middens*—sections of the nest dedicated to storage—for fast sustenance entering the mating season after hibernation through the winter [17, 19].

Despite high foraging specificity for particular plant species, the diversity of fruits of local plants found in preserved nests is high, and represents a small natural library of the mammoth steppe flora. These samples and have proven invaluable in the reconstruction of this novel biome in the past, suggesting the squirrels inhabited a remarkably productive, forb- and graminoid-dominated meadow environment, analogous to the treeless south-facing hillslopes that enjoy a high amount of radiant sun energy during summer months in the region today[17, 19]. Squirrel populations underwent repeated habitat fragmentation and range contractions/expansions during the repeated glacial-interglacial cycles that affected their Nearctic range in the Pleistocene [2, 9]. Nevertheless, the contents of the squirrel nests over time appears to remain comparatively static [5], suggesting their biology and the composition of the flora in their well-drained meadow and hill-slope habitats has adapted well

to glacial-interglacial range shifts—a central theme in chapter 5.

3.1.2 Ancient DNA Preservation

The primary goal of the methods described here is to extract DNA from permafrost-preserved seeds to enable genetic study (Chapters 4–5). The challenges and hazards of working with ancient DNA (aDNA) are more fully described in section 4.1.1. Briefly, the DNA is expected to be highly fragmented, very low in quantity, and mixed with a large amount of contaminant DNA from microbial and other sources.

One critical aim was to keep the samples cold such that post-collection DNA fragmentation is minimised, since this reduces the amount of endogenous DNA that is amenable to sequencing and analysis [4].

A related aim was to avoid contaminating the samples with human DNA or other sources of modern DNA from the environment. Pollen DNA contamination was a distinct worry, since some of the taxa sampled have close relatives or even descendants in the Klondike area, and the samples are collected during the flowering season. Being located within a botanical garden, the ACAD ancient DNA laboratory also bears an elevated risk of contamination by a wide range of plant taxa. Contaminant human DNA is unlikely to affect the result of a study into plant genetics, but samples in an aDNA laboratory may be used in many different projects over several decades, and such contamination could certainly affect the reliability of a future metagenomic study involving non-plant specimens [8, 10, 11, 14].

A final aim was to keep the samples themselves (or genetic material deposited on them during handling) from contaminating an aDNA laboratory. The aDNA laboratory is shared with researchers working on many projects, and so DNA sources introduces to the laboratory must avoid affecting any other study [8, 10]. This is particularly important given the metagenomic studies taking place at ACAD [16]. Metagenomic studies make a broad survey of all the DNA in their samples, and occasionally draw inferences based on the presence of novel sequences in small quantities, and such inferences can only be trusted when cross-contamination between samples in different projects can be adequately ruled out.

3.2 Methods

3.2.1 Field Handling

Working a mine site in the Yukon requires coordination with the miners to ensure safety standards are met, and with the Yukon Department of Paleontology, which *"serves as the Expert Examiner for the exportation of all Yukon fossil remains (including fossil ivory) from Yukon under the Canadian federal Cultural Property Export and Import Act"* [1].

While retrieving permafrost-preserved squirrel nests for aDNA work, I developed and applied the following procedure, which aims to keep the nests largely frozen (for optimal DNA preservation), while minimising avenues for contamination [8]. Safety gear (helmet, reflective vest etc.) was worn at all times, and the following equipment was used:

- Large cold box.
- Backpack.
- 3–4 frozen freezer blocks per nest.
- Large industrial zip-lock sample bags, at least three per nest pre-
- labeled in permanent marker in two separate places each.
- 1% bleach solution, at least 1 L.
- Disposable laboratory paper towel.
- A club hammer.
- At least one chisel. Cold chisels, bolster chisels, and paring chisels are all useful.
- Laboratory gloves.
- Laboratory facemask.

Nests in permafrost can be identified as tufts of red-brown nesting material protruding from the permafrost [19]. Burrows appear as round tubes with a diameter of ~8–15 cm, and often appear in the vicinity of nests. Burrows may partially collapse and be filled in with mud, leaving characteristic circular divets at the surface. Alongside regular field notes for each sample, factors that may affect DNA preservation and contamination were recorded, including notes on the time since exposure, the temperature, and any likely sources of contamination.

A face mask and gloves were worn whenever working with or near the nests. All equipment was bleached and wiped with laboratory towel before use. The outer surface of the nests are often thawed; In such cases the thawed surface was scraped into a sample bag and stored for morphological work only, since thawing negatively impacts the survival of DNA.

The thawed surface was washed away by pouring bleach solution over the exposed nest surface. This helps to reveal the shape of the nest. The nests were then removed from the permafrost using the hammer and chisel. Difficult nests can be chipped into large fragments by driving the chisel directly into the permafrost beside the nest. The flakes were rinsed in bleach to remove loose mud and placed in the sample bag. With the bulk of the nest in the sample bag, the outside of the bags were rinsed again in bleach and a second bag was placed around the first and sealed. Placed together in a cooler box with freezer blocks, nests in bags will remain frozen for over ten hours.

3.2.2 Processing For Transport

Before transport, the inner bags containing the nests were washed thoroughly with a hose, then rinsed with bleach solution, to ensure no leftover material was stuck to the outside of the bags. A large (20–50 L) tub of 1% bleach solution was prepared in advance, and transferred to the bags with a plastic bottle. The nests were then packed into cleaned and bleached cooler boxes filled with around 1/3rd volume with freezer blocks and 2/3rds with nests.

3.2.3 Laboratory Methods

3.2.3.1 Entry Procedure

Transferring nest samples into an aDNA clean laboratory (see chapter 4) is a particular challenge. With bone samples, it is standard practice to remove the sample from its container and bleach the surface as part of the entry procedure, but this would no doubt cause more contamination than it saved for the nest material, which had at this stage already been rinsed in bleach solution when first placed in the sample bag.

The inner bags were therefore left closed throughout the procedure. At first, the cooler boxes were cleaned with 3% bleach solution outside the laboratory, the contents sprayed lightly with bleach solution, and the freezer blocks removed. The bags were passed one at a time into the entry room of the lab, where the outer bags were discarded, replaced, and then treated with bleach and ultraviolet light before storage in a freezer.

3.2.3.2 Retrieving Nest Contents

Defrosting and filtering the nests is necessary to work with the contents. Chilled ethanol was used to defrost the fragments and wash the loess from the biological contents of the nest. This was intended to keep the sample cool while defrosting, since ethanol remains liquid at sub-zero temperatures.

This is a tricky procedure, and preparation is key. The bone grinding room of the aDNA laboratory was used, which is standard procedure for higher-contamination-risk activities. This area permits access to Reverse Osmosis (RO) filtered water, a sink, and much bench space.

The loess surrounding and within the nests is very mobile and viscous when wet, but dries quickly to become extremely fine dust. Occasional spills were sprayed directly with bleach solution, then wiped up with paper towel.

The following was set up in the workspace:

- Empty bins.

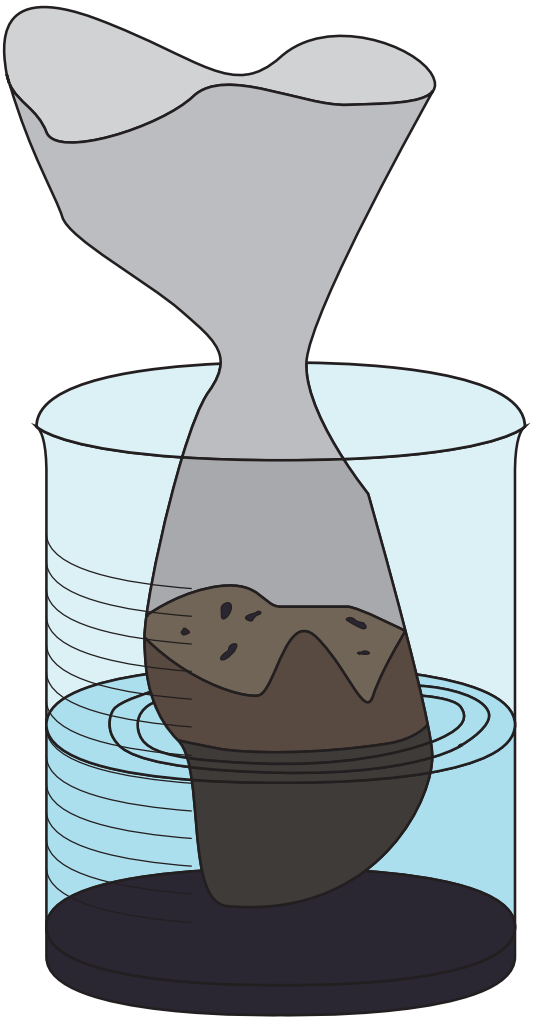

FIGURE 3.2: Filtering the sample in a bath of sub-zero ethanol. A few minutes of agitation by moving the mesh around releases the fine loess, which accumulates in the bottom of the beaker.

- Workspaces covered in paper towel.
- Water purifier containing at least 5L of water.
- A full 1–1.5L spray bottle of 3% bleach solution and 3+ rolls of paper towel. At least half a litre of ethanol bench spray.
- A screwdriver or very small chisel, and light hammer.
- A 0.5L bottle of ultrapure 100% ethanol kept at ~-20.
- Two large plastic tubs (~10L each; bleached and ultraviolet treated). These are always useful for temporarily storing dirty sample material while working.
- Thermometer (bleached and ultraviolet treated).

- The drying tray: a small, washing-basket style tray (about 15×20 cm, with a flat bottom and holes in the sides to permit air flow; bleached and ultraviolet treated). Five to ten layers of paper towel should line the bottom, with another two sheets put aside to place over the top once the sample is added.
- 1 mm fiberglass mesh circle, cut into a circle with a radius of ~15–20 cm (bleached and ultraviolet treated).
- Large ziplock bags.
- Liquid waste containers.
- Shelf space in a ~4°C cool room.

The hammer and screwdriver were used to break up nest without removing it from its bags. Doing this in a tub ensured a mineral risk of sediment escaping and contaminating the room. Aiming through the bag's opening, the screwdriver was positioned against the nest in the inner bag with the intention of breaking off large portions (~100 cm^3). The other two bags were held tightly around the shaft of the screwdriver to seal off the nest from the room. Portions of nest were broken away by tapping the screwdriver's handle lightly many times with the hammer. This was repeated until approximately 200 g of a nest was reduced to small chunks. The bags were checked throughly and damaged bags were replaced.

Approximately ~2–5 nest fragments were chosen for defrosting. A fragment was considered promising if the exposed surface suggested the fragment may contain the desired subfossils. When a nest is located, this is usually evident from the densely-packed fruits extending from the surface of the fragment, often with a single species dominating.

The cold ethanol was brought to the working room and 200–500 ml poured into the beaker. The mesh was folded into a cone and placed with the tip in the ethanol. Nest fragments were transferred to the cone. The circumference of the mesh sheet was 'clumped' together to form a loose bag containing the sample fragments, immersed in the ethanol (FIG. 3.2). The bag was agitated for 5–10 m, allowing thawed sedimentary material to pass through the mesh. The material left in the mesh bag after filtration was usually only a small portion of the original size. To keep the ethanol's temperature low, it was placed upon a cold block from the freezer. The thermometer was used to monitor the temperature. Before repeating the process with the remaining sample fragments, filtrate in the mesh was wrung out by squeezing from the top using a few layers of paper towel, and the filtered material was transferred to the drying tray, using the mesh to spread it evenly over the paper towel. The drying tray was also placed on a cold block.

Once all the sample fragments were processed, the drying tray was moved to the coolroom (with paper towel covering). The cleanup procedure had to ensure that no loess was disposed of in sinks, in case of blockage. Therefore all loess was transferred to the bins using paper towel and bleach spray.

FIGURE 3.3: **Left:** Selected samples from the Klondike nests used in the current study. Scale bars represent ~5mm. **Right:** High-resolution images of samples from the same taxa, published in [17, 18], and used to aid in identify the various taxa (see text).

Drying was usually complete after two nights, and the nest contents were transferred to a sample bag or a 50 ml plastic tube.

3.2.3.3 Microscopy And Sample Selection

Identifying nest contents was performed by sifting through small amounts (2–3 g) of material on a microscope dish under a dissecting light microscope. The nest contents were kept cool by placing a microscope dish on a freezer block. A pair of scalpel blades were found to be better than tweezers for manipulating the nest contents: Even touching a fruit gently with the tip of the blade can make it stick there for easy transfer to a 1.5 mL Eppindorph tube, and this usually does less damage to the sample than tweezers.

The tubes containing samples can be stored in a coolroom, and used in the extraction step directly. This procedure is described in chapter 4.

3.3 Conclusions

Using the methods described, it is possible to handle permafrost-preserved squirrel nests and extract their contents in an aDNA laboratory, while keeping the contents at low temperatures and without contravening normal aDNA contamination standards [8]. Future work on this kind of material might benefit from increased throughput, either by scaling up the procedure, or by processing samples in parallel. Further research into the most common contaminants that affect this material may allow some precautionary measures to be skipped, further improving throughput, for instance, filtration may be safe to perform outside a laboratory environment if decontamination is effective in cleaning the samples afterwards. The benefits of keeping the samples frozen, as opposed to simply allowing them to reach room temperature, also needs assessment, but is likely indisposable if the samples are to be used in the promising new pursuit of ancient botanical RNA extraction [6, 15]. This is especially important given the findings of the following chapters, which recommend screening many samples for genetic work since DNA preservation is nest- and taxon-dependent—and the best genera for sequencing may be rare.

Chapter 3 Bibliography

[1] Web Page. URL: http://www.tc.gov.yk.ca/palaeontology.html.

[2] Aren A Eddingsaas et al. "Evolutionary history of the arctic ground squirrel (Spermophilus parryii) in Nearctic Beringia". In: *Journal of Mammalogy* 85.4 (2004), pp. 601–610. ISSN: 0022-2372.

[3] Duane G Froese et al. "The Klondike goldfields and Pleistocene environments of Beringia". In: *GSA Today* 19.8 (2009), p. 5. ISSN: 1052-5173.

[4] Marc Garcia-Garcera et al. "Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics". In: *PLoS One* 6.8 (2011), e24161. ISSN: 1932-6203.

[5] Elizabeth A Gillis et al. "Evidence for selective caching by arctic ground squirrels living in alpine meadows in the Yukon". In: *Arctic* (2005), pp. 354–360. ISSN: 0004-0843.

[6] Paul L Guy. "Prospects for analyzing ancient RNA in preserved materials". In: *Wiley Interdisciplinary Reviews: RNA* 5.1 (2014), pp. 87–94.

[7] C Richard Harington. "Pleistocene vertebrate localities in the Yukon". In: *Late Cenozoic history of the interior basins of Alaska and the Yukon. US Geological Survey Circular* 1026 (1989), pp. 93–98.

[8] Bastien Llamas et al. "From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era". In: *STAR: Science and Technology of Archaeological Research* 3.1 (2016), pp. 1–14.

[9] Charles F Nadler and Robert S Hoffmann. "Patterns of evolution and migration in the arctic ground squirrel, Spermophilus parryii (Richardson)". In: *Canadian Journal of Zoology* 55.4 (1977), pp. 748–758. ISSN: 0008-4301.

[10] Elena Pilli et al. "Monitoring DNA contamination in handled vs. directly excavated ancient human skeletal remains". In: *PLoS One* 8.1 (2013), e52524.

[11] Hendrik N Poinar et al. "Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA". In: *Science* 311.5759 (2006), pp. 392–394. ISSN: 0036-8075.

[12] Shari J Preece et al. "Characterization, identity, distribution, and source of late Cenozoic tephra beds in the Klondike district of the Yukon, Canada". In: *Canadian Journal of Earth Sciences* 37.7 (2000), pp. 983–996. ISSN: 0008-4077.

[13] AS Sandhu et al. "Glass-fission-track ages of Late Cenozoic distal tephra beds in the Klondike district, Yukon Territory". In: *Yukon exploration and geology* (2000), pp. 247–256.

[14] Pontus Skoglund et al. "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pp. 2229–2234.

[15] Oliver Smith et al. "A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus". In: *Scientific reports* 4 (2014), p. 4003.

[16] Laura S Weyrich et al. "Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus". In: *Nature* 544.7650 (2017), pp. 357–361.

[17] G.D. Zazula et al. "Arctic ground squirrels of the mammoth-steppe: paleoecology of Late Pleistocene middens (~24000–29450 ^{14}C yr BP), Yukon Territory, Canada". In: *Quaternary Science Reviews* 26.7 (2007), pp. 979–1003. ISSN: 0277-3791.

[18] G.D. Zazula et al. "Vegetation buried under Dawson tephra (25,300 ^{14}C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 242.3 (2006), pp. 253–286. ISSN: 0031-0182.

[19] Grant D Zazula et al. "Early Wisconsinan (MIS 4) Arctic ground squirrel middens and a squirrel-eye-view of the mammoth-steppe". In: *Quaternary Science Reviews* 30.17 (2011), pp. 2220–2237. ISSN: 0277-3791.

[20] Grant D Zazula et al. "Palaeobotany: Ice-age steppe vegetation in east Beringia". In: *Nature* 423.6940 (2003), pp. 603–603. ISSN: 0028-0836.

Chapter 4

Ancient DNA from Permafrost-Preserved Plant Material

The following chapter describes investigations into the quantity and quality of ancient DNA present in the fruits of Beringian plants, collected in the field and processed in the laboratory according to the methods described in the previous chapter. This was achieved by extracting, shotgun sequencing, and reference-mapping the DNA, then relating many different variables (e.g. sample age, species, laboratory methodology) to the end results. The results are summarised in a number of ways, including visualising the length distributions of reads, and summarising their useful information content using a new metric, the coverage potential score. *Relevant background on ancient DNA and High Throughput Sequencing (HTS) and is given. A basic understanding of Sanger sequencing and PCR analysis are assumed.*

4.1 Introduction

4.1.1 Ancient DNA

The field of ancient DNA (aDNA) first gained recognition in the early eighties with the sequencing of a 229 nt DNA amplicon from an extinct zebra [30], and has since grown to become a prominent source of information for paleontologists, archaeologists, and evolutionary biologists [31]. Today, over 70 aDNA labs worldwide routinely sequence partial and occasionally complete ancient genomes of organisms ranging from microbes to human ancestors [10]. The challenges of working with aDNA stem from the joint processes of contamination and degradation [46, 57]. Commonly, the bulk of the DNA in an ancient sample originates not from the sample organism itself, but from microbes that colonise the tissue post mortem. Other sources of contaminant DNA are the surrounding environment (often soil), researchers handling the sample, airborne DNA (in pollen, fungal spores, or microdroplets—which are ubiquitous in DNA laboratories), and reagent contamination [17].

The swamping of endogenous DNA with contamination is compounded by the endogenous DNA's degradation. Hydrolytic, oxidative, ultraviolet radiation, and enzymatic degradation all affect DNA molecules [27, 28, 67]. After the death of an organism, DNA preservation and repair mechanisms decay, leaving the DNA exposed to degradation by nucleases. Continual fragmentation gives aDNA a short average fragment length, which correlates approximately with the thermal age of a sample [62], and is suspected to impose an upper limit on DNA retrieval of perhaps slightly over one million years [52]. Alongside oxidative, hydrolytic, and radiative damage to the phosphodiester backbone, a principle mechanism of DNA fragmentation is depurination—the hydrolytic excision of purine bases—which interferes with replication when the DNA is amplified [21]. Hydrolysis also deaminates cytosine nucleotides, cleaving off the 4′ amine group and effectively converting them to uracils [62]. Depurination and cytosine deamination have very specific consequences when the DNA is sequenced, as discussed below.

4.1.1.1 High Throughput Sequencing

When Sanger sequencing aDNA, cloning can be used to achieve independent replication of multiple copies of a sequence, allowing fragments with damage-induced substitutions to be distinguished from fragments retaining the true biological sequence [31]. The advent of High Throughput Sequencing (HTS) rapidly advanced the field, offering a simple mechanism for sequencing many fragments simultaneously while massively increasing sequence output [65]. The Illumina MiSeq platform (reagent kit v3, 300 cycle paired end) can produce up to 15 gigabases of sequence data in around two days, while the HiSeq 2500 (SBS v4, 125 cycle paired end, dual flow cells) may produce up to a terabase in under a week [2].

The Illumina platforms that currently dominate the field of aDNA use *Sequencing-By-Synthesis* (SBS)-based visualisation [65]. In brief, DNA strands are extended one base at a time using terminating bases, each labelled with a fluorescent tag. At each extension, the tags are photographed, before being chemically removed, along with the terminator, in preparation for subsequent cycles.

To interact with the sequencing hardware, DNA fragments must be flanked with double-stranded DNA *adapter* sequences, yielding a *library* of sequenceable DNA fragments. Ancient DNA often has nicks in the phosphodiester backbone, and many fragments have overhanging 3′ and 5′ single stranded regions (in which deamination is particularly common). In the standard aDNA protocol pubished by Meyer et al. in 2010 [50], prior to blunt-end adapter ligation, the sequences are repaired by polymerases that displace and copy DNA downstream from nicks, extend over 5′ overhangs, and remove 3′ overhangs. Enzymatic phosphorylation of the terminal 5′ hydroxyl groups allows a ligase enzyme to join this strand to the blunt end of an unphosphorylated adapter. To ensure the correct polarity, adapters are added with one blunt and one overhanging end. To ensure that fragments are primarily ligated to adapters—and not to one and another—the concentration of adapters is calculated to greatly exceed that of DNA fragments. A strand-displacing polymerase is then

used to fill the nick that remains on the unligated strand, and further PCR amplifications fill in the overhanging ends of the adapters.

During sequencing, adapter sequences provide sites for *sequencing primers* to anneal in order to begin extension, along with regions that bind the DNA to a *flow cell*. The flat surface of the flow cell allows a high-resolution camera to detect the light emitted from "colonies" of identical DNA sequences that extend simultaneously, one base at a time, for a set number of cycles. An algorithm interprets the camera's images and calls a base for each colony based upon the frequency and position of light emitted from the flowcell. Since colonies may overlap, go out of focus, or lose synchronisation between strands, there is a degree of uncertainty in each call, which is quantified as a quality score for each base. After the cycles are completed, the sequence from each original fragment of DNA is termed a *read* [13, 49].

Since fragments are often sequenced from both ends, it is common for the paired reads to overlap, in which case they can be merged by an algorithm that detects the overlapping part [64]. DNA fragments that are shorter than the read length cause the sequencing read to extend right through the fragment, and into the opposing adapter. Adapter contamination can largely be removed from the data by sequence recognition algorithms. It is common also to remove sequences, or ends of sequences, that have low quality scores, since quality scores tend to wane as cycles progress [13, 49].

Adapters are often designed to include unique identifying sequences, typically 5–7 nt, that are useful for distinguishing between samples when they are sequenced simultaneously in a multiplexed run. Such a sequence is known as an *index*. Classifying and separating the reads according to their indicies is known as *demultiplexing*.

4.1.1.2 Mapping and Assembly with Shotgun Sequencing Reads

Shotgun sequencing is the name given to sequencing fragments of the total DNA in a sample somewhat randomly, that is, without targeting any particular genes or loci. This randomness is occasionally an advantage, for instance, in metagenomic applications, where identifying sequences in a mixed-origin DNA sample can function as a proxy for the composition of organisms in the sample [45].

Shotgun sequencing also allows the reconstruction of long stretches of a genome by combining overlapping sequence fragments. When enough data are available, *de novo* assembly algorithms can combine short reads to infer longer—often very much longer—contiguous sequences, known as *contigs* [83]. More commonly, the reads are aligned to a *reference* genome—the sequenced genome of a closely-related organism—that is assumed to be only slightly different to the sequenced organism [43]. Mapping the reads to a reference genome is an effective way to identify homologous fragments and reconstruct parts of a novel genome.

PCR amplification of the libraries produces multiple copies of the same fragment, and putative clonal (or duplicate) reads can be identified in the mapping alignment by their shared mapping positions. Duplicate reads are common in aDNA owing to the low amounts of

starting template, and are normally removed to avoid biasing downstream steps. Mapped reads are commonly assigned a mapping quality score [42], which reflects the confidence of the algorithm that the read sequence has been correctly aligned to its homolog. Mapping quality scores can assist a variant-calling algorithm to list mutations (SNPs or indels) that differ between the reference organism and the sequenced organism, and from these variants a consensus sequence can be made, which ideally represents the entire homologous genome of the sequenced organism. In practice, gaps, errors, and ambiguities in the consensus sequence can be very common.

4.1.1.3 NGS and DNA Degradation

Mapping reads to a reference genome offers unique opportunities to investigate deamination and depurination in aDNA [15, 51, 53, 62] (see section 4.1.1). In fact, evidence of DNA degradation is now a standard test for establishing the authenticity of aDNA. Cytosine deamination occurs primarily in single-stranded overhangs. The 5′ overhangs are copied during library preparation so, since deamination converts cytosine, C, to uracil , U (an analogue of thymine, T), the polymerase copying the overhang pairs these converted bases with adenosine, A, rather than the original guanine, G. In later rounds of amplification, this A is naturally paired with T. The mapped reads therefore show increasing numbers of C-to-T substitutions toward the 5′ end of the fragments, and G-to-A mismatches in the 3′ ends [62].

Furthermore, depurination converts purines (A and G) to abasic sites. Single stranded overhangs can be filled up to the first abasic site on the template strand. Since purines are complementary to pyrimidines, a depurination event in a 5′ single-stranded overhang will result in the extending strand terminating immediately before a pyrimidine (C or T). As a result, reads mapped to a reference genome have an increased tendency to end before a pyrimidine in the reference, or, when the reverse strand is sequenced, to begin after a purine (G or A) [15].

Scripts such as MapDamage 2.0 [34] are used to visualise these patterns graphically (see figure 4.2, chapter 4).

4.1.1.4 Enrichment with Liquid Baits

Regular shotgun sequencing does not target particular loci, which is a disadvantage when much of the total DNA in a sample is not useful for the desired study. Hybridisation enrichment is one of several similar means of targeting particular sequences, for instance, mitochondrial DNA, or Coding DNA Sequences (CDS). Enrichment can be of great use in aDNA studies, where the DNA of interest is often heavily diluted by contaminant DNA [14, 16].

The liquid RNA baits method of sequencing library enrichment relies upon binding the total extracted DNA to RNA oligonucleotide *baits*, which are designed (using a reference genome) to be similar to sequences of interest. An incubation period, during which the temperature of the reaction is often slowly decreased, ostensibly ensures that the baits preferentially

bind those DNA fragments with the most similar sequences. The DNA-RNA duplexes are then immobilised on streptavidin magnetic beads, by means of biotin molecules bound to the RNA. The immobilised duplexes are washed to remove non-target DNA. The remaining DNA is then amplified. The post-enrichment DNA pool usually represents a very small fraction of the starting DNA. Amplifying from these ultra-low concentrations results in highly clonal libraries with strong biases in coverage that may favour sequences with particular sequence motifs, lengths, or G/C contents [19, 20]. The method is comparatively new and experimental data are limited. Optimisation of factors such as the bait length, the hybridisation conditions, and the stringency of the washing steps is often required to perfect the specificity and improve highly-variable results [14, 16, 18, 19, 23, 26, 47, 54, 73].

4.1.1.5 Ancient DNA from Plants

Amplicon sequencing has been successfully performed on ancient plant materials including wood, seeds, rhyzomes, and pollen [27, 36, 55, 63]. A surge of interest in ancient plant DNA beginning in the mid-21st century has focused heavily on domesticated plants such as barley, maize, gourds, and melons [32, 37, 48, 56]. Shotgun sequencing of DNA from individual plants has been largely confined almost exclusively to Holocene-age samples, but HTS has been applied to barcoding amplicons from environmental and coprolite samples to yield metagenomic profiles of plant communities in samples dating to the Late Pleistocene [69, 76, 77]. The results of such studies are characterised by high variablility between samples: currently, the only study to attempt whole genomic analysis using ancient plant shotgun sequence [48] estimated endogenous DNA content in mid-Holocene barley grains from the same cave site to range between 0.4 to 96.4%, with those having high endogenous content having a mean mapped fragment length ~40–75 nt longer than the low-content samples. To date, several ancient chloroplast genomes have been recovered using enrichment techniques, all dating to within the Holocene epoch (< 10 kya) [37].

4.2 Methods

This chapter describes the shotgun sequencing of plant aDNA from fruits prepared as described in chapter 3.

4.2.1 Sample Selection And Radiocarbon Dating

Eighteen nests in total were wholly or partially processed as described in chapter 3, of which eight contributed to the datasets used in subsequent chapters. Nests whose identification codes begin with "QC" were collected as part of this project, and all the techniques and precautions described in chapter 3 were applied. The remainder of the nests were collected on previous expeditions, and the details of the field handling and DNA lab entry procedures are unknown. Subsequent procedures were performed as described in chapter 3. All eight

Nest ID	OxCal ID	δ^{13}C	δ^{14}C Date	Error	95% CI Upper	95% CI Lower	Mid-point
QC5	OxA-30860	-26.91	44000	1200	47878	43580	45729
GZ0907	OxA-30933	-25.54	25530	240	28458	27094	27776
GZ1103	OxA-30934	-25.02	>49000				
DF1042	OxA-32762	-27.49	32090	280	34646	33391	34018.5
TK1127	OxA-32764	-26.23	31940	340	34597	33111	33854
TK1026	OxA-32766	-25.66	>47800				
QC8	OxA-32767	-27.06	>49000				
GZ0908 no EtOH	OxA-32763	-24.39	25170	160	27661	26861	27261
GZ0908 with EtOH	OxA-32945	-24.62	25150	180	27679	26812	27245.5

TABLE 4.1: Radiocarbon dating results from squirrel nest material.

nests contributing to the final dataset originated from Quartz Creek mine. Nesting material or midden contents from these contributing nests were then sent to Oxford Radiocarbon Unit for carbon-14 dating (Table 4.1), and calibrated using OxCal [3]. The dates are consistent with the age of the sediments at Quartz Creek, whose dates can be constrained to periods before or after ~30 ka and ~80 ka, owing to the presence of the Dawson and Sheep Creek tephra beds, respectively[24, 33, 74]. In order to assure that the use of ethanol to filter and dry the nest was not affecting the dating process by introducing contaminant sources of carbon, two separate preparations of nesting material GZ0908 were made, with and without ethanol treatment. As table 4.1 shows, the resulting dates are effectively identical.

Species were chosen for inclusion in the study based upon their presence in multiple nests (to allow comparison beween nests of different ages), the homogeneity of their morphologies across nests (to ensure they were the same species), the numbers in which they occurred (to allow replication within each nest), and the desire to use the same set of nests for coverage of all species. Three taxa were selected for experimentation using high-throughput sequencing (FIG. 3.3), identified morphologically as *Draba sp.* (Brassicaceae), *Ranunculus sp.* (Ranunculaceae), and *Bistorta vivpara* (Polygonaceae). Squirrel remains were also retrieved from three of the nests. In the course of assessing the viability of continued study on various taxa, short aDNA amplicons from several other species were also sequenced, including the squirrel, and these results are summarised in the following chapter (section 4.2). Radiocarbon dates associated with the ancient samples are also given (Table 4.1).

Many herbarium specimens were also included in the study, and metadata on these samples is available in the data supplement (DS:Samples). These genetic investigations made use of 112 extracted samples from the genera *Bistorta* (n_{total}=18, $n_{ancient}$=9), *Draba* (n_{total}=47,

$n_{ancient}$=4), and *Ranunculus* (n_{total}=47, $n_{ancient}$=10). From these extracts, 119 sequencing libraries were ultimately produced, 11 of which were used for enrichment experiments. Sequencing was perfomed in 11 sequencing runs, 4 of which used the Illumina HiSeq platform, and 7 the Illumina MiSeq. Detailed information is provided in DS:Data and in the remainder of this section. The laboratory workflow is summarised in figure 4.1.

Part of the workflow involves PCR and Sanger sequencing assays performed on the three study taxa. In early experiments, PCR bands from several other taxa (*Spermophillous parryii*, *Potentilla sp.*, *Erysimum sp.*, *Anemone sp.*) were Sanger sequenced in the course of assessing different species' viability for further study. For completeness, the outcome from these assays is included in the results section (TBL. 4.2; TBL. 4.3).

The sequence data was first used to investigate the quality and quantity of aDNA in these samples by mapping them to appropriate reference genomes. This allowed us to characterise the levels of DNA degradation, estimate the amount of useful genetic information in the samples, and to explore using read length distributions the possible reasons that different samples and treatments yield the results they do. The data produced was then used for phylogenetic analysis, as described in chapter 5.

4.2.2 Molecular Methods

4.2.2.1 DNA Extraction

DNA extraction was performed in the Australian Centre for Ancient DNA (ACAD) ultraclean laboratory facilities, observing all accepted standards for aDNA work [67]. The extraction protocol was based upon an ancient plant DNA extraction method recommended by Kistler [36]. The volumes were altered to reflect the small size of the samples. Beta-mercaptoethanol was not used being considered unnecessary safety risk in the ultraclean laboratory. The recovery by ethanol precipitation was replaced with a silica solution recovery, which is standard practice in aDNA studies owing to its relatively consistent high yield compared with other recovery methods [60].

Selected fruits were rinsed twice in 100% ethanol in a 1.5 ml Eppendorf tube using a 1 ml pipette. After removing the ethanol, the tubes were secured open inside a UV oven, which dried the samples for 10 m. Two methods of homogenisation were used in an attempt to release any surviving aDNA and thus improve the yield. The unusually small size of the samples made this task a challenge, and several attempted methods in fact failed to yield enough homogenate to be considered worth continuing the extraction, or which never yielded a PCR band or useable sequencing library. Among these failed methods was an attempt to homogenise the fruits in a mortar and pestle, either at room temperature, or cooled by -20°C freezer, laboratory freezing spray (dimethyl ether), or liquid nitrogen (LN_2). The following methods (see figure 4.1) produced extractable homogenate:

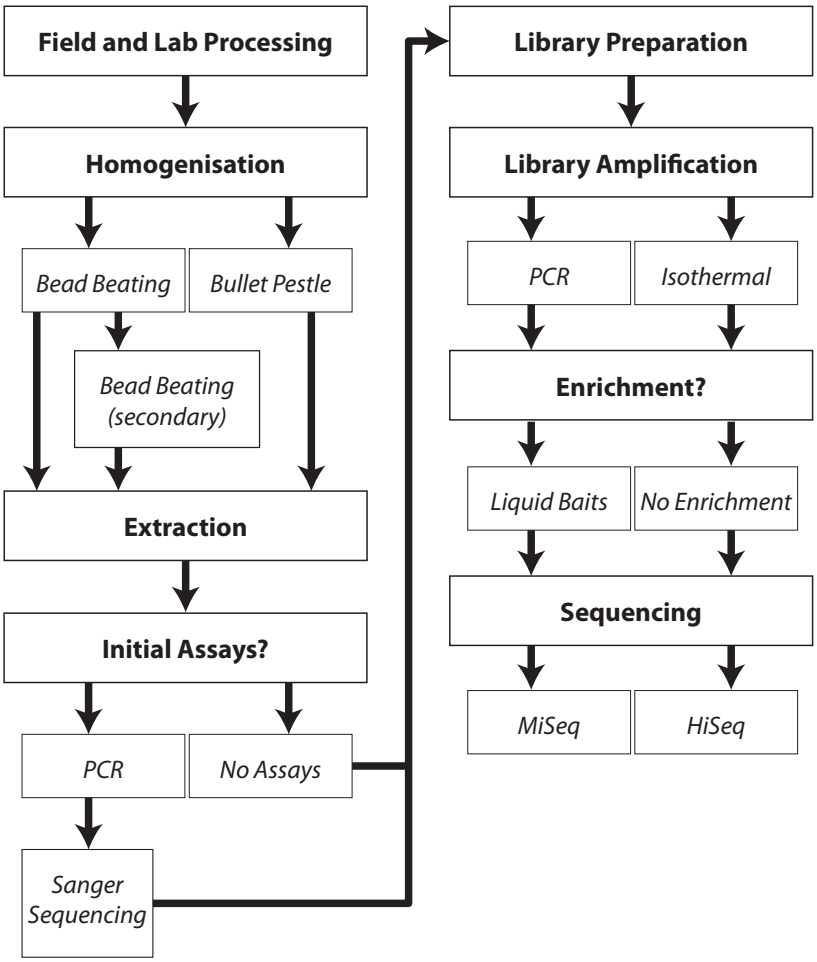

FIGURE 4.1: Laboratory workflow, with steps in bold, and variations on these steps shown underneath each in italic. See section 4.2 for details.

1. **Bullet Pestle.** The fruits were imbibed in a 2% CTAB extraction buffer (CTAB buffer hereafter; 2% w/v hexadecyltrimethylammonium bromide (CTAB), 1% w/v polyvinylpyrrolidone (PVP), 1.4 M NaCl, 20 mM EDTA, 100 mM Tris-HCl pH 8) for approximately one hour, to soften them. A sterile bullet pestle (Eppendorf Europe, manufacturer no. 0030 120.973) was used to grind the sample.

2. **Bead beating.** Homogenisation and lysis can be made more effective by freezing the sample with liquid nitrogen (LN_2), which causes cell walls and tissues to shatter under mechanical force. LN_2 was tested for contamination by adding a small quantity to control tubes in library preparation and allowing the LN_2 to evaporate. After amplification, these controls failed to produce a visualisable sequencing library in all tested cases, suggesting very low levels of DNA contamination. Fruits were added without buffer to tubes taken from the PowerPlant Pro DNA Isolation Kit (MoBio, catalog no. 13400), which contain stainless steel beads and are designed to be shaken by a laboratory tissue homogeniser. The tubes were immersed in LN_2 until boiling slowed, and then beaten using a FastPrep 120 (Thermo Savant, product code FP120) for 30 s

with speed set to 4.5 m/s. Following beating, the tube tops were bleached and tightened before further handling, in anticipation of the possibility that some exchange of air through the lid seal could occur under thermal expansion/contraction. Following beating the samples were completely powdered, with only very few individual fragments visible.

3. **Second-round bead beating.** It is often suggested (especially in studies on ancient bone) that the portion of the sample that resists digestion after the homogenisation and incubation steps may, owing to its toughness, be better suited to preserve aDNA than the more easily digestible portion of the sample [25, 38, 61]. To investigate whether the same effect could be possible in plant samples, several samples were subjected to an analogous treatment: Pellets left over after the treatment described above (which still contained beating beads) were subjected to the same freezing and beating a second time before CTAB buffer was added again.

The homogenate from each method was made up to 200 µL with CTAB extraction buffer and incubated for 1–2 hr at 37°C on a rotor. The debris was pelleted by centrifugation at 12,000 rpm for 120 s. The supernatant was removed and the pellet archived. The DNA was extracted using 25:24:1 phenol:chloroform:isoamyl alcohol, which was added in equal volume to the supernatant, and then intermittently vortexed over a period of 10 min. The phases were separated by centrifugation at 12,000 rpm for 1 min. The aqueous phase was removed and added to 1.2 mL of a guanidinium binding buffer (93% v/v Buffer QG [Qiagen, catalog no. 19063], .5% v/v 5 M NaCl, 1.22% v/v Triton-X 100 [Sigma Aldrich, product no. 234729), and 5.6% v/v NaOAc], which also contained 15 µL silica suspension [12% w/v silica in H_2O]. The DNA was allowed to bind the silica for 1 hr on a rotor at 37°C. The silica was pelleted by centrifugation at 13,000 rpm for 1.5 min, and washed by resuspension in 900 µL of 80% v/v ethanol. After vortexing, the silica was pelleted again (centrifugation at 13,000 rpm for 1.5 min), and the ethanol removed by pipette tip. Remaining ethanol was allowed to evaporate by placing the tube in a heat block at 37°C, with a sterile laboratory tissue covering the open top. The DNA was released by resuspending the pellet in 200 µL of an aqueous elution buffer (10 mM Tris-HCl ph 8, 0.05% v/v Tween-20 (Sigma Aldrich, product no. P9416), 1 mM EDTA), and incubating at 50°C for 10 min with intermittent vortexing. After pelleting the silica by centrifugation at 13,000 rpm for 2 min, the supernatant containing the DNA was removed by pipette and stored at ~-20°C.

Herbarium samples were prepared identically to the permafrost samples in a standard DNA extraction laboratory, with the exception that fruits were substituted for ~200–500 mg of foliar and stem material.

Alongside ancient fruits, several samples of Arctic ground squirrel were retrieved from the nests. DNA was extracted from two to three cheek teeth from each of three skulls using a standard aDNA protocol reported by Rohland and Hofreiter (2007) [59]. The teeth were reduced to fragments averaging ~0.2–0.5 mm in diameter using a hammer before being digested and extracted.

4.2.2.2 PCR and Sanger sequencing assays

To assay for the presence of endogenous plant aDNA, a 158 nt fragment of the chloroplast gene rbcL was PCR amplified using angiosperm-specific primers (laboratory ID A179: CGTCCTTTGTAACGATCAAG; laboratory ID A181: GAGGAGTTACTCGGAATGCTGCC). The squirrel samples were assayed with rodent-specific primers targeting a ~130 nt fragment of the mitochondrial 12S rRNA gene (laboratory ID 835: GAAACCCCTAATGACAAACA; laboratory ID 836: AGAGAGCCAAAGTTTCATCA). PCR assays were conducted in 25 µL reactions with 2 µL of undiluted DNA extract and 21 µL reaction mixture (0.4 mM each forward and reverse primers, 0.25 mM each mixed dNTPs, 3 mM MgSO4, 1X HiFi PCR Buffer, 2 mg/ml Rabbit Serum Albumen (RSA), 1.25 U/µL HiFi Polymerase). The polymerase was activated with a two-minute incubation at 94°C, followed by 55 cycles of denaturation (30 s at 94°C), annealing (30 s at 58°C), and extension (30 s at 68°C). The final extension was performed at 68°C for 10 min, and the products were visualised under UV light after electrophoretic separation (typically run for ~1 hr at 100 v on a 3% agarose gel, subsequently immersed in a 3X Gel Red solution for ~20 min, followed by washing in water for 10–20 min). PCR bands (TBL. 4.2) produced for several taxa putatively identified to genus by comparisons to photographs in the literature [78–82] were Sanger sequenced [4] to help verify their authenticity. After manually checking and editing the sequences using the chromatogram editing tools in Geneious 7 [35], a BLAST search for the sequence was performed against Genbank's nucleotide database [12] using default parameters.

4.2.2.3 Library Construction

Illumina sequencing libraries were constructed using a modification of the protocol outlined by Meyer and Kirscher (2010) [50], beginning with 20 µL of undiluted DNA extract as a template. Herbarium samples were sheared using a focussed ultrasonicator (Covaris, model no. S220), following the manufacturer's recommended settings for a mean fragment size of 180 nt (lib_mm, lib_ms) or 600 nt (lib_lon). Each repair reaction (4 µL 10X NEB2 Buffer, 0.3 µL 25 mM each mixed dNTPs, 4 µL 10 mM ATP, 0.8 µL 10 mg/ml RSA, 7.4 µL ultrapure H_2O, 2 µL 10 U/µL T4 Polynucleotide Kinase (PNK), 1.5 µL 3U /µL T4 DNA Polymerase) was incubated at 25°C for 30 m. The repaired DNA was purified on silica columns using the MinElute Enzymatic Reaction Kit (catalog no. 28204): Each repair reaction was mixed with 300 µL ERC buffer and spun through the column using a centrifuge for 60 s at 13,000 rpm. The bound DNA was washed with 700 µL of PE buffer, using the same centrifuge settings. Excess EB buffer was removed from around the edges of the column membrane using a 10 µL pipette. The membrane was soaked in 22.5 µL EB elution buffer at 37°C for ~10 m, then the DNA was eluted in the buffer by centrifugation at maximum speed (16,000 rpm) for 60 s. Adapters were ligated to the repaired, purified samples by incubation in ligation reaction (22 µL DNA from previous step, 1 µL each 12 mM P5 and P7 truncated adapter DNA with barcodes specific to each sample, 4 µL 10X T4 Ligase Buffer, 4 µL 50 Polyethylene Glycol

(PEG) solution, 1 µL 10X T4 DNA Ligase, 9 µL ultrapure H_2O) for 60 m at 22°C. The truncated adapters used were in standard Illumina format with 5, 6, or 7 bp barcodes at one or both ends (P5 truncated: ACACTCTTTCCCTACACGACGCTCTTCCGATCT[barcode], P7 truncated: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT[barcode]). The barcodes for each samples are given in DS:Data. The samples were again purified on silica columns as described above. Nick repair was performed by incubating the nick repair reaction (20 µL adapter-ligated DNA from previous step, 4 µL 10X Thermopol Buffer, 0.3 µL 25 mM each mixed dNTPs, 1.5 µL 8 U/µL Bst DNA Polymerase, 14.2 µL ultrapure H_2O) for 60 m at 37°C. The Bst Polymerase was then heat-deactivated by incubation at 80°C for 10 m. Libraries were amplified by two subsequent PCR reactions, with one exception (lib_fw; see below). In both rounds of amplification, each sample was amplified in eight separate reactions as a means to reduce clonality. The first set of reactions (2 µL library DNA, 1.25 µL forward primer ACACTCTTTCCCTACACGAC, 1.25 µL reverse primer GTGACTGGAGTTCAGACGTGT, 2.5 µL 10X Amplitaq Gold Buffer, 2.5 µL 25 mM MgCl2, 0.625 µL 10 mM each mixed dNTPs, 0.25 µL 5 U/µL Amplitaq Gold, 14.625 µL ultrapure H_2O; Enzyme activation 94°C for 6 m, 13 cycles of melting—94°C for 30 s, annealing—60°C for 30 s, and extending—72°C for 40 s, final extension at 72°C for 10 m) serves mainly to amplify the sample. The second round provides minimal further amplification but uses fusion primers to add full-length adapters to the fragments, which include a 7bp index in the P7 adapter (P5 full: AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT, P7 full: CAAGCAGAAGACGGCATACGAGAT[index]GTGACTGGAGTTCAGACGTGT). The second amplification was also performed in octuplicate on each sample, with identical chemistry with the exception of the primers, and for 7 PCR cycles. After each round of amplification, the octuplicate reactions for each sample are pooled and the reaction purified and concentrated using magnetic beads using the Agencourt AMPure XP system (Beckman Coulter, part no. A32782), according to the manufacturer's directions, using a 1.8:1 ratio of AMPure bead solution to sample. This ratio affects the lower size cutoff of the purified sample, and is important in removing dimer products that often occur in sequencing libraries made from low-concentration DNA. The purified DNA was eluted in 40µL Tris-HCl pH 8 + 0.05% Tween-20.

Samples in the library preparation batch lib_fw (see DS:Data) underwent a single round of recombinase-polymerase amplification using the TwistAmp Basic Kit (TwistDX, product code TABAS03KIT), according to the manufacturer's protocol, beginning with 12.5µL starting DNA. The reaction was run for 25 m and stopped by the addition of 10 µL 0.5 M EDTA.

Libraries were visualised and quantified using the TapeStation 2200 (Agilent, product no. G2964AA), according to the manufacturer's protocol. Prior to sequencing, libraries were pooled, and the pool was again visualised on the TapeStation allowing estimation of the average fragment size. The pool was then more accurately quantified using the KAPA qPCR Quantification Kit for Illumina Platforms (KAPA Biosciences, product code KR0405v6.14), according to the manufacturer's instructions, and diluted to 2nM.

4.2.2.4 Enrichment

Selected libraries were enriched using the MYcroarray MYbaits Custom Target Capture Kit, with 20,000 custom baits to target the chloroplast genomes of the three study genera based upon the published chloroplast genomes of three closely-related taxa, *Ranunculus macranthus* (NCBI Accession: NC_008796.1), *Draba nemorosa* (NC_009272.1) and *Fagopyrum esculentum* (NC_010776.1). With 120 nt baits, a tiling density of five was achieved, meaning each nucleotide position in the three reference sequences was included in five bait sequences. Short adapter ligated libraries were pooled and concentrated to provide ~100 ng starting DNA. The protocol followed the manufacturer's directions in version two of the manual (legacy version available at http://www.mycroarray.com/pdf/MYbaits-manual-v2.pdf). The extended hybridization time of 36 h recommended for aDNA by the manufacturers was used, beginning at 65°C and decrementing the temperature by 2°C every four hours. In the final washing steps, three washes were used, each including a 5 m incubation in Wash Buffer 2 on a rotor at 55°C. Two subsequent library amplifications, visualisation, and quantification, were performed as described at the end of the library construction method (section 4.2.2.3).

4.2.2.5 Sequencing

Five libraries were sequenced on the Illumina MiSeq platform (150 cycles, paired end) at the Australian Genome Research Facility (AGRF) in South Australia [4], two libraries were sequenced on an Illumina HiSeq (125 cycles paired end) at the Queensland Biology Institute (QBI) [5], and two libraries were sequenced on an Illumina HiSeq (50 cycles single end) at the Australian Cancer Research Foundation's (ACRF) Cancer Genomics Facility in South Australia [6]. All sequencing was funded by ACAD. Sequencing statistics and other information for each library are given in the DS:Data.

4.2.3 Computational Methods

4.2.3.1 Read Processing

Reads were demultiplexed based upon their 3′ and 5′ barcodes using the script MTRW_-demultiplex.pl (DS:Code:2.7), which tolerates a given number of mismatches in each barcode—so long as there is little ambiguity about the most likely assignment—and which also allows for barcodes of differing lengths, even on either end of the same fragment. Three sets of reads were produced for each library, each suitable for different investigations (see section 4.2.3.2):

- Single-End Truncated (henceforth SET) reads were produced for investigations where the outcome might be biased by read size and quality score patterns (see section 4.2.3.4), which are largely different between merged and single-end reads because the merged sections have higher quality scores owing to the confirmation of one sequence by the

other. Truncated reads were created by taking all reads (the first read of each pair in the case of paired-end libraries) with length >= 50 and truncating them to 50 nt if longer.

- Paired-End Untrimmed (PEU) sequencing reads were used when investigation of the original DNA fragment size was important, or when an investigation depended upon positional information with respect to the ends of DNA fragments. PEU reads were derived from all paired-end libraries. Reads were merged and adapter sequences trimmed using AdapterRemoval v2 [64] with default settings (which does not trim poor-quality bases from read ends). All merged reads of length 20 nt or over were retained.
- Quality Trimmed (QT) reads were used when sequence coverage and reliability were of utmost importance. To create QT reads, PEU and raw single-end reads were quality-trimmed at both ends using BBduk [1], which also more stringently removes putative adapter contamination from the reads, using the *qtrim=lr* option and the parameters *trimq=25* and *minavgquality = 25*.

4.2.3.2 Mapping to Reference Sequences

Wherever mapping was performed, the short read aligner BWA [41], the sorting tool in the SamTools package [44], and the duplicate removal tool MarkDuplicates in the Picard Tools package [7] were used. The mapping steps were performed using commands shown in the DS:Code:1.4.1, in which BWA's mapping parameters are set to standard parameters for aDNA (*-l 1024 -n 0.01 -o 2*).

To create draft chloroplast reference genomes appropriate to the taxa under study, the program MITObim [29] was used. This program is designed to help reconstruct new plastid sequences when the reference genome is highly diverged from the sequenced genome. Reads are first mapped to the divergent guide genome to identify homologous portions of this genome in the available reads. These portions—contigs—are then extended by astringently searching the available reads for those that contain sequences similar to the ends of the homologous regions, allowing the contig to be extended. In the final consensus sequence, these regions are often joined by a length of 'N' (no call) nucleotides, meaning the resultant genome includes many gaps, hence "draft". A "generic consensus" chloroplast was generated for each taxon under study, using the pooled reads from all samples of each. The initial reference genomes used were identical to those used for bait design (see section 4.1.1.4), but with the large inverted repeats that occur in many chloroplast genomes [68] removed manually with the use of dotplots and sequence manipulation tools in Geneious 7. The MITObim commands are given in DS:Code:1.4.2.

Several subsequent mapping-based analyses were performed, and the resultant summaries (section 4.2.3.4) are combined and contrasted in the results (section 4.3):

- To investigate the approximate relative amounts of nuclear DNA in each sample (see sections 4.3.2.3 and 4.3.3), nuclear Coding DNA Sequences (CDSs) from genome assemblies of closely-related taxa were used. *Draba* sequences were mapped to *Arabidopsis thaliana* (both family Brassicaceae), TAIR10 release [39]. *Bistorta* sequences were mapped to *Beta vulgaris* (both order Caryophyllales), RefBeet1.2 assembly [22]. *Ranunculus* sequences were mapped to *Aquilegia caerulea* (both family Ranunculaceae), annotation 1.1 [8]. To investigate the approximate relative amounts of Repetitive Element (RE) DNA in each sample, sequences from all study genera were mapped to the PGSB Repeat Database PGSB-REdat_v9.3p [9].
- To characterise degradation patterns in the DNA (see section 4.3.2.1), PEU reads were mapped to the generic consensus chloroplasts for each study genera.
- To investigate factors that influence the success of botanical aDNA sequencing (see sections 4.3.2.1, 4.3.2, 4.3.2.2, and 4.3.3), the PEU reads were mapped to a combined reference containing the generic chloroplast and CDSs for the genus, and the RE sequence database. The combination of the sources was intended to eliminate sequences that map to similar sequences in more than one source. Sequences mapping closely to more than one source are given a low mapping quality score, and a cutoff of $maq \geq 10$ was imposed to help eliminate such confounding.
- To investigate how the relative amounts of DNA from each source varied between samples (see sections 4.3.2.3 and 4.3.3), the SET reads were mapped to the same combined database, with the same quality cutoff imposed. SET reads were also used where any comparison was made that included the herbarium samples sequenced in fifty-cycle, single-end HiSeq runs. Truncation was deemed important based on the observation that shorter reads had a much greater tendency to map to the RE database. This is likely because this database contains many short sequences representing the single repeat units of a tandem repeat sequence: A read deriving from a tandem repeat region will therefore map only if a significant portion of the fragment includes the repeat unit; A read that overlaps the end of the unit may not map, and this is much more likely for long reads. Homogenising the read length helps avoid such sources of bias.
- For phylogenetic reconstruction and assessment of the useful data available in libraries (see coverage potential score below), QT reads were mapped to the chloroplast references (see chapter 5).

4.2.3.3 Damage Characterisation

DNA degradation patterns were characterised in the PEU reads using MapDamage 2.0 [34], using default parameters (DS:Code:1.4.4). Many libraries did not generate enough mappable reads to produce a reliable MapDamage profile. As a compromise, reads from each genus were pooled into a permafrost or herbarium group to produce an aggregated MapDamage profile for each type of preservation.

4.2.3.4 Novel Library Summary Statistics: Motivation and Calculation

Reads, mapped reads, and quality score strings were all parsed using MTRW_strings_summary.pl (DS:Code:2.8), a generalised script that summarises the desired properties of many strings, whether they be nucleotides or quality scores. The script was used to output the total number of reads, mean read length, the mean read quality, the mean read G/C content, the mean base quality and G/C content at each position relative to the beginning of the read, a frequency histogram of read lengths for both shotgun reads and reads that mapped to various references (CDSs, REs, and generic chloroplast consensuses), and the Shannon entropy of these length distributions.

For PEU reads, these histograms were normalised (divided by the total number of reads) to give an approximation of the read length distribution.

Using the combined output from these two scripts, the data were manipulated in R and Excel to produce the following measurements:

- **The relative endogenous content estimate:** This was taken to be the proportion of SET reads mapping to the generic chloroplast reference. This was chosen for two reasons. Firstly, the chloroplast reference sequences were thought a better way to compare between genera, since less bias was introduced by differences in the sizes or completeness of other possible references, the sequence divergence between the references and the study taxa, differences in ploidy between the genera and their members, or differential degradation rates of different types of DNA. The use of truncated reads was considered necessary based upon informal investigations that showed that short or truncated reads would map to a reference genome more often than longer or full-length reads, and that making the length homogeneous helped greatly in reducing this bias when comparing libraries. Undeduplicated read counts were used when calculating this proportion: Assuming that the duplication rate is similar in mapped and unmapped sequences, this ought to provide a more accurate estimate of endogenous DNA content than deduplicated read counts, because it avoids confounding between the dual processes of PCR duplication and endogenous DNA degradation.
- **The duplication factor:** A measurement of the degree of clonality seen in the sequences owing to PCR duplicates being sequenced, and resulting in identical sequences.

Since putative PCR duplicate fragments are identified by the coincident mapping positions of the read ends, the mapped PEU reads were used to generate this figure. The statistic gives the proportion of mapped reads that are removed as putative duplicates, that is, $1 - (\frac{\text{number of mapped reads with duplicates removed}}{\text{total number of mapped reads}})$.

- **The coverage potential score:** This score is an attempt to summarise what a shotgun aDNA study may attempt to optimise. While a high proportion of mapped reads, or enough quality DNA to produce PCR bands reliably, or a low duplication rate, may all be indicators of improvement, the ultimate goal will usually be to reconstruct the greatest possible amount of ancient sequence. The final sequence yield is the result of multiple interacting factors [26]. Consider for instance the effect of the read length distribution of a library, which is influenced by the degradation of the DNA, and the size-selecting properties of purification steps in the extraction and library preparation methods. Small fragments may be more likely to be derived from the original source. However, they may map less reliably, and each one that maps results in less coverage than a long read. Targeting small reads may also cause more of the library to be discarded (or not sequenced at all), as a greater amount of unsequenceable dimer product is inadvertently included in the library.

  Rather than assuming any particular property is desirable, the coverage potential score reports, based upon the proportion and length distribution of mappable reads, the proportion of a 1,000,000 nt reference sequence that is expected to be covered at least once by 1,000,000 sequencing reads having the same length distribution and unique mapping rate (deduplicated QT reads mapped to generic chloroplast consensus). The numbers here are chosen to give sensible results in aDNA studies (where the unique mappable rate might be around 0–2% and the mean read length between 15 and 100 nt), but can be easily adapted to suit other studies. The probability of a randomly-mapped read on the reference sequence not covering a particular site on the reference is approximately $1 - \frac{\text{length of read}}{\text{length of reference}}$, so the probability of the site not being covered by a unique reads is $(1 - \frac{\text{length of read}}{\text{length of reference}})^a$. If 1,000,000 reads are sequenced, then the number of unique mapped reads of a particular length will be $1,000,000 \times$ deduplicated mapping rate $\times$ proportion of reads with this length. The coverage potential score is therefore one minus the intersection of these probabilities calculated over all the possible lengths of reads:

$$1 - \prod_{l=\text{length of reads}} \left(1 - \frac{l}{1,000,000}\right)^{1{,}000{,}000 \times \text{deduplicated mapping rate} \times \text{proportion of reads with length } l}$$

  The coverage potential score was calculated for the QT reads mapped to the chloroplast generic references, and is used as the basis for assessing the importance of various other characteristics to the aDNA application potential of different factors in figures 4.6 and 4.4. The score was calculated from the read length histograms, with a Perl

one-liner given in DS:Code:1.4.3.

4.3 Results and Discussion

4.3.1 PCR and Sanger Sequencing Assays

		ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	ext	bands	TOTALS	
		tc		*pow*		*nd*		*cmp*		*dar*		*ran*		*rod*		*ste*		*tan*		*tun*		*extracted*	*bands*
PCR BLANK		*1*		*1*				*2*		*1*		*3*		*2*		*2*	*2*	*4*		*1*		**17**	**2**
EXT BLANK		*1*		*1*		*1*		*2*		*1*		*1*		*2*		*1*		*1*		*1*		**12**	**0**
PCR POS		*1*	*1*					*3*	*3*	*1*	*1*	*1*	*1*	*1*	*1*	*1*		*2*	*2*	*2*	*2*	**12**	**11**
Ranunculus	*beadbeat LN2*	*2*	*1*	*2*	*2*	*2*	*1*	*3*	*1*													**9**	**5**
Ranunculus	*bullet*							*3*	*1*	*1*	*1*	*9*	*4*			*3*	*2*	*2*	*2*			**18**	**10**
Ranunculus	*other*																					**0**	**0**
Bistorta	*beadbeat LN2*			*2*	*1*	*2*	*1*	*3*	*2*													**7**	**4**
Bistorta	*bullet*							*3*	*1*	*6*	*6*					*4*	*2*			*3*	*3*	**16**	**12**
Bistorta	*other*																					**0**	**0**
Draba	*beadbeat LN2*	*1*		*1*	*1*	*1*																**3**	**1**
Draba	*bullet*															*2*		*2*	*2*	*2*	*2*	**6**	**4**
Erysimum	*beadbeat LN2*	*6*	*6*							*1*	*1*											**7**	**7**
Potentilla	*beadbeat LN2*	*1*	*1*																			**1**	**1**
Anemone	*beadbeat LN2*	*1*																				**1**	**0**
Spermophillous	*other*													*2*	*2*							**2**	**2**
TOTALS		**14**	**9**	**7**	**4**	**6**	**2**	**19**	**8**	**11**	**9**	**14**	**5**	**7**	**3**	**13**	**6**	**11**	**6**	**9**	**7**	**111**	**59**

TABLE 4.2: Results of PCR assays for plant aDNA, giving the number of assays done in each extraction, and the number that produced PCR bands of any strength in the expected size range. The second row gives the extraction identification codes (see *Conventions* and Table 4.4 for details.)

The PCR assays (TBL. 4.2) and Sanger sequencing results (TBL. 4.3) provide the first evidence for botanical aDNA being preserved in the permafrost-preserved samples. The amplicon targeted in the plant samples is within the rbcL gene, which is highly conserved, making it an ideal target for assaying the presence of plant DNA from a diverse range of species, but less useful for identifying the samples to low taxonomic rankings. Nevertheless, using sequence homolgy searches (TBL. 4.3), assignments could be made with enough accuracy to provide strong evidence that the source of the amplified DNA sequence was likely to be endogenous. The results from positive and negative controls indicate that overall the false positive rate was around 7% (2 positive results / [17 extraction blank controls + 12 PCR blank controls]), and the false negative rate around 8% (1 negative result / 12 PCR positive controls). The aggregate success rate (where "success" is taken as a band in the expected size range as opposed to no band or inconclusive amplification results) is around 59% (47/80), though this does not necessarily guarantee any abundance of aDNA fragments as long as the targeted sequence, since template-jumping of fragmented DNA between PCR cycles can serve to resynthesize complete target sequences from overlapping fragments.

4.3.2 Mapping-Based Analyses

4.3.2.1 DNA Degradation By Age, Genus, and Nest

MapDamage profiles (FIG. 4.2) were used to confirm the presence of DNA degradation signals in DNA that maps to a plant reference. The profiles demonstrate that such signals are only prominent in the permafrost-preserved samples, verifying that the samples contain DNA of ancient origin. The pooling of reads from multiple samples means that the plots

Genus (Morphology-based Assignment)	Sequence (after manual editing)	Notes
Ranunculus	CATTCCGAGTAACTCCTCAACCGGGAGTTCC ACCTGAAGAAGCGGGGGCTGCTGTAGCTGC CGAATCTTCTACAGGTACATGGACAACTGTGT GGACCGATGGACTTACCAGCCTTGATCGTTA CAAAGGACG	100% match (138/138 nt) with > 30 spp. in genus *Ranunculus.*
Bistorta	TCGTCCTTTGTAACGATCAAGGCTGGTAAGTC CATCAGTCCACACAGTTGTCCATGTACCAGTA GAAGATTCGGCAGCTACCGCGGCTCCTGCTT CTTCTGGTGGAACTCCAGGTTGAGGAGTTAC TCGGAATGCTGCC	99% matches (both 137/139 nt) with multiple accessions of *Bistorta vivipara*. A single specimen of *Drosophyllum lusitanicum* (L01907.2) also matches at 138/139 sites, presumably due to misassignment: *Drosophyllum* (order Caryophyllales) are only very distantly related to the Ranunculaceae (order Ranunculales), making convergence unlikely.
Draba	TCGTCCTTTGTAACGATCAAGGCTGGTAAGCC CATCGGTCCACACAGTTGTCCATGTACCAGTA GAAGATTCAGCAGCTACCGCAGCCCCTGCTT CTTCAGGTGGAACTCCGGGTTGAGGAGTTAC TCGGAATGCTGCC	100% match (139/139 nt) with > 30 spp. in family Brassicaceae, including members of genus *Draba*
Erysimum (possibly E. cheiranthoides)	TCATGTACCAGTAGAAGATTCAGCAGCTACCG CAGCCCCTGCTTCTTCAGGTGGAACTCCAGG TTGAGGAGTTACTCGGAATGCTGCC	100% match (87 nt) with > 30 spp. in family Brassicaceae, including members of genus *Erysimum*
Potentilla	CAGTCGATCCATGTACCAGTAGAAGATTCCGC AGCTACCGCTGCTCCTGCTTCCTCGGGCGGA ACTCCAGGTTGAGGAGTTACTCGGAATGCTG CC	99% match (89/90 nt) with > 30 spp. In genus *Potentilla,* and one species in *Wetria* (*W. insignis*; AB267934.1)
Anemone (possibly A. naricissiflora)	GTTGTCCATGTACCTGTAGAAGATTCGGCAGC TACAGCAGCTCCTGCTTCTTCAGGTGGAACT CCAGGTTGAGGAGTTACTCGGAATGCTGCC	99% match (92/93 nt) to >5 species in family Ranunculaceae, in genera *Anemone* (including *Anenome narcissiflora*, Accession KF602167.1), *Caltha,* and *Psychrophyla.*
Spermophillous (S. parryii)	ATGACAAACATCCGCAAAACTCACCCTTTAATT AAAATCGTCAACCACTCCTTTATCGACTTACC TGCACCTTCCAACATTTCTGCATGATGAAACT TTGGCTC	99% match (107/108 nt) to multiple specimens of *Spermophillous parryii plesius* and *S. parryii lyratus*.

TABLE 4.3: Summarised results of Sanger sequencing parts of the rbcL gene (all plants) and the 12S rRNA gene (*S. parryii*) in the various taxa found in the nests.

represent an overview of the damage patterns in each genus, with the samples that produced more mapped reads having more influence on the outcome. However, generating profiles separately on high-yielding libraries (results not shown) confirms that the deamination and depurination profiles summarised in figure 4.2 are reflected in individual members of the pool. Read counts for the samples are given in DS:Data.

DNA survival is expected decrease over time, influenced by temperature, by the presence of water and exposure to sunlight, by the chemistry of the environment, and by repeated freeze-thaw cycles [66, 70]. The findings seem to show, however, that while ancient permafrost samples contain less chloroplast DNA than the herbarium samples, DNA survival is more heavily dependent upon the genus or the particular nest studied (FIG. 4.3). This probably owes to a combination of degradation and displacement by contaminant DNA. Overall, the proportion of endogenous chloroplast DNA is remarkably low in all cases, falling in the realm of log(-8)–log(-7) (around 0.03–0.09%) for permafrost samples, and log(-4)–log(-7) (0.09–1.8%) for herbarium samples.

Since there is little or no time-dependence, the important processes that reduced the endogenous DNA content probably occurred before or after long-term freezing—and before is more likely, given the care taken to keep the samples frozen and uncontaminated until extraction. Since many of the nests are found water-logged, it is likely that any given nest spent several seasons in the soil zone that froze and unfroze each year, being water-logged in summer and frozen each winter. Hydrolytic damage, the cause of both deamination and depurination, is enhanced in liquid water, and the heating of soil moisture by high levels of summertime insolation (see section 1.3.3), while the breakdown of organic soil matter decreased soil pH and provides abundant H+ donors. Repeated freezing and thawing fragments DNA [66]. Seasons spent in the active layer would therefore likely have a more severe effect on DNA preservation than many subsequent years spent continuously frozen, and is the best explaination as to why endogenous DNA content depends more on the specific nest used than on the absolute age of the nest, as clearly shown in figure 4.3. This correlation suggests future studies may be well advised to begin with a broad screening of as many nests as possible, followed by focussing effort into those that give the most promising initial results.

It is apparent from figure 4.2 that *Bistorta* produced more heavily deaminated reads than *Draba* and *Ranunculus*, and that *Ranunculus* incurred lower rates of depurination-mediated fragmentation. This could be explained by differences in age and sequence representation between the samples in each genus, but this fails to explain why different levels of deamination are seen between *Bistorta* and *Draba* despite similar depurination levels, or why depurination affects modern *Draba* and *Ranunculus*, but not *Bistorta*. These facts suggest preservation properties differ between genera. This may owe to the different sample types: *Bistorta* an asexual bulbil, *Draba* a silicle comprising two leaf-like valves and a full or partial complement of seeds (usually 5–10), and *Ranunculus* a tough achene (FIG. 3.3). While most of the fruits from all genera were notably brittle and often bearing signs of physical damage, the *Bistorta* bulbils were notably so, often appearing distorted by drying, or hollow with the inside scales having decayed leaving only the tunic.

The inter-taxon differences in sample homogenisation may also be relevant: *Bistorta* bulbils are extremely brittle and can be ground into very small fragments, almost imperceptible by eye, and this may have aided the release of endogenous DNA in extraction more than it facilitated the invasion of contaminants *in situ*. *Draba* silicles can be similarly ground to a lesser degree, but *Ranunculus* achenes prove very difficult to grind, even after ten minutes

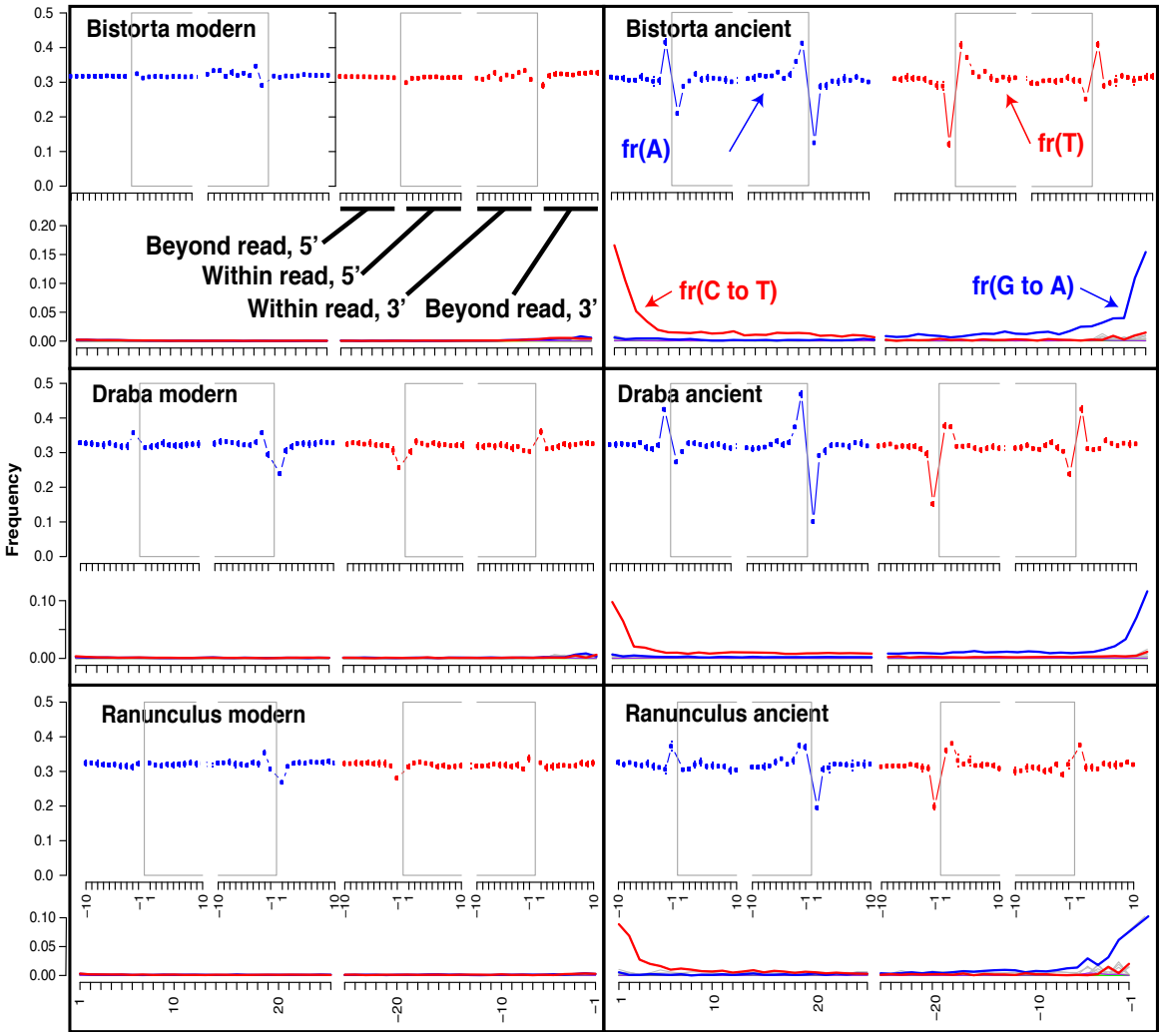

FIGURE 4.2: MapDamage-derived damage profiles of PEU pools for each study genus, for permafrost-preserved and herbarium specimens. The units displayed in the lower two panels apply to all others. The top of each panel represents the frequency of A and T nucleotides within the reads (positions -1 to -10) and in the reference sequence beyond the ends of the reads (positions 1–10; see labels in top panels). In degraded samples, depurination-mediated strand breaks cause an increased frequency of A nucleotides at the first position upstream (3') of the read, and an increased frequency of T nucleotides at the first position downstream. Within the read, frequency deviations are confounded by deamination-induced changes, as well as by imperfections in the adaptor trimming process, which may leave residual adapter sequence, or trim excess insert sequence non-randomly. The bottom of each panel displays the frequency of C-to-T and G-to-A mismatches between a read and the reference sequence, relative to the ends of the read. The same adapter-trimming effects noted above apply here.

of pestle work, which poses an undesirable cross-contamination risk in an aDNA laboratory environment.

Log Coverage Potential Score	Sample Genus	Nest ID	Age (cal. kyr BP)	Ancient/ Modern (Herb-arium)	Homogen-isatio Method	Extraction ID	Library ID	Enrich-ment	Amplifi-catio Method	Tech	Duplication Factor	Log Relative Endo-genous Chloroplast Content	Lower 25th Percentile	Median	Upper 25th Percentile	Entropy	Mean GC Content
-2.036	Ranunculus	TK1127	34597	A	Bullet pestle	tan_6	dar_6	N	PCR	MiSeq	0.49	-4.23	79.33	109.71	166.80	4.67	0.52
-2.030	Ranunculus	TK1127	34597	A	Bullet pestle	tan_6	kw_6	N	PCR	MiSeq	0.35	-4.02	60.49	81.58	119.68	4.24	0.57
-2.013	Ranunculus	TK1127	34597	A	Bullet pestle	dar_6	dar_6	Y	PCR	MiSeq	0.75	-3.81	77.95	107.13	161.87	4.55	0.49
-1.893	Draba	GZ1103	65000	A	Bullet pestle	tan_9	fw_5	N	Isothermal	MiSeq	0.51	-3.79	92.09	125.06	199.49	4.85	0.43
-1.666	Draba	GZ1103	65000	A	Bullet pestle	tan_8	fw_1	N	Isothermal	MiSeq	0.35	-3.51	72.08	104.81	170.70	4.69	0.43
-1.648	Draba	GZ1103	65000	A	Bullet pestle	tan_9	fw_5	Y	Isothermal	MiSeq	0.66	-3.60	96.19	131.60	201.66	4.80	0.41
-1.569	Ranunculus	GZ0907	28458	A	Bullet pestle	tan_5	kw_5	N	PCR	MiSeq	0.10	-3.61	49.47	64.81	95.10	3.96	0.46
-1.495	Ranunculus	GZ0907	28458	A	Bullet pestle	tan_5	fw_4	N	Isothermal	MiSeq	0.38	-3.44	65.95	93.88	150.79	4.54	0.42
-1.398	Draba	GZ1103	65000	A	Bullet pestle	tun_4	kw_8	N	PCR	MiSeq	0.13	-3.28	47.74	62.04	89.16	3.90	0.50
-1.387	Draba	GZ1103	65000	A	Bullet pestle	tan_8	fw_1	Y	Isothermal	MiSeq	0.52	-3.34	85.17	120.66	190.78	4.77	0.40
-1.295	Draba	GZ1103	65000	A	Bullet pestle	tan_8	dar_7	N	PCR	MiSeq	0.07	-3.55	79.47	113.33	182.15	4.75	0.39
-1.282	Ranunculus	GZ0907	28458	A	LN2 beatbeat x2	nd_14	cm_17	N	PCR	MiSeq	0.02	-3.40	121.25	154.97	251.04	4.48	0.53
-1.277	Bistorta	GZ0907	28458	A	Bullet pestle	tun_3	kw_9	N	PCR	MiSeq	0.26	-3.09	59.18	79.23	119.59	4.24	0.48
-1.265	Ranunculus	GZ0907	28458	A	Bullet pestle	tan_5	fw_4	Y	Isothermal	MiSeq	0.69	-3.25	79.50	108.65	173.46	4.63	0.41
-1.212	Ranunculus	GZ0908	27661	A	Bullet pestle	dar_4	fw_2	N	Isothermal	MiSeq	0.73	-2.70	54.76	86.49	158.25	4.55	0.40
-1.149	Bistorta	GZ1103	65000	A	Bullet pestle	tun_2	kw_7	N	PCR	MiSeq	0.27	-2.90	48.88	64.92	95.22	3.96	0.50
-1.010	Bistorta	QC8	65000	A	Bullet pestle	dar_5	dar_1	Y	PCR	MiSeq	0.49	-2.91	71.96	98.16	152.64	4.49	0.38
-0.951	Ranunculus	GZ0908	27661	A	Bullet pestle	dar_4	fw_2	Y	Isothermal	MiSeq	0.91	-2.49	70.25	105.17	175.60	4.65	0.38
-0.918	Bistorta	TK1026	65000	A	Bullet pestle	dar_7	dar_8	N	PCR	MiSeq	0.09	-2.89	76.34	107.71	179.57	4.73	0.42
-0.909	Ranunculus	GZ0907	28458	A	Bullet pestle	tan_5	fw_4	Y	Isothermal	MiSeq	0.91	-2.94	73.11	107.61	178.96	4.75	0.40
-0.834	Draba	GZ1103	65000	A	Bullet pestle	tun_4	kw_8	N	PCR	HiSeq	0.56	-3.40	80.57	109.89	169.91	4.66	0.48
-0.805	Draba	GZ1103	65000	A	Bullet pestle	tun_4	dar_4	N	PCR	MiSeq	0.10	-2.89	92.08	137.80	224.29	4.98	0.41
-0.803	Bistorta	QC5	43580	A	Bullet pestle	dar_6	dar_5	N	PCR	MiSeq	0.05	-2.83	87.28	131.36	216.92	4.93	0.39
-0.795	Bistorta	QC8	65000	A	Bullet pestle	dar_5	dar_1	N	PCR	MiSeq	0.14	-2.81	73.36	101.52	159.71	4.61	0.39
-0.777	Bistorta	TK1026	65000	A	Bullet pestle	dar_8	dar_8	Y	PCR	MiSeq	0.46	-2.65	75.69	104.84	173.20	4.61	0.40
-0.761	Bistorta	GZ1103	65000	A	LN2 beatbeat x2	nd_3	cm_16	N	PCR	MiSeq	0.07	-2.97	118.71	149.15	261.63	4.31	0.50
-0.759	Bistorta	TK1127	34597	A	LN2 beatbeat	pow_7	cm_13	N	PCR	MiSeq	0.03	-2.84	115.86	143.12	220.68	4.33	0.50
-0.682	Bistorta	QC5	43580	A	Bullet pestle	dar_6	dar_5	Y	PCR	MiSeq	0.52	-2.60	78.48	111.76	186.63	4.70	0.38
-0.680	Bistorta		0	M	other	kmga_5	bm_5	N	PCR	HiSeq	0.37	-2.68	92.21	120.06	176.90	4.58	0.43
-0.664	Ranunculus		0	M	Bullet pestle	mm_2	mm_2	N	PCR	MiSeq	0.10	-2.92	90.89	131.50	203.06	4.87	0.42
-0.536	Ranunculus	GZ0908	27661	A	Bullet pestle	dar_4	dar_2	Y	PCR	MiSeq	0.86	-2.15	66.22	90.47	145.90	4.45	0.38
-0.533	Bistorta		0	M	other	kmga_7	bm_7	N	PCR	HiSeq	0.26	-2.50	90.59	116.94	165.79	4.49	0.41
-0.530	Draba	GZ1103	65000	A	Bullet pestle	tun_4	dar_4	Y	PCR	MiSeq	0.53	-2.43	82.12	118.35	194.07	4.77	0.40
-0.502	Ranunculus	GZ0908	27661	A	Bullet pestle	dar_4	dar_2	N	PCR	MiSeq	0.44	-2.51	66.64	98.42	161.99	4.65	0.40
-0.313	Bistorta		0	M	other	kmga_3	bm_3	N	PCR	HiSeq	0.34	-2.21	91.01	118.39	175.87	4.56	0.42
-0.310	Ranunculus		0	M	Bullet pestle	mm_2	lon_2	N	PCR	MiSeq	0.04	-2.34	126.48	194.37	326.56	5.27	0.40
-0.256	Draba		0	M	Bullet pestle	mm_1	mm_1	N	PCR	MiSeq	0.39	-2.21	86.68	129.13	199.80	4.91	0.42
-0.244	Bistorta		0	M	other	kmga_1	bm_1	N	PCR	HiSeq	0.39	-2.13	90.49	118.39	173.98	4.54	0.41
-0.207	Bistorta		0	M	other	kmga_4	bm_4	N	PCR	HiSeq	0.45	-2.07	92.34	122.83	185.00	4.61	0.43
-0.202	Bistorta		0	M	other	kmga_2	bm_2	N	PCR	HiSeq	0.47	-2.06	89.50	117.21	172.90	4.55	0.42
-0.135	Draba		0	M	Bullet pestle	mm_5	mm_5	N	PCR	MiSeq	0.20	-1.95	97.92	135.04	205.28	4.86	0.41
-0.110	Ranunculus		0	M	Bullet pestle	mm_7	lon_7	N	PCR	MiSeq	0.04	-2.15	166.13	256.79	391.26	5.43	0.41
-0.104	Bistorta		0	M	other	kmga_6	bm_6	N	PCR	HiSeq	0.69	-1.88	91.05	119.72	179.04	4.60	0.43
-0.043	Ranunculus		0	M	Bullet pestle	mm_4	mm_4	N	PCR	MiSeq	0.31	-1.99	96.87	135.77	207.81	4.89	0.42
-0.019	Draba		0	M	Bullet pestle	mm_1	lon_1	N	PCR	MiSeq	0.10	-1.68	139.64	229.09	409.46	5.42	0.40
-0.017	Draba		0	M	Bullet pestle	mm_5	lon_5	N	PCR	MiSeq	0.16	-1.59	140.42	215.52	364.01	5.36	0.40
-0.002	Ranunculus		0	M	Bullet pestle	mm_4	lon_4	N	PCR	MiSeq	0.09	-1.54	156.11	234.02	368.86	5.38	0.41

TABLE 4.4: Summarised results of mapping generated using PEU reads mapped to the generic chloroplast consensus references. Only libraries for which *more than 50 PEU reads mapped to the reference* are included. Libraries are ordered by coverage potential score (left column) with selected statistics and factors to compare. The derivations of chosen statistics are given in section 4.2.3.4. Categorical data are arbitrarily coloured for visual distinction. In the *Ancient/Modern* column, permafrost samples are categorised as ancient (A), and herbarium samples as modern (M). The *Enrichment* column reads (Y)es or (N)o depending upon whether the library was enriched (see also figure 4.6). The *Median*, *Upper/Lower 25th Percentile*, and *Entropy* columns refer to the read length distribution, with Shannon Entropy given in nats. The dataset is available in the DS:Data.

4.3.2.2 Homogenisation Method And Ancient DNA Degradation

The success of efforts to increase the relative endogenous DNA yield by altering the homogenisation method were assessed by comparing these two factors in figure (FIG. 4.4). The results are counterintuitive: It was initially reasoned that endogenous DNA was likely to be more prevalent in the insides of the samples, and that completely powdering the samples by LN_2-cooled beat-beating would allow access to a greater amount relative to contaminant DNA from the parts of the sample that were exposed to soil and water.

Furthermore, the parts of the sample that were robust enough to withstand pestle grinding were thought to be more likely to have also withstood invasion from contaminant DNA sources, that is, the same logic that motivates the secondary re-homogenisation of undigested debris. The data, however, suggest that bead beating probably results in a lower average endogenous rate than grinding with a bullet pestle, and only very weakly suggest that a modest improvement is made when re-homegenising undigested material (FIG. 4.4; note that in table 4.4 most of these samples are not shown, having not met the inclusion criteria).

There are several possible explanations. If the bead beating was causing excessive damage to the DNA we might expect more of the longer contaminant reads to be sheared down into the size range of more ancient reads, which may then be preferentially amplified downstream. Excess fragmentation could also feasibly "downshift" both the endogenous and contaminant DNA fragments in average size, causing more small ancient fragments to be discarded in library purification, in the bioinformatic pipeline (at the merging or length filtering steps), or to fail to map to the reference genome. However, fragmentation-based explanations do not explain the possible increase in endogenous DNA content following secondary bead beating, if this increase is not just a stochastic anomaly. The secondary homogenisation results do at least suggest that the first homogenisation does not extract all the available DNA, and future studies may benefit from trying multiple homogenisations using less harsh methods than bead beating.

4.3.2.3 Relative Survival Of Nuclear and Chloroplast DNA

Comparing the number of reads that map to CDS with the number that map to the generic chloroplast reference allowed some assessment of the degradation rates of DNA from these different sources. These measures are heavily confounded by the differences between the reference sequences used for each genus, making comparisons between genera problematic. However, it is noteworthy that an approximately linear relationship is observed in each, and that ancient samples of *Draba* uniquely appear to contain more DNA that maps to the CDSs than to the chloroplast. This is the reverse of an observations made by Allentoft et al. in 2012 [11], who describe an increased rate of nuclear aDNA decay compared with organellar aDNA in moa bones. In plants the nuclear DNA has been held to degrades at a similar rate to the chloroplast [55, 62, 71, 72, 75]. The observations highlighted in figure 4.3 are in

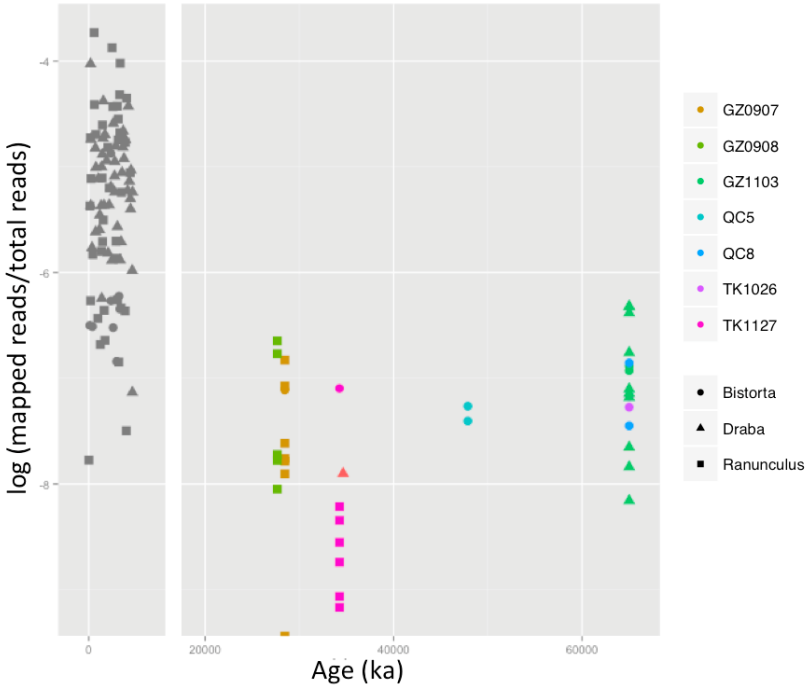

FIGURE 4.3: Approximate changes in relative endogenous DNA content with time, genus, and nest, calculated using SET reads mapped to the generic chloroplast consensus references. Herbarium samples (grey) are all from Age 0 ka, but are jittered randomly for easier visualisation.

agreement with those shown in figure 4.2, suggesting that DNA decay trends are taxon- and genome-dependent, and helping to explain some of the mixed findings in previous studies.

4.3.3 Length Distribution-Based Analyses

Examining the length distributions of mapped and unmapped shotgun reads can help illuminate what portions of shotgun libraries are contributing to coverage of a target genome. The interpretation must be treated with caution: The length distribution of reads may not reflect the true distribution of DNA fragments sizes in the sample, which (at least in the case of degraded DNA) is expected to skew further towards the short end as time progresses. The lower end of the read length distribution is largely controlled by size-selecting processes in the library preparation protocol, especially the magnetic beat purification step that is used to remove dimer products at the expense of some of the shorter fragments in the library. The upper end of the distribution (>~50 bp) is therefore the most informative on the fragments size distribution in ancient samples, but is merely a reflection of the DNA shearing process for modern samples.

Considered in light of other variables, such as those previously discussed in section 4.3.2, the length distributions suggest many interesting—often contradictory—results (FIG. 4.6). Firstly, consider the expectation that mappable fragments will be largely ancient, and hence fall in the shorter end of the length distribution. This expectation is not reliably met: it is

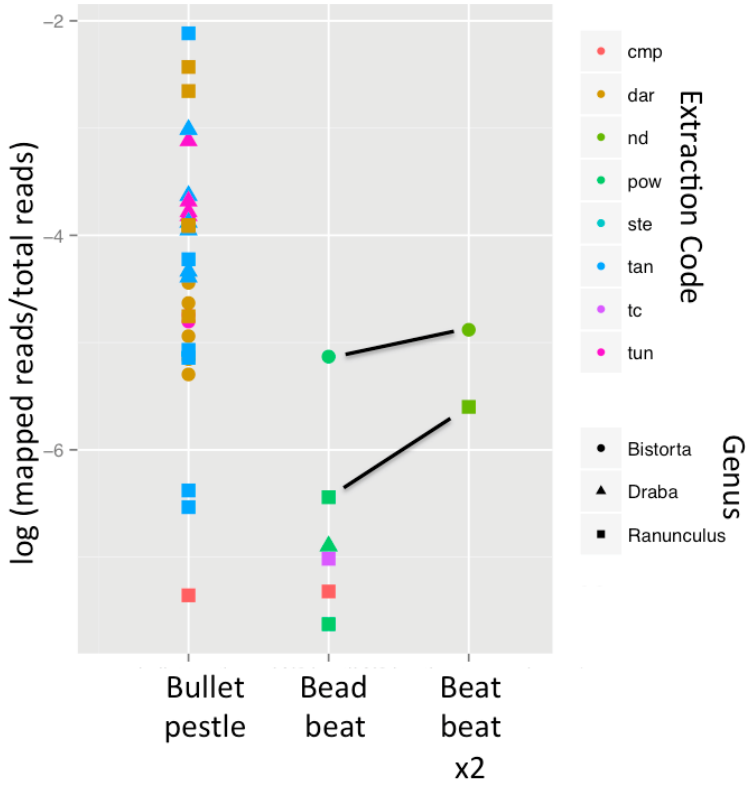

FIGURE 4.4: Comparing the influence of modifications to the extraction method upon the relative approximate endogenous proportion of DNA recovered. Calculated using SET reads mapped to the generic chloroplast consensus references. Samples that were sequenced before and after secondary extraction (see section 4.2.2.1 above) are connected with black lines.

observed for all references in lib_dar_2 and lib_fw_2, and exclusively for the RE and CDS references in lib_dar_4, 5, 7, and 8, and lib_fw_1, 5, and 4. The reverse (longer reads being more likely to map) is seen for the chloroplast reference in lib_dar_1, 4, 5, and 8, and possibly lib_fw_4. Reads mapping to RE and CDS references show this tendency in lib_kw_5, 7 and 9.

Identifying such patterns is made more difficult by 'spikes' in the RE and CDS mapped read distributions, prominent especially in lib_fw samples of both *Draba* and *Ranunculus*. These are probably the result of abundant short reads mapping to short sequences in the reference, possibly representing genes or repetitive elements with a high copy number in the genome, but which occur only once in the reference sequences (as described in section 4.2.3.4). These spikes tend to develop where an abundance of short reads are produced, notably in lib_-fw, which may have been affected by this library's isothermal amplification, while the other libraries were amplified with multiple rounds of parallel PCRs (see section 4.2.2.3). The effort is further impeded by the relatively small number of mapped reads in many cases, making the shape of the distributions difficult to ascertain, though the well-characterised distributions of modern libraries (lib_mm_1 and 5) suggest that when the DNA is of this quality, the reads that map to each source have very similar distributions. The consistent

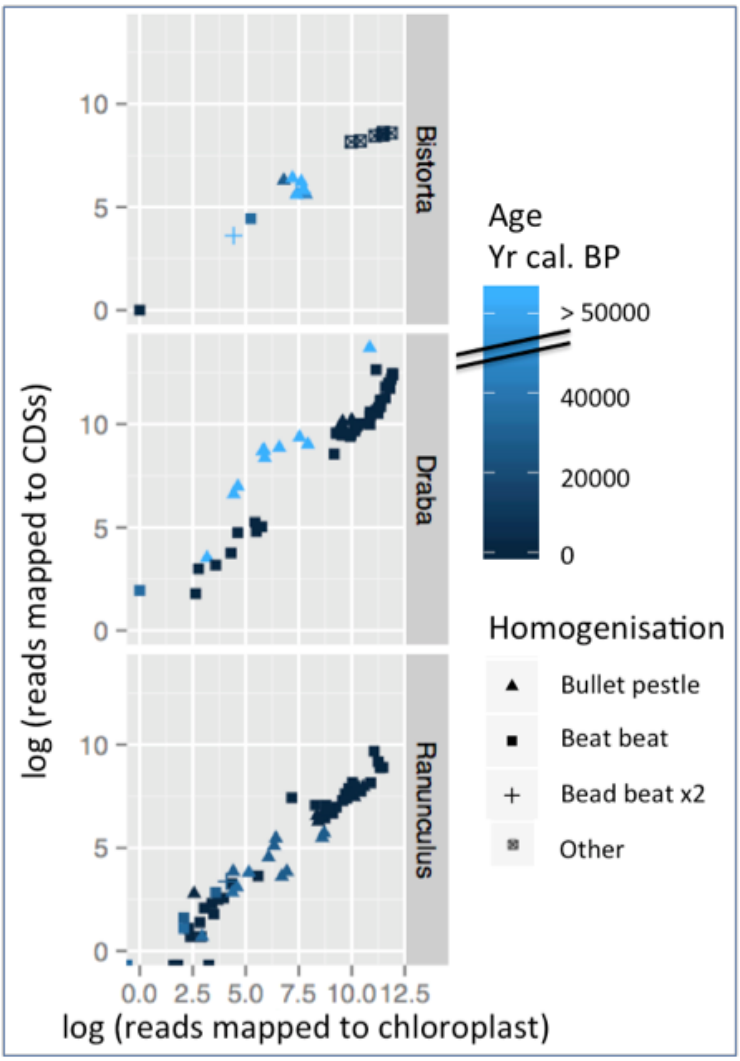

FIGURE 4.5: The relationship between the number of reads mapping to different sources, and the influences of extraction method modification and sample age. Calculated using SET reads mapped to the generic chloroplast consensus references and the nuclear CDS references.

differences between mapped read length distributions and total read length distributions suggests that the latter does not exert a strong effect on the former: the mappable reads in a library have a very specific length distribution *a priori*, and a library that sequences a large number of reads with lengths near the modes of this distribution may benefit in terms of ancient sequence data yield.

Secondly, we might expect that multiple libraries made from the same extract will yield similar results. This is only very weakly suggested by the observations. The anomalously high endogenous chloroplast rate and smaller mapped fragment size are retained across lib_dar_2 and lib_fw_2, which both derive from ext_dar_4, despite the differences in amplification method and aforementioned 'spikes' in the length distribution. In fact the estimation is remarkably similar between them, while the coverage potential is severely reduced in lib_fw. It is worth reiterating here that the coverage potential score is calculated based upon the

length distribution and proportion of mappable unique merged and quality-trimmed (QT) reads, while the relative endogenous chloroplast content is estimated using mapped truncated SET reads, without deduplication. The difference may therefore be a consequence of duplication favouring chloroplast DNA, but this seems unlikely in an unenriched library, despite lib_fw_2's moderately high duplication rate (see table 4.4).

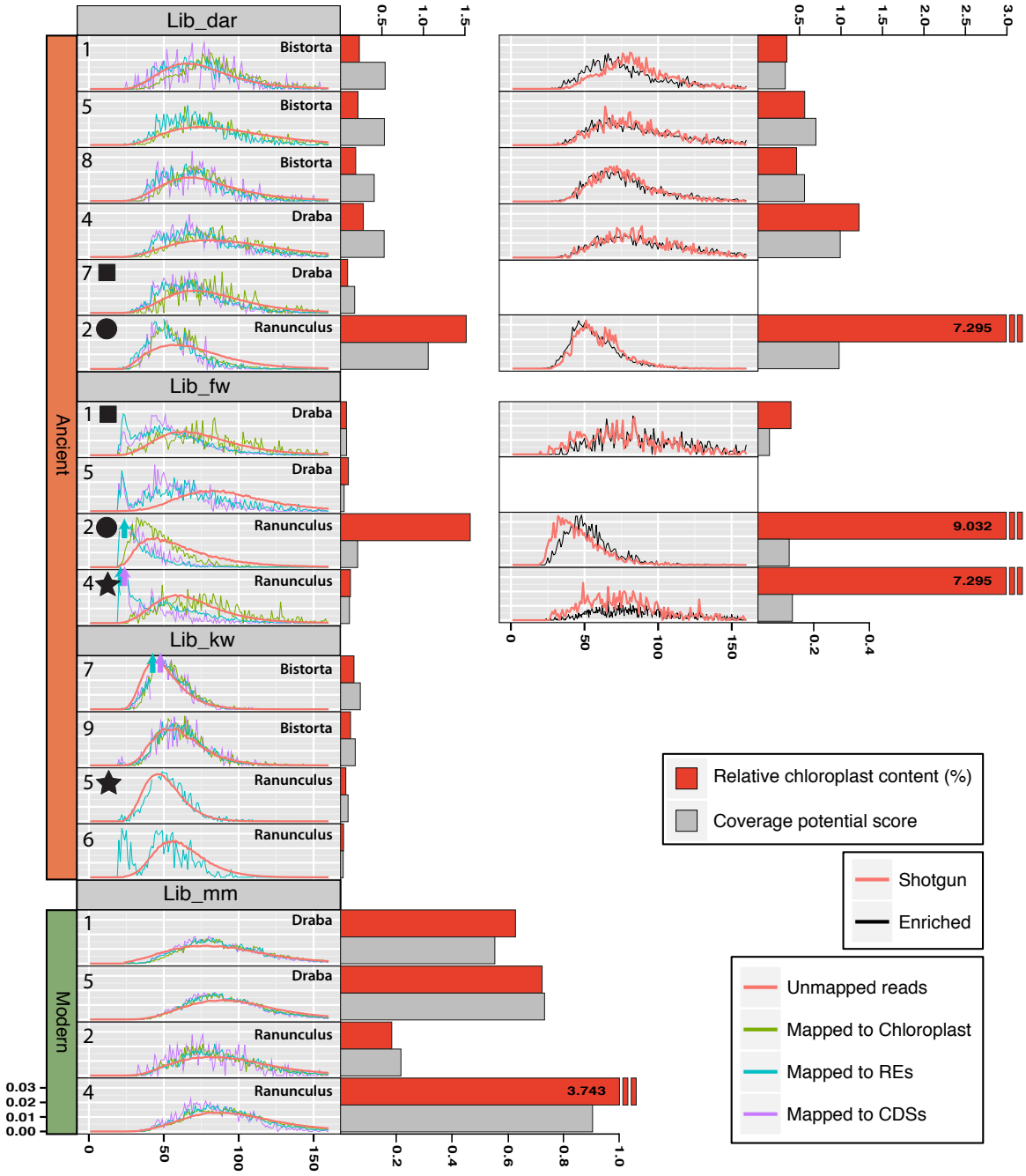

FIGURE 4.6: Investigations involving the length distributions of mapped and unmapped sequences (PEU sequences mapped to all source references). The left column shows the fragment length distributions of shotgun sequenced samples, and of those subsets of the libraries that mapped to different sources, with the library ID numbers and genera given on the left and right of each panel respectively. The right column shows the length distribution of reads mapped to the chloroplast before and after enrichment. Enriched samples are lined up with the shotgun library they were enriched from on the left. All panels are scaled identically, with the density (vertical axis) given at the bottom left of the figure, and length (horizontal axis) given at the bottom of each column of panels. The library identification numbers are given at the top left of each panel, and black shapes (stars, circles, squares) indicate libraries that derive from the same DNA extract (see table 4.4). Other library statistics are given by the length of bars to the right of each panel. The scale for the approximate relative endogenous chloroplast content (a percentage, calculated using SET reads mapped to the generic chloroplast reference sequences) is given at the top of the column, and the scale for the coverage potential score (a proportion, see section 4.2.3.4) is given at the bottom. Truncated bars are labelled with their true values.

More likely the matching endogenous rates are not artifacts, but the coverage potential is limited in lob_fw_2 by the length distribution of mapped reads, the mode of which differs by ~15—20 nt between the two libraries, and which itself is influenced by the purification steps and by the affinity of the amplification reactions for small molecules [58].

Too few reads from lib_kw_5 mapped to the chloroplast and CDS references to characterise a length distribution, with the length distribution of RE-mapped reads difficult to compare owing to the low-length spike in its counterpart library, lib_fw_4. Both derive from ext_tan_5. Lib_dar_7 and lib_fw_1 both derive from ext_tan_8. Both have low endogenous chloroplast rates and coverage potential, and the RE- and CDS-mapped reads in both are on average shorter than those mapped to the chloroplast. This pair of libraries is useful in demonstrating that where libraries have longer average read sizes, and the low-length spike makes up less of the read length distribution, then the comparison of results across libraries improves. Comparing the mapped read length distribution patterns seen in these libraries with those seen in the modern libraries reinforces the source-dependent differential DNA degradation noted for *Draba* specimens in figure 4.5, demonstrating that not only is the nuclear DNA in older samples less abundant, it is also more fragmented.

Thirdly, since enrichment is expected to target endogenous aDNA closely matching the sequence of the baits [40], and since this sequence is similar to the generic chloroplast reference sequences, the process might be expected to alter the length distribution of samples to favour that sequence length range that mapped most commonly to the target genome in the shotgun sample [54]. This expectation is not met: in fact in all cases, the chloroplast-mapped enriched read length distribution seems to tend towards the unmapped shotgun distribution. This is particularly notable in lib_dar_1 and lib_fw_2, where the post-enrichment distributions shift from their pre-enrichment counterparts in opposite directions, to resemble better the shotgun distribution. This result may of course be accidental, arising from stochastic variation, suggesting highly variable results dependent upon either minor variations in treatment in the laboratory protocol (e.g. the order in which samples are processed, variations in reagent concentrations when pipetting very small amounts of a reagent), or upon other sources of variation between samples (e.g. purity, raw concentration, interactions between mixed samples, or the composition of contaminant DNA). However, it is possible that the hybridisation conditions were not stringent enough for the RNA-DNA binding to distinguish strongly between target and non-target DNA [40], and as a result, the baits hybridised with a more random sample of the shotgun reads, which would be expected to have a similar length distribution. If the target length distribution differs from the shotgun distribution, then this sampling would sway the target length distribution towards that of the shotgun distribution, even after washing removed much of the non-target DNA. This possibility highlights the importance of stringency gradient experiments in setting up enrichment protocols for untested samples or bait sets [19]. I note further that where possible, gradient experiments should include in their range, conditions that fall well beyond the levels at which the procedure could be expected to work at either end of the gradient. For an enrichment protocol, this means that one group of replicates should be subjected to

such stringent conditions that no DNA is expected to hybridise, and another replicate group subjected to such astringent conditions that no enrichment ought to occur. Including such samples is invaluable to confirming that the procedure is working as expected, to assessing the range of expected outcomes, and to demonstrating the soundness of the assumptions that underlie the interpretation.

4.4 Conclusions and Recommendations

The work demonstrated conclusively that ancient plant DNA is preserved in the fruits collected by Arctic Ground Squirrels, and that while it is very low in quantity, the quality shows no sign of deteriorating as a function of time even going back to ~50–80 ka. The greatest determinants of endogenous DNA content appear to be the sampled taxon and the idiosyncratic preservation conditions of the nest under study—and *not* the appearance of the samples, the library amplification method, or the sequencing platform. Room temperature homogenisation of the samples may be prefereable to freezing, as well as size selection or amplification techniques that favour the endogenous DNA's size range, without allowing too much contamination by small products including adapter/primer dimers. Enrichment may improve the yield in some cases but has variable results, and careful optimisation will likely be required. In general, future work on similar samples should focus on screening a large number of samples from a broad range of nests to effectively identify those which nests and samples bear promising results, according to some metric such as the coverage potential score introduced in section 4.2.3.4, at the same time characterising the length distribution of desired fragments as demonstrated in section 4.3.3. This introduces the possibility of post-screening size selection of libraries, which was not attempted in the current study. As the next chapter demonstrates, with appropriate screening and optimisation, the prospect of phylogenetic work using multiple whole ancient chloroplasts from individual permafrost-preserved fruits will be challenging, yet is certainly within reach.

Chapter 4 Bibliography

[1] Web Page. URL: https://sourceforge.net/projects/bbmap/files/.

[2] Web Page. URL: http://www.illumina.com/systems/sequencing.html.

[3] Web Page. URL: https://c14.arch.ox.ac.uk/oxcal.html.

[4] Web Page. URL: www.agrf.org.au/services/sanger-sequencing.

[5] Web Page. URL: www.qbi.uq.edu.au/.

[6] Web Page. URL: www.sapathology.sa.gov.au/.

[7] Web Page. URL: http://picard.sourceforge.net.

[8] Web Page. URL: http://genome.jgi.doe.gov/AqucoeGoldsmith_FD/AqucoeGoldsmith_FD.info.html.

[9] Web Page. URL: http://pgsb.helmholtz-muenchen.de/plant/recat/.

[10] Web Page. URL: https://palaeogenomics.wordpress.com/ancient-dna-labs/.

[11] Morten E Allentoft et al. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils". In: *Proceedings of the Royal Society B: Biological Sciences* 279.1748 (2012), pp. 4724–4733. ISSN: 0962-8452.

[12] Stephen F Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410. ISSN: 0022-2836.

[13] Wilhelm J Ansorge. "Next-generation DNA sequencing techniques". In: *New biotechnology* 25.4 (2009), pp. 195–203.

[14] Maria C Avila-Arcos et al. "Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA". In: *Scientific reports* 1 (2011).

[15] Adrian W Briggs et al. "Patterns of damage in genomic DNA sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences* 104.37 (2007), pp. 14616–14621. ISSN: 0027-8424.

[16] Meredith L Carpenter et al. "Pulling out the 1%: Whole-Genome Capture for the Targeted Enrichment of Ancient DNA Sequencing Libraries". In: *The American Journal of Human Genetics* 93.5 (2013), pp. 852–864. ISSN: 0002-9297.

[17] A. Cooper and H.N. Poinar. "Ancient DNA: do it right or not at all". In: *Science* 289.5482 (2000), pp. 1139–1139. ISSN: 0036-8075.

[18] Richard Cronn et al. "Targeted enrichment strategies for next-generation plant biology". In: *American Journal of Botany* 99.2 (2012), pp. 291–311. ISSN: 0002-9122.

[19] Diana I Cruz-Davalos et al. "Experimental conditions improving in-solution target enrichment for ancient DNA". In: *Molecular Ecology Resources* (2016). ISSN: 1755-0998.

[20] J. Dabney and M. Meyer. "Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries". In: *Biotechniques* 52.2 (2012), pp. 87–94. ISSN: 1940-9818. DOI: 10.2144/000113809. URL: http://www.ncbi.nlm.nih.gov/pubmed/22313406.

[21] Jesse Dabney, Matthias Meyer, and Svante Paabo. "Ancient DNA damage". In: *Cold Spring Harbor perspectives in biology* 5.7 (2013), a012567. ISSN: 1943-0264.

[22] Juliane C Dohm et al. "The genome of the recently domesticated crop plant sugar beet (Beta vulgaris)". In: *Nature* 505.7484 (2014), pp. 546–549. ISSN: 0028-0836.

[23] Jacob M Enk et al. "Ancient whole genome enrichment using baits built from modern DNA". In: *Molecular Biology and Evolution* 31.5 (2014), pp. 1292–1294. ISSN: 0737-4038.

[24] Duane G Froese et al. "The Klondike goldfields and Pleistocene environments of Beringia". In: *GSA Today* 19.8 (2009), p. 5. ISSN: 1052-5173.

[25] Cristina Gamba et al. "Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing". In: *Molecular Ecology Resources* 16.2 (2016), pp. 459–469. ISSN: 1755-0998.

[26] Gema Garcia-Garcia et al. "Assessment of the latest NGS enrichment capture methods in clinical context". In: *Scientific reports* 6 (2016).

[27] Felix Gugerli, Laura Parducci, and Remy J Petit. "Ancient plant DNA: review and prospects". In: *New Phytologist* 166.2 (2005), pp. 409–418. ISSN: 1469-8137.

[28] Erika Hagelberg, Michael Hofreiter, and Christine Keyser. "Ancient DNA: the first three decades". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1660 (2015).

[29] Christoph Hahn, Lutz Bachmann, and Bastien Chevreux. "Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach". In: *Nucleic acids research* (2013), gkt371. ISSN: 0305-1048.

[30] Russell Higuchi et al. "DNA sequences from the quagga, an extinct member of the horse family". In: (1984).

[31] Michael Hofreiter et al. "Ancient DNA". In: *Nature Reviews Genetics* 2.5 (2001), pp. 353–359. ISSN: 1471-0056.

[32] Viviane Jaenicke-Despres et al. "Early allelic selection in maize as revealed by ancient DNA". In: *Science* 302.5648 (2003), pp. 1206–1208. ISSN: 0036-8075.

[33] Britta JL Jensen et al. "150,000 years of loess accumulation in central Alaska". In: *Quaternary Science Reviews* 135 (2016), pp. 1–23. ISSN: 0277-3791.

[34] Hakon Jonsson et al. "mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters". In: *Bioinformatics* (2013), btt193. ISSN: 1367-4803.

[35] Matthew Kearse et al. "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data". In: *Bioinformatics* 28.12 (2012), pp. 1647–1649. ISSN: 1367-4803.

[36] Logan Kistler. "Ancient DNA extraction from plants". In: *Ancient DNA: Methods and Protocols* (2012), pp. 71–79. ISSN: 1617795151.

[37] Logan Kistler et al. "Transoceanic drift and the domestication of African bottle gourds in the Americas". In: *Proceedings of the National Academy of Sciences* 111.8 (2014), pp. 2937–2941. ISSN: 0027-8424.

[38] Petra Korlevic et al. "Reducing microbial and human contamination in DNA extractions from ancient bones and teeth". In: *BioTechniques* 59.2 (2015), pp. 87–93. ISSN: 0736-6205.

[39] Philippe Lamesch et al. "The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools". In: *Nucleic acids research* 40.D1 (2012), pp. D1202–D1210. ISSN: 0305-1048.

[40] Chenhong Li et al. "Capturing protein-coding genes across highly divergent species". In: *BioTechniques* 54.6 (2013), pp. 321–326. ISSN: 0736-6205.

[41] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. ISSN: 1367-4803.

[42] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome research* 18.11 (2008), pp. 1851–1858. ISSN: 1088-9051.

[43] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 1367-4803.

[44] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 1367-4803.

[45] Bo Liu et al. "Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences". In: *Genome biology* 12.1 (2011), p. 1. ISSN: 1474-760X.

[46] Bastien Llamas et al. "From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era". In: *STAR: Science and Technology of Archaeological Research* 3.1 (2016), pp. 1–14.

[47] Lira Mamanova et al. "Target-enrichment strategies for next-generation sequencing". In: *Nature methods* 7.2 (2010), pp. 111–118. ISSN: 1548-7091.

[48] Martin Mascher et al. "Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley". In: *Nature Genetics* 48.9 (2016), pp. 1089–1093. ISSN: 1061-4036.

[49] Michael L Metzker. "Sequencing technologies—the next generation". In: *Nature reviews genetics* 11.1 (2010), pp. 31–46.

[50] M. Meyer and M. Kircher. "Illumina sequencing library preparation for highly multiplexed target capture and sequencing". In: *Cold Spring Harbor Protocols* 2010.6 (2010), pdb. prot5448. ISSN: 1940-3402.

[51] Matthias Meyer et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (2012), pp. 222–226. ISSN: 0036-8075.

[52] Ludovic Orlando et al. "Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse". In: *Nature* advance online publication (2013). ISSN: 1476-4687. DOI: 10.1038/nature12323. URL: http://dx.doi.org/10.1038/nature12323.

[53] Ludovic Orlando et al. "True single-molecule DNA sequencing of a pleistocene horse bone". In: *Genome research* 21.10 (2011), pp. 1705–1719. ISSN: 1088-9051.

[54] Johanna LA Paijmans et al. "Impact of enrichment conditions on cross-species capture of fresh and degraded DNA". In: *Molecular ecology resources* 16.1 (2016), pp. 42–55.

[55] Sarah A Palmer, Oliver Smith, and Robin G Allaby. "The blossoming of plant archaeogenetics". In: *Annals of Anatomy-Anatomischer Anzeiger* 194.1 (2012), pp. 146–156. ISSN: 0940-9602.

[56] Harry S Paris. "Overview of the origins and history of the five major cucurbit crops: issues for ancient DNA analysis of archaeological specimens". In: *Vegetation History and Archaeobotany* (2016), pp. 1–10. ISSN: 0939-6314.

[57] Gabriel Renaud et al. "Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA". In: *Genome biology* 16.1 (2015), p. 1. ISSN: 1474-760X.

[58] Stephen Malone Richards. "Investigation and Application of Methods for Ancient DNA Research". Thesis. 2014.

[59] Nadin Rohland and Michael Hofreiter. "Ancient DNA extraction from bones and teeth". In: *Nature protocols* 2.7 (2007), pp. 1756–1762. ISSN: 1754-2189.

[60] Nadin Rohland and Michael Hofreiter. "Comparison and optimization of ancient DNA extraction". In: *Biotechniques* 42.3 (2007), p. 343. ISSN: 0736-6205.

[61] C Sarkissian et al. "Shotgun microbial profiling of fossil remains". In: *Molecular ecology* 23.7 (2014), pp. 1780–1798. ISSN: 1365-294X.

[62] Susanna Sawyer et al. "Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA". In: *PloS one* 7.3 (2012), e34131–e34131. ISSN: 1932-6203.

[63] Angela Schlumbaum, Marrie Tensen, and Viviane Jaenicke-Despres. "Ancient plant DNA in archaeobotany". In: *Vegetation History and Archaeobotany* 17.2 (2008), pp. 233–244. ISSN: 0939-6314.

[64] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". In: *BMC research notes* 9.1 (2016), p. 1. ISSN: 1756-0500.

[65] Stephan C Schuster. "Next-generation sequencing transforms today's biology". In: *Nature* 200.8 (2007), pp. 16–18.

[66] Wen Shao, Sonny Khin, and William C Kopp. "Characterization of effect of repeated freeze and thaw cycles on stability of genomic DNA using pulsed field gel electrophoresis". In: *Biopreservation and biobanking* 10.1 (2012), pp. 4–11. ISSN: 1947-5535.

[67] Beth Shapiro and Micheal Hofreiter. *Ancient DNA: Methods and Protocols*. Humana Press, 2012. ISBN: 161779516X.

[68] Joey Shaw et al. "Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III". In: *American journal of botany* 94.3 (2007), pp. 275–288. ISSN: 0002-9122.

[69] JH Sonstebo et al. "Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate". In: *Molecular Ecology Resources* 10.6 (2010), pp. 1009–1018. ISSN: 1755-0998.

[70] Julien Soubrier et al. "Early cave art and ancient DNA record the origin of European bison". In: *Nature Communications* 7 (2016), p. 13158. ISSN: 2041-1723.

[71] Martijn Staats et al. "DNA damage in plant herbarium tissue". In: *PLoS One* 6.12 (2011), e28448. ISSN: 1932-6203.

[72] Martijn Staats et al. "Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens". In: *PLoS One* 8.7 (2013), e69189. ISSN: 1932-6203.

[73] Gregory W Stull et al. "A Targeted Enrichment Strategy for Massively Parallel Sequencing of Angiosperm Plastid Genomes". In: *Applications in Plant Sciences* 1.2 (2013). ISSN: 2168-0450.

[74] Derek G Turner et al. "Middle to Late Pleistocene ice extents, tephrochronology and paleoenvironments of the White River area, southwest Yukon". In: *Quaternary Science Reviews* 75 (2013), pp. 59–77. ISSN: 0277-3791.

[75] Clemens L Weiss et al. "Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens". In: *bioRxiv* (2015), p. 023135.

[76] Eske Willerslev et al. "Ancient biomolecules from deep ice cores reveal a forested southern Greenland". In: *Science* 317.5834 (2007), pp. 111–114. ISSN: 0036-8075.

[77] Eske Willerslev et al. "Fifty thousand years of Arctic vegetation and megafaunal diet". In: *Nature* 506.7486 (2014), pp. 47–51. ISSN: 0028-0836.

[78] Matthew J Wooller et al. "The detailed palaeoecology of a mid-Wisconsinan interstadial (ca. 32 000 14C a BP) vegetation surface from interior Alaska". In: *Journal of Quaternary Science* 26.7 (2011), pp. 746–756. ISSN: 1099-1417.

[79] G.D. Zazula et al. "Arctic ground squirrels of the mammoth-steppe: paleoecology of Late Pleistocene middens (˜24000–29450 ^{14}C yr BP), Yukon Territory, Canada". In: *Quaternary Science Reviews* 26.7 (2007), pp. 979–1003. ISSN: 0277-3791.

[80] G.D. Zazula et al. "Vegetation buried under Dawson tephra (25,300 ^{14}C years BP) and locally diverse late Pleistocene paleoenvironments of Goldbottom Creek, Yukon, Canada". In: *Palaeogeography, Palaeoclimatology, Palaeoecology* 242.3 (2006), pp. 253–286. ISSN: 0031-0182.

[81] Grant D Zazula et al. "Early Wisconsinan (MIS 4) Arctic ground squirrel middens and a squirrel-eye-view of the mammoth-steppe". In: *Quaternary Science Reviews* 30.17 (2011), pp. 2220–2237. ISSN: 0277-3791.

[82] Grant D Zazula et al. "Paleoecology of Beringian "packrat" middens from central Yukon Territory, Canada". In: *Quaternary Research* 63.2 (2005), pp. 189–198. ISSN: 0033-5894.

[83] Daniel R Zerbino and Ewan Birney. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". In: *Genome research* 18.5 (2008), pp. 821–829. ISSN: 1088-9051.

Chapter 5

Evolutionary Ecology Of The Mammoth Steppe Flora

Phylogenetic Investigations Of Low-Coverage Shotgun Data From Non-Model Arctic Plants

---

*This chapter describes efforts to glean reliable phylogenetic information from poor-yielding shotgun samples of ancient DNA, and from non-model organisms where closely-related reference sequences are unavailable. Chloroplast sequence data for the three study taxa (*Bistorta vivipara, Draba spp., *and* Ranunculus spp.*) was generated as described in chapters 3 and 4, and used to produce phylogenies. This work also produced the two oldest known draft chloroplast genomes to date. Part of the analysis pipeline used in this chapter is presented in chapter 6, in the context of the new software tool developed to accomplish it.*

5.1 Introduction

The modern boreal plant community is thought to have begun its establishment with the gradual cooling of the polar regions beginning at the end of the Miocene epoch, five million years ago. Their evolution has since been fundamentally affected by Pleistocene glacial cycles, with fragmentation and recolonisation acting repeatedly upon the diversity of plants first introduced from the undergrowth of Miocene forests and alpine chains to the south [1, 38].

The modern flora of the Yukon Territory, Canada, is typically Beringian. The region Beringia (see chapter 1) was first named by the Swedish phytogeographer Eric Hultén, whose seminal work *Outline of the History of Arctic and Boreal Biota during the Quarternary Period* [38] compared the distributions of arctic plant species, and concluded that a great northern refugium for plant species during ice ages must have existed on the land-bridge between the Lena River to the west, and the Mackenzie River to the east. His postulation of this and other far

northern refugia contradicted the *tabula rasa* vision of previous naturalists, who envisioned ice sheets advancing fairly uniformly by longitude, and "cleaning the slate" of the northern plant and animal communities [1, 4, 19, 20]. Observing many species' ranges all radiating outwards from a few single far-northern points of high species diversity, Hultén concluded that these centres must have escaped glaciation in order to accumulate and maintain this diversity.

Later geological findings were to vindicate the thinking of Hultén and contemporary biogeographers, revealing that not only did Beringia escape glaciation, but also that thousands of square kilometers of continental shelf north of Siberia were exposed and ice-free during the full glacial [24]. The advent of molecular genetics saw the use of isozymes, Amplified Fragment Length Polymorphisms (AFLPs), and the sequencing of chloroplast and rRNA genes [2, 5–7, 13, 25, 26, 38, 39, 42, 47, 48, 51, 54, 56, 57, 59, 60, 65] to confirm plant survival in refugia, both regional such as Beringia, and perhaps on mountain tops protruding from the ice sheets (Nunataks), or low-altitude coastal refugia. The molecular geneticists also dispelled former skepticism about the viability of long-distance dispersal, including transatlantic crossings by seeds with no apparent adaptations for anemochory (wind dispersal) or hydrochory (water dispersal). Historically, studies on individual species have been difficult to unite into broad generalisations. In 1963, the phytogeographer Rolf Berg [9] stated of the Norwegian Alpine flora:

> *"It is my opinion that no single explanation can account for all the arctic-alpine disjunctions in Scandinavia. A great deal of argumentation has resulted from a futile search for the one universal cause. Each species area should be regarded as a problem per sé. For future advances to be made in this field, more exact descriptive and experimental data must be accumulated, species by species."*

The need for this taxon-by-taxon approach is highlighted by the results of phylogeographic studies that have revealed species groups with both ancient and recent origins, speciation events via both fragmentation and sympatrically via spontaneous genetic barriers, and evidence for strong dispersal barriers in some species being readily overcome by others [1].

Even the genetic methods available during Hultén's life also aided significantly in revealing some of the striking features of the arctic flora. Examination of the chromosomes of the forbs, grasses, sedges, and dwarf shrubs that make up much of the boreal angiosperm community revealed that their notorious taxonomic complexity is reflected at the genetic level. Perhaps owing to the plants' need to accumulate diversity inbetween glacial periods of isolation and stress, the arctic flora are notorious for their reproductive eccentricity, exhibiting variable ploidy, frequent hybridisation, vegetative reproduction, selfing, and agamospermy among other strategies: While species diversity declines with productivity along a south-to-north gradient, the frequency of polyploidy increases, most markedly in regions that were once glaciated, confirming that variable ploidy could be part of a mechanism to generate and maintain genetic/phenotypic variation, and hence a competitive edge in the repeated recolonisation events necessitated by glacial-interclacial cycles [1, 5, 11, 17, 33, 61, 64].

5.1.1 Aims and Challenges

The main goal of the work described here was to recover enough homologous chloroplast sequence from enough modern plant samples to place the ancient plants into some useful context (see chapter 4). We aimed in particular to:

- Demonstrate that extemely sparse and low-quality data could be purged of sources of bias and putative error, so as to produce reliable phylogenetic information.
- To use this information to examine the relationships between members of mammoth steppe species complexes, and to relate the findings to their evolutionary histories and survival strategies.
- To discover the most closely related modern species to the ancient samples, and relate the biology of these species to the environments of the Late Glacial.

Initial attempts to map sequence data to reference sequences, generate consensus sequences, and align the consensus sequences, suggested that the low volume and quality of data would prove the largest challenges. Figure 5.1 illustrates a portion of such an alignment (generated according to the methods outlined in section 5.2.1), to demonstrate the incompleteness of most of the ancient consensus sequences, though several herbarium samples in each group had similarly poor coverage.

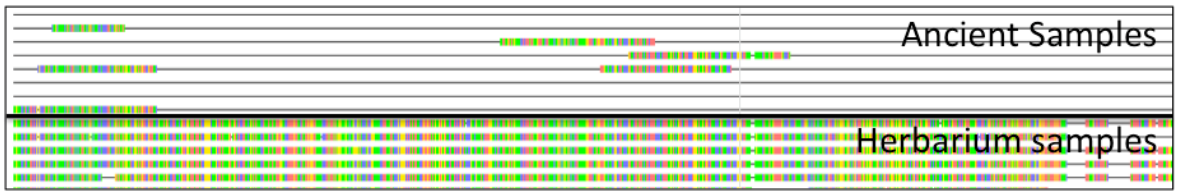

FIGURE 5.1: Aligned consensus sequences for ancient and herbarium samples (taken from the *Ranunculus* alignment before cleaning, positions 63,600–63,620). Horizontal bars represent sequences, with nucleotides coloured T, G, A, C, and N here and in subsequent figures. Horizontal lines indicate gaps in the sequence. See section 5.1.1.

The alignments evidently contain many gaps. The combination of poor coverage, high divergence from the reference, adapter contamination, sequencing errors, and degradation-induced damage (see figure 4.2), causes many fragments to align in a way that clearly does not reflect homology (more details in section 5.3.1 and figure 5.2). Poorly-aligning sections may also represent the confounding influence of sequences transferred horizontally between organelles and the nuclear genome in plants: A 2008 study, for instance, located >40% of the grape vine chloroplast sequence in the mitochondrial genome of the same organism [31].

These challenges prompted development of methods to 'clean' the alignments (Section 5.2), enough to extract simple distance-based phylogenetic information that can be verified against the findings of previous phylogenetic studies.

	Modern Total		*Modern In Final Analysis*		*Ancient Samples*
	Samples Sequenced	**Species**	**Samples Sequenced**	**Species**	
Bistorta	9	1	7	1	9
Draba	44	31	35	26	4
Ranunculus	37	30	22	17	8

TABLE 5.1: Sample counts for modern specimens used in phylogenetic analysis. The species counts refer to the total number of species represented, since in many cases mutiple specimens of the same species were included.

5.2 Methods

5.2.1 Samples

This chapter focuses on three study taxa (see chapters 3 and 4): The alpine bistort, *Bistorta vivipara* (Polygonaceae), a member of the Whitlow-grass genus *Draba* (Brassicaceae; these are not true grasses), and a buttercup from genus *Ranunculus* (Ranunculaceae). The three study genera are all typical of circumboreal arctic-alpine species complexes that weathered glacial-interglacial cycles in the north, and make excellent representatives of the mammoth steppe forb communities. One-hundred-and-eleven modern and ancient samples were sequenced (Table 5.1), as described in chapter 4 (see also DS:Data for detailed information on the sequencing runs). The herbarium samples (Table 5.2.1) representing *Draba* and *Ranunculus* species present from the Yukon were collected in Northern Canada and Alaska, with one sample (*Ranunculus repens*) being collected in Adelaide, Australia. *Bistorta vivpara* samples were acquired from Canada, Northern USA, and Eastern Siberia, which was linked to the Northeastern North America via the Bering land bridge during full Glacials such as during the LGM (see section 1.3). More information on the samples is included in the data supplement (DS:Samples). Previous research on the biology, phylogeography, and genetics of these taxa are described in the context of new results in section 5.3.3.

Genus	**Species**	**Library ID**	**Location Notes**
Bistorta	vivipara	bm_1	Moosehorn Lake B.C.
Bistorta	vivipara	bm_2	Talaya Pass, Russia
Bistorta	vivipara	bm_3	Tiffany Lake Washington
Bistorta	vivipara	bm_4	Terminal Lake B.C.
Bistorta	vivipara	bm_5	Little Blue Sheep Lake, B.C.
Bistorta	vivipara	bm_6	Kindersley Pass, B.C.
Bistorta	vivipara	bm_7	Magadan, Russia
Bistorta	vivipara		Quartz Creek
Draba	albertina	ms_46	Kluane National Park, St Elias Trail

Draba	alpina	ms_47	Ivvavik National Park, Clarence Lagoon
Draba	arabisans	ms_48	Thunder Bay, Lake Superior
Draba	aurea	ms_49	Ruby Mountains, trail to Pika camp
Draba	aurea	ms_51	Euchre Mountain
Draba	aurea	ms_52	Wolf Lake
Draba	cana	ms_53	Wind River
Draba	cinerea	mm_-1,lon_1	Gravel Pit Rd
Draba	cinerea	ms_54	Ivvavik National Park Ptarmigan Bay
Draba	corymbosa	ms_55	Mackenzie Mountains, Canyon Range plateau
Draba	glabella	ms_59	Kent Peninsula
Draba	groenlandica	mm_-5,lon_5	Ellesmere Island
Draba	incana	ms_60	Kenora, Penn Island fuel cache, Hudson Bay coast, 16km SE of Manitoba border
Draba	incerta	ms_61	Bonnet Plume Drainage, Gillespie Lake
Draba	incerta	ms_75	Brute Mountain
Draba	juvenalis	ms_62	Keno Hill
Draba	lactea	ms_58	South Canol rd, Groundhog Creek.
Draba	lactea	ms_63	Nunavut Gravel Pit Road
Draba	lactea	ms_85	Ivvavik National Park Ptarmigan Bay
Draba	lonchocarpa	ms_64	Bonnet Plume Drainage, Pinguicula Lake
Draba	lonchocarpa	ms_84	Kotaneelee Range
Draba	lonchocarpa	ms_86	Mackenzie Moutains, Dall Sheep Shoulder Site D
Draba	nemorosa	ms_66	Aishihik Lake north shore
Draba	nivalis	ms_67	Mount McIntyre
Draba	oblongata	ms_70	Mount Klotz camp
Draba	ogilviensis	ms_71	Summit Creek Flats
Draba	palanderiana	ms_73	Wind River, Deception Mountain

Draba	pilosa	ms_68	Nunavut, 7.5km along the road to Mount Pelly
Draba	praealta	ms_69	Mackenzie Mountains, Backbone Ranges: Little Dal Lake
Draba	reptans	ms_74	Victoria, Canada
Draba	simmonsii	ms_76	Nunavut, 7.5km along road to Mount Pelly
Draba	stenoloba	ms_77	Vuntut National Park, vicinity of Snowdrift camp
Draba	stenopetala	ms_78	Yukon-Charley National Preserve Mt Frosty 1
Draba	subcapitata	ms_79	Victoria Island Trunsky Lake
Draba	verna	ms_80	Victoria. Royal BC Museum
Draba			Quartz Creek
Ranunculus	abortivus	ms_26	Alta Glenevis
Ranunculus	acris	ms_23	York Factory
Ranunculus	aquatilis	ms_22	Wolf River Camp 2
Ranunculus	bulbosus	ms_27	Niagara, Wainfleet
Ranunculus	cooleyae	ms_28	Haines Alaska, Chilkoot Valley, pond ridge
Ranunculus	cymbalaria	ms_29	Dawson City
Ranunculus	eschscholtzii	ms_5	Telkwa Microwave Tower
Ranunculus	flammula	ms_31	Mackenzie Mountains, Careajou Lake drainage
Ranunculus	gelidus	ms_32	Gates of the Arctic National Park, Castle Mountain
Ranunculus	gmelinii	ms_35	LaBiche between LaBiche and Liard Rivers
Ranunculus	Iapponicus	ms_36	Shingle Point area, mainland opposite
Ranunculus	pygmaeus	ms_11	Victoria Island; 30 Mile Creek. Nunavut
Ranunculus	repens	ms_12	South Surrey, Whitehorse
Ranunculus	sabinei	ms_14	Cape Bathurst
Ranunculus	sceleratus	ms_13	Atlin, SE of Telegraph Ranch
Ranunculus	sulphureus	ms_15	Printer's Pass
Ranunculus	turneri	mm_-4,lon_4	Yukon-Charley National Preserve

Ranunculus	turneri	mm_-7,lon_7	Nuntun National Park
Ranunculus	turneri	ms_18	Vuntut National Park
Ranunculus	turneri	ms_19	Engineer Creek
Ranunculus	turneri	ms_20	Yukon-Charley National Preserve, Mount Casca camp
Ranunculus	turneri	ms_21	Vuntut National Park, Dog Creek camp
Ranunculus			Quartz Creek

TABLE 5.2: Sample details for all herbarium specimens appearing in the phylogenetic analysis (see section 5.3).

5.2.2 Data Handling and Mapping

Merged and trimmed ("QT"; see section 4.2.3.1) reads from each sample were mapped to the corresponding generic consensus chloroplast sequence using BWA as described in section 4.2.3.2 (commands given in DS:Code:1.5.1). The QT reads were chosen because sequence accuracy was considered a priority. PCR duplicate removal was performed with PicardTools MarkDuplicates (DS:Code:1.5.1). Mapped reads from different libraries originating from the same original extract were pooled (DS:Code:1.5.2). Consensus sequences were then generated with bcftools (DS:Code:1.5.3). Long stretches of 'N's were also removed, as these tended to cause crashes and poor performance in the multiple alignment steps (DS:Code:1.5.3). An outgroup chloroplast sequence was added to the multi-species datasets, with *Arabis alpina* (NCBI Genbank Accession HF934231) as the outgroup to *Draba*, and *Aconitum chiinsanense* (NC029829) for *Ranunculus* (DS:Code:1.5.4). Empty sequences were removed manually, and multiple alignments were performed using Clustal Omega [58] (DS:Code:1.5.5).

5.2.3 Alignment Cleaning

The script MTRW_alignment_cleaner.pl (DS:Code:2.4) was written to purge alignments of data that are likely to be misleading in the context of this and similar studies (see section 5.1.1). Such a procedure is necessary wherever standard read-trimming is inadequate to remove noise from the final dataset, or where significant noise is introduced subsequent to read processing, errors or inconsistencies in the multiple alignment step. In such circumstances, this script represents an advance on other alignment cleaning tools such as Gblocks [14], which remove whole columns only and do not operate on individual sequences, whereas MTRW_alignment_cleaner.pl considers both rows and columns in the

alignment. These sequence-wise (row) operations are crucial for removing errors when they are detectable only in scans over sequences rather than over columns or blocks of columns, and when informative columns are sparse, meaning that their removal should be avoided where possible to retain the highest possible amount of quality data (DS:Code:1.5.6).

The script first removes likely candidates for adapter contamination and ancient DNA damage by requiring any 5 nt block of sequence adjoining a length of 3+ gaps to be free of variant sites. Any ends breaking this rule are deleted progressively until they meet this criterion. To achieve this, the script first represents each sequence as a string of symbols (a match-string, or m-string) stating whether a particular site is gap/invalid nucleotide/N ('-'), a variant site at which multiple different bases occur in the alignment ('v'), or a match, that is, identical to all the other characterised bases in the alignment ('m'). Regions covered by only one sequence are considered uninformative and deleted. A sequence whose m-string reads '-----mmvmmmmmmvmmmmmv------' would be recognised in this step, and the ends deleted to yield a shorter sequence with m-string '--------mmmmmmvmmmmm-------'. The comparison shown between the upper and lower panels of figure 2 demonstrates the effectiveness of this method at removing misplaced sequences without removing many putative informative variant sites, or matching regions that might contribute to judging the distance between pairs of sequences. Secondly, the script removes sites that might be caused by sequencing errors or damage-induced substitution, by considering only non-singleton variants.

5.2.4 Distance-Based Phylogenetics

Distance matrices for each alignment were made using MTRW_distmat.pl (DS:Code:2.3). This script generates pairwise percentage distances that count only sites that are properly characterised (that is, not a gap, and N, or an IUPAC ambiguous code such as R, K, U...) in both sequences being compared. A threshold of 150 was set for the number of sites that must be shared between a pair of sequences to generate a distance estimate. Further investigation of the effect of coverage upon the distance matrix is given in the following chapter. Sequences with very low coverage did not meet the threshold for distance estimates in most comparisons, and were progressively removed beginning with the lowest, until all remaining pairs had distance estimates.

The relationships between groups were assessed usings PCA and visual inspection of the sequences. For each study genus, a PCA was produced using the script MTRW_phylo.R (DS:Code:2.2), which calls upon the plot3D package to show the first three components of the PCA in three dimensions. This was found to be highly preferable to 2D plots: very rarely did 2D representations ever succeed in showing the grouping of samples without superimposing samples that were broadly separated on subsequent dimensions. The 3D PCA plots were rotated manually to best show the grouping of samples. The sizes of the datapoints were made proportional to a measure of sequence completeness, calculated as the percentage of variant sites in the alignment represented in the sequence (see also chapter

6). The datapoints were coloured according to their classifications in earlier studies (see figure text for figs 5.4, 5.6, and 5.8, and section 5.3).

5.2.5 Testing Robustness to Sequence Damage

To ensure the success of the methods used to purge noise from the data and glean useful phylogenetic information, the entire method was repeated on an alignment to which artificial DNA damage was added. This is covered in more detail in the following chapter. Briefly, the new program SimWreck (chapter 6) was employed to add uniform deamination-like damage to the original alignments (DS:Code:1.5.7). The damage was applied under a "worst case scenario" principle: it is unlikely in the extreme that so many bases in the final alignment could be altered by deamination, with on average every third C or G undergoing a transition to T and A respectively. The alignments were then cleaned, the phylogenetic analysis performed, and the results compared to those using the original data. The results demonstrated that the distance matrices were highly robust to sequence damage after alignment cleaning, and there were no significant changes to the grouping of taxa or the relationships between groups in any of the three genera studied. More details are given in chapter 6.

5.3 Results and Discussion

5.3.1 Alignment Cleaning

Visual inspection of the sequence alignments before and after cleaning clearly shows how the cleaning script successfully removes dubious regions that would otherwise confound the analysis.

The top panel in figure 5.2 highlights instances where the alignment is probably affected by such confounding errors: A red square surrounds a cluster of mismatches on the end of a read, probably caused by adapter contamination. The rectangle surrounds an "orphan" section of a consensus sequence, aligned to a region to which it obviously does not share homology. The green C-to-T change in the red circle is consistent with possible deamination or adapter contamination. Such errors may artificially create the appearance of synapomorphy between sites on sequences where errors converge on the same nucleotide, a probable example of which is indicated by a red arrow.

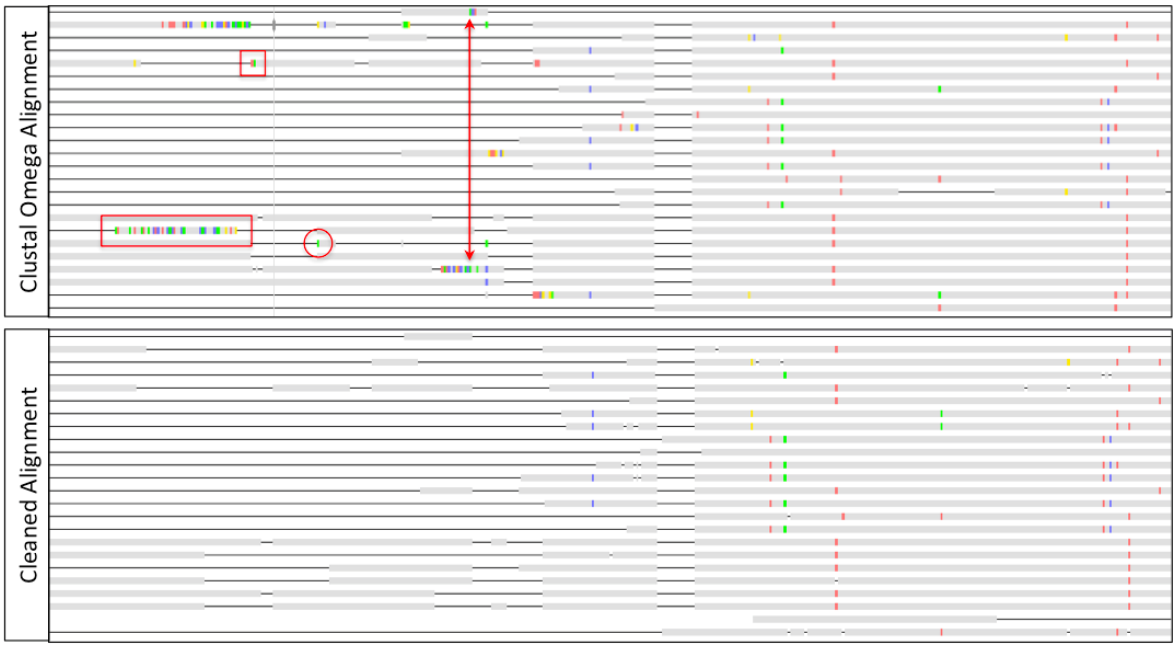

FIGURE 5.2: Comparison of a portion of the *Ranunculus* alignment (positions 1–130) before and after cleaning with MTRW_alignment_cleaner.pl. Nucleotides are coloured only when they disagree with the majority consensus base for the site. **Red arrows/boxes**: refer to section 5.3.1 text.

5.3.2 Complete Ancient Draft Chloroplast Genomes

Two ancient specimens (Draba_GZ1103_exttun4 and Ranunculus_GZ0908_extdar4) yielded enough sequence data to achieve near-complete coverage of the characterised parts of the draft generic consensus sequence. The script MTRW_cov_from_bam.pl (DS:Code:2.5) was written to calculate coverage and depth for samples where the reference contains no-call 'N' bases (these alignment positions are simply excluded from the calculations). After combining all sequencing runs from all libraries made from each of these extracts, Draba_GZ1103_exttun4 shows 99.01% coverage of the generic reference with a mean read depth of 61.2, which owes to deep-sequencing the sample on an Illumina HiSeq platform to produce a very large volume of data. Ranunculus_GZ0908_extdar4 with 98.17% coverage and mean depth 10.25, owes mostly to an anomalously high endogenous DNA rate (see chapter 4, table 4.4 and figure 4.6). To my knowledge, these two samples represent the oldest draft chloroplast genomes sequenced to date. Using HTS and targeted enrichment, a previous study has recovered 83.6% of the Long Single Copy (LSC) region of an an ~8000 ka Bottle Gourd chloroplast genome[43], and the same study also generated a >99% complete LSC for a ~900 ya sample. The draft chloroplast genomes sequenced here extend this time record by an order of magnitude.

5.3.3 Phylogenetics

5.3.3.1 The Effects Of Alignment Cleaning On Distance-Based Phylogenetics

The removal of singleton variants (Section 5.2.3) may be a contentious choice, as it represents a trade-off: variants that do not occur in at least two individuals are purged along with noise in the data. The results from this chapter (Sections 5.3.3.3, 5.3.3.4) show that the retained data are sufficient to gain ample insight into natural systems when the samples form clear groups. There are, however, certain limitations and considerations.

Variable columns where fewer than 2 (one variant) + 2 (an alternative variant) = 4 sequences are characterised are inevitably lost, and those variants lost will therefore include most of those singletons that occur on longer terminal branches of the phylogeny being investigated. To illustrate, consider the hypothetical phylogeny in figure 5.3. Branches along which substitutions will be retained using the current method are indicated with red lines. The majority of this hypothetical tree will be largely "hidden". However, even in the taxon-poor hypothetical tree below data could still reveal the relationships between three groups: A, B, and C+D (which can be identified as basal based upon other information, allowing the tree to be rooted).

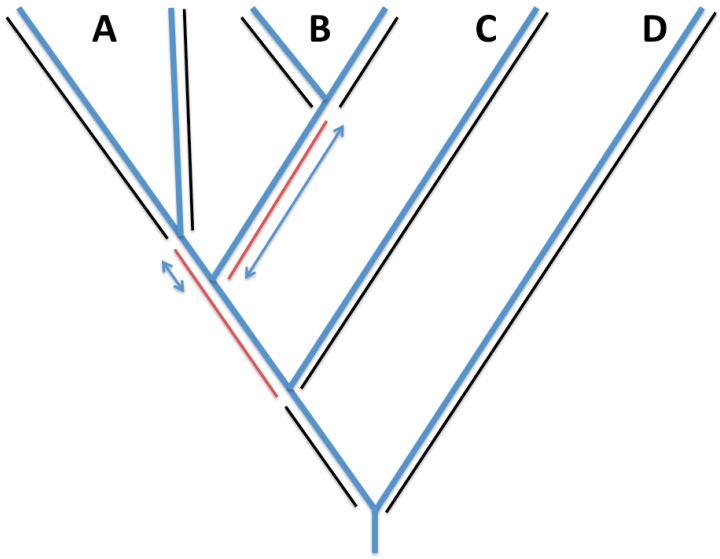

FIGURE 5.3: Hypothetical phylogenetic tree demonstrating the implications of retaining only shared variants. See section 5.3.3.1.

As figure 5.3 demonstrates, no absolute estimate of some of the branch lengths may be possible, but it may be at least possible to estimate the relative lengths of the two branches marked with blue arrows: substitutions occurring along the left of the two are expected to occur less often than on the right, owing to the earlier divergence of taxa within A compared to taxa within B. As a result, the number of variants unique to group A relative to the number unique to group B ought to reflect this difference, as should the number of variants shared A + (C and D) relative to those variants shared B + (C and D). Given enough data,

and an amenable tree topology, the same type of reasoning can even produce means to explore the distances between basal single-taxon branches, such as C and D in figure 3. Some similar procedures, including the popular *ABBA-BABA* test [23], have been used as a means of inferring introgression between diverged populations. I return briefly to this theme in discussing experimental findings, but since branch lengths were considered unreliable and of secondary importance, the difference between groups of closely-related samples makes up the main focus of the chapter.

5.3.3.2 *Bistorta vivipara*: Dispersal and Persistence

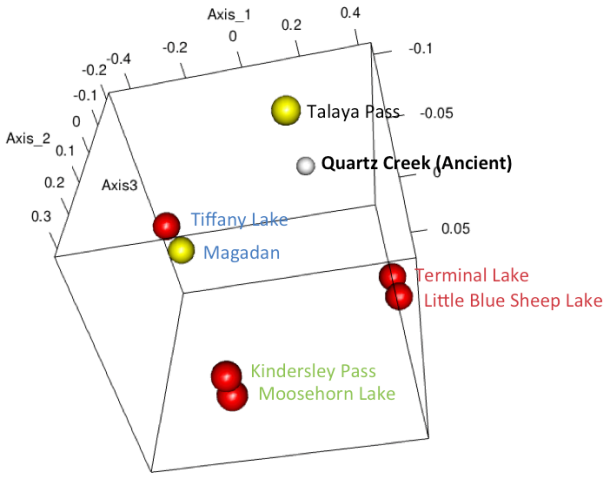

FIGURE 5.4: PCA showing relationships between *Bistorta vivipara* samples (first three principal components). Size is proportional to sequence completeness, with the smallest being the pooled ancient sample with 25.53% of the variable sites characterised, and the largest being Magadan and Tiffany Lake, both with 100%. **White:** Ancient sample (East Beringia). **Red:** North American samples. Tiffany Lake WA, NW USA; Terminal Lake, BC, NW Canada; Little Blue Sheep Lake, BC, NW Canada; Kindersley Pass, BC, W Canada; Moosehorn Lake, BC, NW Canada. **Yellow:** Eurasian Samples. Magadan, Magadan Oblast, SE Russia; Talaya Pass, Magadan Oblast, SE Russia.

Bistorta vivipara is a cold-adapted arctic-alpine species that grows in a range of environments including high-altitude meadows, moist fens, and other nutrient-rich sites. The plants propagate vegetatively via rhyzomes, and asexual bulbils that emanate from a spike-like terminal raceme [16]. The bulbils often sprout before separating from the parent plant, giving the species its name. While pink or white flowers form along the top of the spike, the stamens largely abort before seeds are produced, with various critical steps in the seed development process failing in the bulk of experimental plants [21]. The bulbils provide food for fauna including ptarmigan, reindeer, and arctic ground squirrels, while the rhyzomes were preserved and consumed by First Nations populations. *B. vivipara* is common and widespread with a holarctic distribution, and a geographically-separate population exists in Colorado [12, 48]. A pilot study using nuclear RAD markers from a single individual in each population suggested that gene flow between the Colorado and Alaska was minimal during the

Last Glacial Period [12]. Chloroplast DNA from a Holarctic sample set of B. vivipara has revealed remarkably low genetic diversity, which is unexpected given its high morphological variability and very large range [10, 48]. Chloroplast DNA also revealed a remarkable lack of genetic structure at a continental scale, with most haplogroups mixing in Colorado, northward along the Rocky Mountains, Ogilvie Mountains, and Alaska Range, and in widely separated locations within Siberia and Eastern/Southeastern Eurasia.

The data presented here (figures 5.4 and 5.5) confirm this same lack of geographical structure, with samples from the west of the Bering strait twice grouping together with samples from the east. The ancient consensus sample from Quartz Creek, YT, Canada, falls closest to the West-Beringian Talaya Pass sample collected in Magadan Oblast, Russia (black text in figure 5.4). Samples from B.C., Canada, fall into two distinct groups (red and green text respectivey), while the sample from Tiffany Lake, WA, USA, south of the North American ice sheets, is grouped with a second Russian sample from near the town of Magadan. This long-range haplotype mixing, is compatible with the interpretation that a lack of strong genetic population structure was present even before the LGM, and perhaps persisted through the course of several glacial-interglacial cycles.

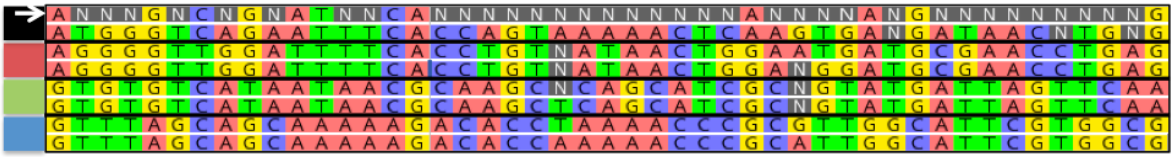

FIGURE 5.5: All variable sites identified in the *Bistorta vivipara* chloroplast alignments. Coloured blocks on the left correspond to groups designated, with colours corresponding to the text in figure 5.4. The ancient consensus is marked with a white arrow.

In agreement with previous studies, this pattern suggests that *B. vivipara*, along with several other arctic taxa, are surprisingly mobile. This mobility must apply to the propagules, rather than just the pollen: chloroplast DNA is usually maternally inherited, and thus passing chloroplast haplotypes long-distance via anaemogamy is usually impossible. While *B. vivipara* can grow lateral rhizomes and then send up new plants, this seems an unlikely method of long-distance travel, especially between land masses and mountain ranges. The human transport of rhizomes is also a poor explanation, given that the ancient samples studied here predate human colonisation of the Americas. The vegetative bulbils therefore seem likely to be the main mode of dispersal. The bulbils are transportable by avian herbivores such as grouse and ptarmigan, which exhibit some migratory behaviour [32], which at very least offers chances for individual birds to be blown long distances in extreme weather events. Dispersal by mammals including ground squirrels is also possible, and migratory mammals such as the megaherbivores of chapter 2 may have allowed some long-distance overland transport.

The long-term survival of geographically-overlapping chloroplast haplotypes suggests that plants remain competitive and hardy no matter what the origin of their chloroplasts. In

other words, the plant's competitive capacity seems decoupled from its maternal ancestry. One explanation could be that the plant specialises in a very particular environment that happens to be widespread and allows the plants to migrate without significant changes in morphology or survival strategy. A homogeneous suite of selective pressures acting on a largely asexual population might allow the accumulation of selectively-neutral mutations in different haploid lineages over time, driven largely by drift, since the homogeneity of selective pressures might have led the local communities toward a stable evolutionary equilibrium, favouring equally endemics or migrants from near-identical environments elsewhere. This explanation is supported by the relative continuity of the alpine and arctic meadows that *B. vivipara* inhabits, but is in direct conflict with the red queen hypothesis (a principle in evolutionary biology stating that even organisms inhabiting a static environment constantly evolve in response to intra- and interspecific competition) [63], and assumes that such a stable environment could exist over multimillenial timescales, despite the climatic and environmental shifts of the Late Quaternary. Also, as previously mentioned, *B. vivipara* does display a great deal of morphological variation, having for instance more gracile and more robust growth forms, variable stamen number, and flowers ranging in colour through pink to white. While some of this variability can be explained by phenotypic plasticity responding to environmental variables[7, 10, 29, 62], it suggests that variability is probably selectively beneficial, and yet that selection has neither rewarded variants of the plant that increase their genetic variability by outcrossing, nor strongly influenced the geography of chloroplast lineages. This morphological variability is perhaps then related to autogenic gross changes in the nuclear genome, and seems likely to be associated with its highly flexible ploidy, with widely varying chromosome counts being recorded suggesting ploidy levels from diploid to decaploid[45]. A large multiplicity of important genes in high-ploid populations may allow for faster generation of potentially useful variants, or for plastic responses to environmental cues based upon complex regulatory networks. HTS data from herbarium specimens sequenced in this study may prove valuable for future investigations into these mechanisms, though the absence of chromosome count data for the specimens is a current limiting factor.

5.3.3.3 *Draba*: Hybrid Survivors

Draba, the most speciose genus in the family Brassicaceae, mainly consists of small perennial herbs growing in arctic, subarctic, and montane environments, with many members specialised for survival in extreme cold[16]. Known for their taxonomic complexity, members of the genus (and of the family in general) form the basis of much research on sympatric speciation via chromosome duplication, hybridisation, and polyploidisation[44]. The effort is complicated, especially in the arctic *Draba*, by the recurrent formation of polyploids, and the existence of cryptic species which, while morphologically identical, produce infertile offspring[44, 46].

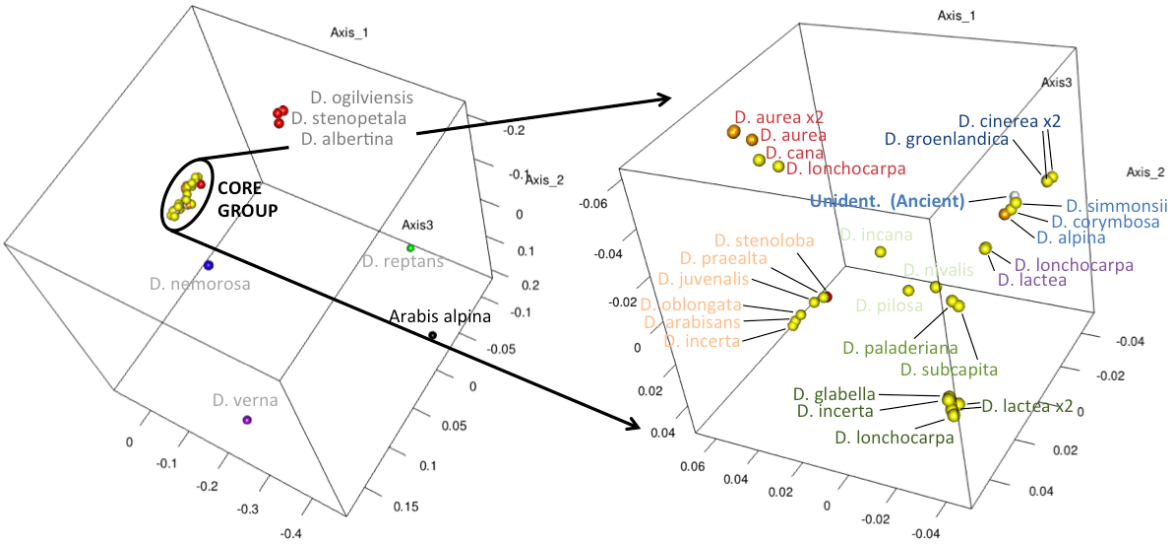

FIGURE 5.6: PCAs showing relationships between *Draba* specimens (first three principal components). The PCA on the right is constructed using only the entries in the distance matrix for those samples identified as the "core group" on the left. Size is proportional to sequence completeness, with the lowest being the pooled ancient sample with 81.71% of the variable sites characterised, and the highest being D. glabella with 99.51%. Colours (besides white and black) correspond to classifications used in Jordon-Thaden et al. (2010) [41]. **White:** Previously unclassified. **Black:** Non-*Draba* (Outgroup). **Green:** Classified as *Draba*, but genetically falling outside the genus. Not found in the Yukon. **Purple:** Basal *Draba*. Not found in the Yukon. **Blue:** Group I. **Red:** Group II. **Yellow:** Group III. **Orange:** Taxa whose members variously fall into groups II and III. See section 5.3.3.3

The present analysis essentially reconstructs the genetic groupings identified by Jordon-Thaden et al. (2010) [41]: *D. verna* falls outside the genus, *D. reptans* is a basal member, and three inner ('core') genetic clades can be clearly identified (figures 5.7 and 5.6, left panel and figure text). Since the separation between the two most basal groups in the phylogeny is expected to the be obscured by the alignment cleaning process and the 'variable sites must occur at least twice' rule (see section 5.3.3.1), the distances between these two taxa and the others are not directly proportional to their true separation, with the distance between the pair representing the expected effect of stochastic convergences in state with members of the non-basal parts of the tree along these longer, basal branches (these convergences probably represent a mixture of genuine substitutions and misaligned sections that escaped the cleaning process affecting divergent chloroplast sequences more badly). The presence of *D. nemorosa* in the Yukon represents an exception for clade I, whose members are usually found in Europe and the Far East. Its placement in the PCA faithfully represents the relationship between the three clades (Basal+Outgroup (I (II + III))) established by Jordon-Thaden et al. (2010), demonstrating the ability of the method used to recover the relative lengths of internal branches, even when the internal branch ends in a single specimen. Clade II is mostly confined to the North American cordillera, with exceptions in Beringia and Asia, three of which are confirmed here. Clade III has some cordilleran members, but is distributed across

the boreal former mammoth steppe (see section 1.3), and northward into once-glaciated areas such as Greenland. The existence of several species that variously fall into groups II and III likely represents hybridisation between members of these clades. The hybridising taxa from each region may have used this ability to capture and store novel variation developed by their counterparts during periods of glacial separation, upon the opening of ice corridors each interglacial. The present study suggests that, failing a specimen misidentification, *D. stenoloba* can be added to this list of hybridising taxa. Misidentification is not a likely cause: with a characteristic sparse raceme of long silicles, *D. stenoloba* is quite distinct from the Yukon clade III *Draba*, though it bears close similarity to *D. albertina* from group II (having been classified as the same species by some authors, e.g. Cody (2000)) suggesting a possible hybridisation partner. Reported chromosome counts suggest both taxa exhibit polyploidy[16], with *D. albertina* probably hexaploid (2n=48, assuming x=8), and *D. stenoloba* a possible uneven polyploid with 2n=40.

This study provides a detailed look inside the structure of clade III, in which the ancient *Draba* from Quartz Creek is placed (figure 5.6, right panel). The PCA reveals four arm-like clusters (henceforth just *clusters*) radiating out from a few central taxa, which demonstrates the ability of this method to display the relationships between groups of taxa internal to the tree. This arrangement thus suggests radiation of chloroplast haplotypes at around the same time. The central taxa (text pale green and medium green) probably represent a valid radiating group, too, attracted approximately equally to the taxa in each cluster, which repel each other resulting in a tetrahedral arrangement. The sequence cleaning process, hybridisation, concerted evolution, and the lack of a characterised whole-chloroplast substitution rate for *Draba* make estimating the divergence time difficult, but Koch & Al-Shehbaz (2002) [44] suspect that this core group (designated clade VI in the study) probably diverged within the last 0.5–1 million years, suggesting the influence of the strong climate fluctuation events of the Upper Pleistocene (see chapter 1, figure 1.4). The five clusters are therefore *D. aurea* + *D. cana* + *D. lonchocarpa*, *D. cinerea* + *D. groenlandica* + *D. simmonsii* + *D. corymbosa* + *D. alpina*, *D. lactea* + *D. glabella* + *D. incerta* + *D. lonchocarpa*, *D. stenoloba* + *D. praealta* + *D. juvenalis* + *D. oblongata* + *D. arabisians* + *D. incerta*, and *D. incana* + *D. nivalis* + *D. pilosa*. *D. palanderiana* and D. *subcapita* fall between the last three groups.

This star-like radiation poses a challenge, since, while the species' ranges are very poorly delineated, inspection of specimen collection maps [16, 18] reveals that the various groups clearly do not observe any obvious differences in range at the scale of the Yukon Territory or more broadly—very similar to the pattern seen with *Bistorta vivipara* haplotypes. Our interpretation in that case involved the role of morphological plasticity, the retention/acquisition of genetic variation via polyploidy, and domination of a homogeneous biome, to producing a species that has the capability to migrate far and remain competitive—and that this quality was decoupled from the chloroplast haplotype. All these qualities are also seen in the Yukon *Draba*.

Variation-management mechanisms have been discussed, and appear to be especially prevalent even in this limited sample of Beringian *Draba* with *D. aurea*, *D. alpina*, known to appear

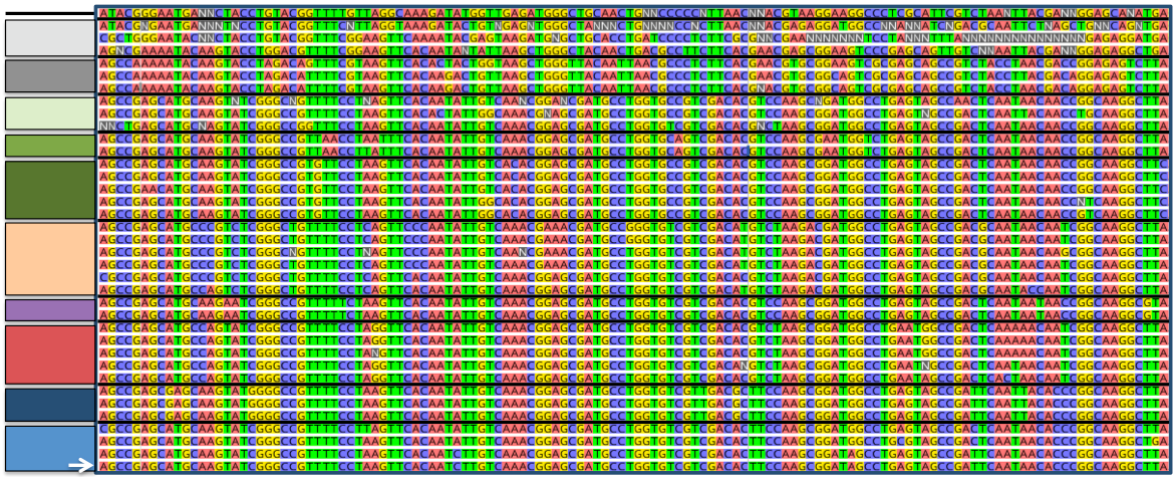

FIGURE 5.7: The first 130 of 2,236 variable sites identified in the *Draba* chloroplast alignments (showing only sites at which the ancient sample's state is known). Coloured blocks on the left correspond to groups designated, with colours corresponding to the text in figure 5.6. The ancient consensus is marked with a white arrow.

in multiple chloroplastic subgroups[41]. The current study demonstrates the same for *D. stenopetala*, and by illuminating groupings within the Yukon clade II, we also note the same for *D. incerta*, *D. lactea*, and *D. lonchocarpa*. The latter occurs in three separate groups, twice with *D. lactea*, suggesting possible bidirectional hybridisation. *D. lonchocarpa* has in fact also been classified as a variety of *D. nivalis*, but none of those sampled falls close to that species in the PCA. A second unexpected finding is that the Flora of North America[18] records only diploid chromosome counts for this largely Cordilleran species, despite its clear affinity for hybridisation. Further investigation of this species' reproductive genetics may help to resolve these intriguing discrepancies.

Domination of a homogeneous biome is certainly arguable for the arctic-alpine *Draba*, indeed the very existence of so many arctic-alpine taxa in general suggests that adaptation to one implies a heightened ability to survive in the other: they are in essence very similar environments and do by their nature occupy long and continuous stretches of Earth's surface, in lines dictated by climate for the former and orogeny for the latter. At the risk of oversimplifying, it is easy to see how this homogeneity could apply in both time and in space, since as climates warm and cool through time, the alpine zone can migrate up or down in altitude, while the arctic zone migrates up and down in latitude. Mountain ranges running longitudinally (such as the Urals and the Rockies) mean the two will continue to intersect at some point at any given time. The study group of Yukon *Draba* have a real affinity for well-drained, cold, and harsh environments, with all habitat descriptions including at least one of the terms rocky, slope, talus (also called scree), or gravel. *Draba* dispersal mechanisms [30, 40] have not been extensively studied, but the environments described above often coincide with nesting sites, implicating bird dispersal. *Draba* seeds have also been identified in mammoth intestines, allowing the possibility of dispersal by migratory mammals.

The ancient *Draba* from Quartz creek appears to be closest to *D. simmonsii* (once classified as *D. alpina* var. *simmonsii*), and *D. corymbosa*. Morphologically, the ancient silicles have been likened to *D. cinerea*, [53] which falls nearby in the same cluster. This is probably a genetically reticulated group itself, with members having been previously categorised variously as variants of one and another, or of unsampled species such as *D. alpina* and *D. murrayi*. A BLAST search of the Genbank Nucleotide database, using a *D. simmonsii* rbcL (accession number KC482639.1) sequence as a query also reveals a 100% match with a specimen of *D. incana*, suggesting further reticulation between clusters. Most members appear to have a distribution centred around Beringia, meaning the ancient lineage dating to over 50 ka has persisted in the same region. To infer the morphology of the ancient *Draba*, we must rely upon congruence between the chloroplast genotype, biology, and the appearance of the silicles. The closeness of the genetic match, the dry meadow habitat preferred by Arctic ground squirrels, and the flattened elliptic-lanceolate silicles (in a raceme arrangement that can be occasionally be observed intact in the squirrel nests) suggest the ancient specimen was more like its closest genetic match in the current dataset, *D. simmonsii*, than other members of this cluster: the next two closest genetic matches sequenced here (*D. corymbosa* and *D. alpina*) are exceptional to the core *Draba* in preferring moist tundra to dry environments, and the ground squirrels inhabit well-drained open hillside meadows, and the ancient silicles are shorter and more elliptic than those of *D. incana* (FIG. 3.3). *D. simmonsii* occurs "mostly on dry, open ground with sparse or open vegetation or on open patches in fresh to dry closed meadow or heath vegetation"[27], consistent with a persistent arctic-alpine niche in the cryoxeric pre-Holocene mammoth steppe as described in chapters 1 and 2. So far as I can find, no identifications of *Draba* from in this cluster have placed them in the non-alpine region of the Klondike that includes Quartz Creek mine where the samples were found. As such, a northward shift in range since the glacial period cannot be ruled out.

5.3.3.4 *Ranunculus*: Adaptive Stalwarts

This chapter has explored a repeated theme of species complexes that migrate along corridors of a homogeneous biome (in both time and space), and to which they are able to adapt via specialised mechanisms for maintaining genetic diversity and morphological plasticity. Some Yukon *Ranunculus* seem to have converged on this strategy suite, too, with some interesting differences.

A more diverse genus than *Draba*, the Boreal *Ranunculus* have been grouped into seven genetic clades, that exhibit some tendencies towards certain habitats and ranges, though almost all are perennial herbs with acrid juices that repel herbivores. Hybridisation is less common in *Ranunculus* than *Draba*, but variable ploidy occurs within all the major clades. The current study successfully replicates some of the findings of previous molecular work, recovering the clades A–F established by Hoffmann et al. (2010) [36] (FIG. 5.8).

R. bulbosus and *R. repens* group together to form clade A, a temperate group with native ranges centred around Europe (figures 5.8 and 5.9).

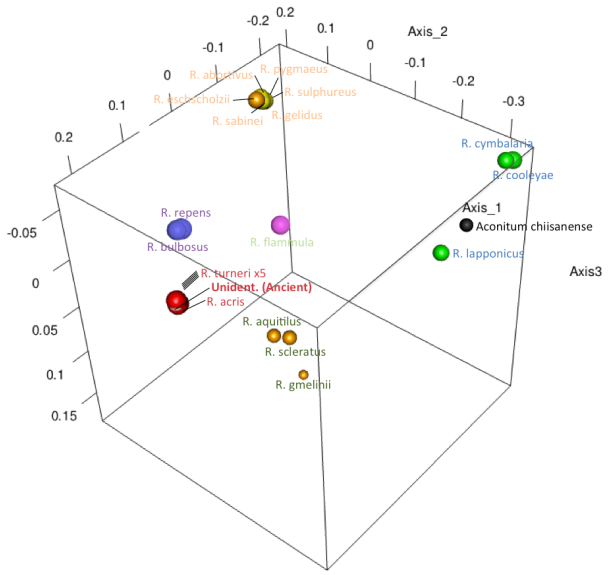

FIGURE 5.8: PCAs showing relationships between *Ranunculus* specimens (first three principal components). Size is proportional to sequence completeness, with the lowest being *R. gmelinii* with 53.17% of the variable sites characterised, and the highest being a specimen of *R. turneri* with 98.43%. Colours (besides white and green) correspond to classifications used in Hoffmann et al. (2010) [36]. **White:** Previously unclassified. **Black:** Non-*Ranunculus* (Outgroup). **Green:** Basal *Ranunculus* (alternately placed in separate genera including *Halerpestes* and *Lapponicum*). **Blue:** Clade A, temperate *Ranunculus* species (introduced in the Beringia region). *R. bulbosus* was not analysed in Hoffman et al. (2010), and was identified as a member of this clade according to Horandl et al. (2005) [37] **Red:** Clade C. **Purple:** Clade D. **Yellow:** Clade E. **Orange:** Clade F, including *R. eschscholtzii*, whose placement has been considered uncertain [36].

No representatives of the Eurasian clade B were included.

Clade C consists largely of microtaxa—species that very closely resemble one and another, consistent with their late radiation in the Quaternary. It also includes the ancient species from Quartz Creek. The placement of this ancient species with *R. acris* and *R. turneri* is reasonably consistent with the morphology of the ancient achenes, which show the correct size, inconspicuous keels on the margin, and a glabrose, minutely reticulate exocarp [8]. Few of the achenes showed the characteristic strongly recurved beak, though most had signs of damage including the beaks being partially removed, perhaps during midden processing, or by the squirrels in collection and transport. The fruits may also have been collected earlier in the growing season, before their development was complete, in fact, intraspecific competition for resources might have encouraged this behaviour in ground squirrels. The sequence data do, however, provide some of the genetic resolution that Hoffmann et al. (2010) suggest is lacking in the ongoing project of resolving whether this clade truly merits division into genetically-distinct species, the distance matrix showing that the *R. turneri* specimens do indeed cluster together very distinctly from *R. acris*. The ancient specimen should be considered a member of *R. turneri*, differing from modern *R. turneri* specimens at, on average,

only 2–3 recovered variable sites across the entire non-repetitive chloroplast, while differing from *R. acris* at, on average, 26–27 variable sites.

R. turneri inhabits meadow environments, especially moist stream banks. Its co-ocurrence *B. vivipara*, which also favours moist alpine/sub-alpine meadows and fens, suggests the squirrels were likely foraging in river valleys on south-facing hillslopes, where these two environments intersect. The preservation of the nests in permafrost, and the observation that they were often water-logged, agrees that they were maintained somewhere near the water table. This unfortunately suggests that the nests recovered may be taxonomically biased towards those that spent time in liquid water, and hence in which DNA preservation will be lower. The Flora of the Yukon Territory[16] records Yukon *R. turneri* only in the far north of the territory, indicating a possible northward range contraction since the last glacial. As such, it seems that three of the ancient species identified in this study—R. turneri, D. simmonsii, and S. parryii plesius (see section 4.3.1)—all contracted northward since the time the ancient samples were deposited.

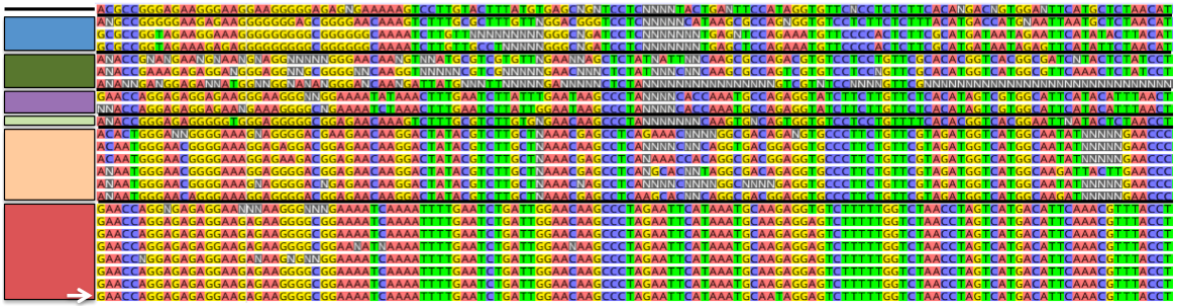

FIGURE 5.9: The first 130 of 2,236 variable sites identified in the Ranunculus chloroplast alignments (showing only sites at which the ancient sample's state is known). Coloured blocks on the left correspond to groups designated, with colours corresponding to the text in figure 5.8. The ancient consensus is marked with a white arrow.

Hoffmann et al.'s clade D is, in this study, represented by the aquatic *R. flammula*, its placement separately from the other groups being further confirmation of the random-convergence mechanism that allows the meaningful positioning of single, divergent taxa despite the singleton filtering steps (as discussed regrding the placement of *D. nemorosa* in section 5.3.3.3).

Like clade C, clade E is dominated by morphological microtaxa, and this is reflected by the tight groupings apparent in figure 5.8. Conversely, the variable morphology and habitat use among clade F is coupled with greater genetic separation between members of that clade. Hoffmann et al. (2010) explicitly note the uncertain placement of *R. eschscholtzii* and *R. gmelinii* variously in clades E and F. The specimens sequenced here suggest both clades are good evolutionary groupings, and R. eschscholtzii is unambiguously associated with clade E while *R. gmelinii* is close to the more early-diverging clade F, however close investigation of the nuclear genomes of multiple specimens may be required to improve the interpretation of data gathered so far.

As was the case with *Draba* and with *Bistorta*, the ancient *Ranunculus* is a part of a species complex with poorly-defined boundaries, which likely radiated during the quaternary. *Ranunculus*, however, is not exclusively an arctic-alpine genus, in fact, even its boreal members inhabit rivers and bogs, moist tundra, meadows, grassy meadows, and shrubland. It is possible that the comparative lack of hybridisation between *Ranunculus* clades is related to this habitat variability, since the range of traits expressed by mixing alleles from broadly different plants might be more likely to lead to outbreeding depression effects whereby the resultant intermediate traits in the hybrid leads to diminished survivorship in either of the habitats inhabited by the parents. The idea that taxa may survive by adapting to a biome that migrates, but does not disappear, during glacial-interglacial cycles may be applicable to *Ranunculus* too, with so many species adapted to life in and around streams, which themselves necessarily stretch unbroken for long distances, and are also necessarily generated wherever glaciated mountain tops reach an altitude (or in the case of polar ice sheets, a latitude) beneath which the ice cannot remain perennially frozen. This certainly applies to the ancient *R. turneri*, which prefers meadows and riverbanks, and whose range in around the northern Yukon area abuts the northwestern reaches of the Laurentide ice sheet, and the slopes of both the Ogilvie and Brooks mountain ranges. We can therefore imagine ancient *R. turneri*, along with close members of its clade, migrating along the rivers and outwash plains flowing from glacial and ice sheet margins, perhaps using the rivers themselves as a means of dispersing to the optimal meadow environments as they waxed and waned throughout the Quaternary. These same environments would no doubt have been shared with grazing herbivores (see chapter 2), which explains *R. turneri*'s prominently beaked achenes, used to hook onto animal fur for dispersal, as well as the need to accumulate acrid *Ranunculin* toxins to avoid consumption by the dominant grazers of the time. The use of hybridisation and polyploidisation as means of preserving diversity seems likely, at least within the two microtaxon-rich clades, in which both diploid and haploid species are recorded.

5.4 Concluding Remarks

The work presented in the last three chapters 3–5 advances the prospect of future studies on plant DNA of Late Quaternary age, providing the field with a standardised ancient DNA methodology, baselines for what results can be expected, and advice for assessing and optimising outcomes.

I remain hopeful that further work will allow genomic studies of such samples, however the difficulty of identifying homologous sequences in a sample with a highly volatile genome of unknown and variable ploidy, high gene copy numbers, and likely with many paralogues and pseudogenes, will be extremely difficult to overcome. It may be possible to glean useful information using analyses that are designed to be robust to such issues. This effort will be aided by the characterisation of the genomes of modern taxa closely related to the ancient—a possible application for some of the better sequence data produced in the current study. A more rigorous reconstruction and annotation of the chloroplast genomes of many

of the samples sequenced, including the ancient samples, may allow the investigation of the structural and functional variation in that genome.

At the time of completion, this work has produced the most ancient draft chloroplast genomes known to the researchers (see section 5.3.2). This represents an important step in botanical palaeogenetics and portends exciting new developments. Ancient chloroplast genomes show all the promise of ancient mitochondrial genomes, with the added benefit that at roughly ten times the length, more sequence information is available to give increased resolution to the results. Whole chloroplast sequences have become increasingly utilised as a phylogenetic tool since the advent of NGS [15, 50, 67], with much of plant evolution now constructed from data including the whole chloroplast sequence. Assembly pipelines that automate the reconstruction of novel chloroplast genomes are becoming publicly available [34, 49, 67]. Phylogenetic and phylogeographic studies on taxa including permafrost-preserved species will benefit from the addition of ancient taxa, both in the reconstruction of the history of these important species complexes [1, 3], and in the addition of internal tree branches which aid greatly in calibrating molecular clocks [55].

The reconstruction of ancient population demographics via phylogenetic methods such as the Bayesian skyline [22] is a further possibility. Such methods infer past population changes using the frequency of coalescent events in a genealogy during past time intervals. As such, a robust tree featuring many individuals, evenly sampled from a population over time and space is needed [35]. While the current study has reinforced the observation that higher-quality data from ancient seeds is best recovered by screening many samples and choosing the most promising, a project of sufficient scale might meet the required number of individuals. If the demographic reconstruction method is properly tested for robustness to missing or inaccurate data, then palaeodemography may even be possible using the distance metrics attainable from scant data, as demonstrated in sections 5.2.3, 5.2.4, 5.2.5, 5.3.1, and 5.3.3.1. The broad range of dates attributed to the samples (see section 4.2.1) should make appropriate sampling for such a study possible, especially given the presence of taxa in species such as *Draba*, *Bistorta*, and *Picea* in permafrost sediments from many locations [66, 68–70].

Furthermore, whole ancient chloroplast sequence data invite longitudinal studies on plastid genome evolution and gene function over time. A possible candidate for such study, for instance, is the RuBisCO gene, which is encoded by the chloroplast, and whose sequence bears a functional relationship to the photosynethetic functions of the plant. These functions are likely to change over time and in the course of migrations, as photoperiods, temperature, moisture, and atmospheric carbon dioxide levels alter [28, 52].

The findings of this chapter reinforce previous interpretations of Arctic plant biology and phytogeography, highlighting the importance of long distance dispersal and survival, and fast morphological development via hybridisation and changes in ploidy. In a conservation context, these findings emphasise the possibility of translocating threatened species and encouraging their adaptation to new environments via hybridisation. However, this depends upon the persistence of enough suitable environments, which is in turn dependent upon the

contraction rate of the high-stress arctic and alpine biomes these plants have specialised in throughout much of the Quaternary.

Chapter 5 Bibliography

[1] R.J. Abbott and C. Brochmann. "History and evolution of the arctic flora: in the footsteps of Eric Hulten". In: *Molecular Ecology* 12.2 (2003), pp. 299–313. ISSN: 1365-294X.

[2] R.J. Abbott and H.P. Comes. "Evolution in the Arctic: a phylogeographic analysis of the circumarctic plant, Saxifraga oppositifolia (Purple saxifrage)". In: *New phytologist* 161.1 (2004), pp. 211–224. ISSN: 1469-8137.

[3] R.J. Abbott et al. "Molecular analysis of plant migration and refugia in the Arctic". In: *Science* 289.5483 (2000), pp. 1343–1346. ISSN: 0036-8075.

[4] Hafdis Hanna Aegisdottir and Dora Ellen Dorhallsdottir. "Theories on migration and history of the North-Atlantic flora: a review". In: *Jokull* 54 (2004), pp. 1–16.

[5] I.G. Alsos et al. "Frequent long-distance plant colonization in the changing Arctic". In: *Science* 316.5831 (2007), pp. 1606–1609. ISSN: 0036-8075.

[6] Nadir Alvarez, Stephanie Manel, and Thomas Schmitt. "Contrasting diffusion of Quaternary gene pools across Europe: The case of the arctic–alpine< i> Gentiana nivalis</i> L.(Gentianaceae)". In: *Flora-Morphology, Distribution, Functional Ecology of Plants* (2012). ISSN: 0367-2530.

[7] Martin R Bauert. "Genetic diversity and ecotypic differentiation in arctic and alpine populations of Polygonum viviparum". In: *Arctic and Alpine research* (1996), pp. 190–195. ISSN: 0004-0851.

[8] Lyman Benson. "A treatise on the North American Ranunculi". In: *American Midland Naturalist* (1948), pp. 1–261. ISSN: 0003-0031.

[9] RY Berg. "Disjunctions in the Norwegian alpine flora and theories proposed for their explanation". In: *Blyttia* 21 (1963), pp. 133–177.

[10] John W Bills et al. "Environmental and genetic correlates of allocation to sexual reproduction in the circumpolar plant Bistorta vivipara". In: *American journal of botany* 102.7 (2015), pp. 1174–1186. ISSN: 0002-9122.

[11] C. Brochmann et al. "Polyploidy in arctic plants". In: *Biological Journal of the Linnean Society* 82.4 (2004), pp. 521–536. ISSN: 1095-8312.

[12] Daniel F Bronny. "Comparative genomics of Bistorta vivipara". In: (2011).

[13] Anne K Brysting, Cecilie Mathiesen, and Thomas Marcussen. "Challenges in polyploid phylogenetic reconstruction: a case story from the arctic-alpine Cerastium alpinum complex". In: *Taxon* 60.2 (2011), pp. 333–347. ISSN: 0040-0262.

[14] J Castresana. "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis". In: *Molecular biology and evolution* 17.4 (2000), pp. 540–552. ISSN: 0737-4038.

[15] Zhiwen Chen et al. "Chloroplast DNA structural variation, phylogeny, and age of divergence among diploid cotton species". In: *PloS one* 11.6 (2016), e0157183.

[16] William J Cody. *Flora of the Yukon territory*. NRC Research Press, 2000. ISBN: 066018110X.

[17] Luca Comai. "The advantages and disadvantages of being polyploid". In: *Nature Reviews Genetics* 6.11 (2005), pp. 836–846. ISSN: 1471-0056.

[18] Flora of North America Editorial Committee. *Flora of North America*. Oxford University Press on Demand, 1993. ISBN: 0195152077.

[19] Eilif Dahl. "On different types of unglaciated areas during the ice ages and their significance to phytogeography". In: *New Phytologist* 45.2 (1946), pp. 225–242. ISSN: 1469-8137.

[20] C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. 1st ed. London: John Murray, 1859.

[21] PAMELA K DIGGLE et al. "Barriers to sexual reproduction in Polygonum viviparum: a comparative developmental analysis of P. viviparum and P. bistortoides". In: *Annals of botany* 89.2 (2002), pp. 145–156. ISSN: 0305-7364.

[22] Alexei J Drummond et al. "Bayesian coalescent inference of past population dynamics from molecular sequences". In: *Molecular biology and evolution* 22.5 (2005), pp. 1185–1192.

[23] Eric Y Durand et al. "Testing for ancient admixture between closely related populations". In: *Molecular biology and evolution* 28.8 (2011), pp. 2239–2252. ISSN: 0737-4038.

[24] J. Ehlers and P.L. Gibbard. "Quaternary Glaciations: Overview". In: *Encyclopedia of Quaternary science*. Elsevier, 2006, pp. 1023–1031. ISBN: 0080547826.

[25] Dorothee Ehrich et al. "Genetic consequences of Pleistocene range shifts: contrast between the Arctic, the Alps and the East African mountains". In: *Molecular Ecology* 16.12 (2007), pp. 2542–2559. ISSN: 1365-294X.

[26] Pernille Bronken Eidesen et al. "Repeatedly out of Beringia: Cassiope tetragona embraces the Arctic". In: *Journal of Biogeography* 34.9 (2007), pp. 1559–1574. ISSN: 1365-2699.

[27] Reidar Elven and Ihsan A Al-Shehbaz. "Draba simmonsii (Brassicaceae), a new species of the D. micropetala complex from the Canadian Arctic Archipelago". In: *Novon: A Journal for Botanical Nomenclature* 18.3 (2008), pp. 325–329. ISSN: 1055-3177.

[28] Jeroni Galmes et al. "Environmentally driven evolution of Rubisco and improved photosynthesis and growth within the C3 genus Limonium (Plumbaginaceae)". In: *New Phytologist* 203.3 (2014), pp. 989–999.

[29] Emmanuel Gardiner. "Comparison of phenotypic plasticity in Bistorta vivipara in topographically rough and flat landscapes". In: (2013).

[30] Bas van Geel et al. "The ecological implications of a Yakutian mammoth's last meal". In: *Quaternary Research* 69.3 (2008), pp. 361–376. ISSN: 0033-5894.

[31] Vadim V Goremykin et al. "Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer". In: *Molecular Biology and Evolution* 26.1 (2009), pp. 99–110. ISSN: 0737-4038.

[32] NE Grulke and LC Bliss. "A note on winter seed rain in the High Arctic". In: *Arctic and alpine research* (1983), pp. 261–265. ISSN: 0004-0851.

[33] H.H. Grundt et al. "High biological species diversity in the arctic flora". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.4 (2006), pp. 972–975. ISSN: 0027-8424.

[34] Christoph Hahn, Lutz Bachmann, and Bastien Chevreux. "Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach". In: *Nucleic acids research* (2013), gkt371. ISSN: 0305-1048.

[35] Rasmus Heller, Lounes Chikhi, and Hans Redlef Siegismund. "The confounding effect of population structure on Bayesian skyline plot inferences of demographic history". In: *PLoS One* 8.5 (2013), e62992.

[36] Matthias H Hoffmann et al. "Sources of the arctic flora: origins of arctic species in Ranunculus and related genera". In: *International journal of plant sciences* 171.1 (2010), p. 90.

[37] Elvira Horandl et al. "Phylogenetic relationships and evolutionary traits in< i> Ranunculus</i> sl (Ranunculaceae) inferred from ITS sequence analysis". In: *Molecular phylogenetics and evolution* 36.2 (2005), pp. 305–327. ISSN: 1055-7903.

[38] Eric Hulten. "Outline of the history of arctic and boreal biota during the Quaternary period". In: (1972).

[39] Hajime Ikeda et al. "Pleistocene climatic oscillations and the speciation history of an alpine endemic and a widespread arctic-alpine plant". In: *New phytologist* 194.2 (2012), pp. 583–594. ISSN: 1469-8137.

[40] Ingrid Jordon-Thaden. "Species and genetic diversity of Draba: phylogeny and phylogeography". In: (2009).

[41] Ingrid Jordon-Thaden et al. "Molecular phylogeny and systematics of the genus< i> Draba</i>(Brassicaceae) and identification of its most closely related genera". In: *Molecular Phylogenetics and Evolution* 55.2 (2010), pp. 524–540. ISSN: 1055-7903.

[42] Maxim V Kapralov et al. "Genetic enrichment of the arctic clonal plant Saxifraga cernua at its southern periphery via the alpine sexual Saxifraga sibirica". In: *Molecular Ecology* 15.11 (2006), pp. 3401–3411. ISSN: 1365-294X.

[43] Logan Kistler et al. "Transoceanic drift and the domestication of African bottle gourds in the Americas". In: *Proceedings of the National Academy of Sciences* 111.8 (2014), pp. 2937–2941. ISSN: 0027-8424.

[44] Marcus Koch and Ihsan A Al-Shehbaz. "Molecular data indicate complex intra-and intercontinental differentiation of American Draba (Brassicaceae)". In: *Annals of the Missouri Botanical Garden* (2002), pp. 88–109. ISSN: 0026-6493.

[45] askell Love and Doris Love. "Cytotaxonomical atlas of the arctic flora". In: *Vaduz: J. Cramer xxiii, 598p.-Map, chrom. nos.. Maps, Chromosome numbers. Geog* 1.2 (1975).

[46] K Marhold and J Lihova. "Polyploidy, hybridization and reticulate evolution: lessons from the Brassicaceae". In: *Plant Systematics and Evolution* 259.2-4 (2006), pp. 143–174. ISSN: 0378-2697.

[47] Kendrick L. Marr, Geraldine A. Allen, and Richard J. Hebda. "Refugia in the Cordilleran ice sheet of western North America: chloroplast DNA diversity in the Arctic–alpine plant Oxyria digyna". In: *Journal of Biogeography* 35.7 (2008), pp. 1323–1334. ISSN: 1365-2699. DOI: 10.1111/j.1365-2699.2007.01879.x. URL: http://dx.doi.org/10.1111/j.1365-2699.2007.01879.x.

[48] Kendrick L Marr et al. "Phylogeographical patterns in the widespread arctic–alpine plant Bistorta vivipara (Polygonaceae) with emphasis on western North America". In: *Journal of Biogeography* (2012). ISSN: 1365-2699.

[49] Michael R McKain et al. "Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes". In: *Bioinformatics* 33.1 (2016), pp. 130–132.

[50] Polina Yu Novikova et al. "Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism". In: *Nature genetics* 48.9 (2016), pp. 1077–1082.

[51] C Oliver, PM Hollingsworth, and RJ Gornall. "Chloroplast DNA phylogeography of the arctic-montane species Saxifraga hirculus (Saxifragaceae)". In: *Heredity* 96.3 (2006), pp. 222–231. ISSN: 0018-067X.

[52] Douglas Orr et al. "Surveying Rubisco diversity and temperature response to improve crop photosynthetic efficiency". In: *Plant physiology* (2016), pp–00750.

[53] Jim Pojar and A. MacKinnon. *Alpine plants of British Columbia, Alberta, and northwest North America*. Lone Pine Publishing, 2013. ISBN: 1551058863.

[54] Christoph Reisch. "Glacial history of Saxifraga paniculata (Saxifragaceae): molecular biogeography of a disjunct arctic-alpine species from Europe and North America". In: *Biological Journal of the Linnean Society* 93.2 (2008), pp. 385–398. ISSN: 1095-8312.

[55] Adrien Rieux and François Balloux. "Inferences from tip-calibrated phylogenies: a review and a practical guide". In: *Molecular ecology* 25.9 (2016), pp. 1911–1924.

[56] Roswitha Schmickl et al. "Phylogeographic implications for the North American boreal-arctic Arabidopsis lyrata complex". In: *Plant Ecology and Diversity* 1.2 (2008), pp. 245–254. ISSN: 1755-0874.

[57] Peter Schonswetter, Reidar Elven, and Christian Brochmann. "Trans-Atlantic dispersal and large-scale lack of genetic structure in the circumpolar, arctic-alpine sedge Carex bigelowii sl (Cyperaceae)". In: *American Journal of Botany* 95.8 (2008), pp. 1006–1014. ISSN: 0002-9122.

[58] Fabian Sievers et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular systems biology* 7.1 (2011).

[59] I Skrede, L Borgen, and C Brochmann. "Genetic structuring in three closely related circumpolar plant species: AFLP versus microsatellite markers and high-arctic versus arctic–alpine distributions". In: *Heredity* 102.3 (2008), pp. 293–302. ISSN: 0018-067X.

[60] Inger Skrede et al. "Refugia, differentiation and postglacial migration in arctic-alpine Eurasia, exemplified by the mountain avens (Dryas octopetala L.)" In: *Molecular Ecology* 15.7 (2006), pp. 1827–1840. ISSN: 1365-294X. DOI: 10.1111/j.1365-294X.2006.02908.x. URL: http://dx.doi.org/10.1111/j.1365-294X.2006.02908.x.

[61] D.E. Soltis and P.S. Soltis. "Polyploidy: recurrent formation and genome evolution". In: *Trends in Ecology and Evolution* 14.9 (1999), pp. 348–352. ISSN: 0169-5347.

[62] O Totland and J Nylehn. "Assessment of the effects of environmental change on the performance and density of Bistorta vivipara: the use of multivariate analysis and

experimental manipulation". In: *Journal of Ecology* 86.6 (1998), pp. 989–998. ISSN: 1365-2745.

[63] Leigh Van Valen. "A new evolutionary law". In: *Evolutionary theory* 1 (1973), pp. 1–30.

[64] M.D. Walker, F.J.A. Daniels, and E. van der Maarel. "Circumpolar arctic vegetation". In: *Journal of Vegetation Science* 5.6 (1994), pp. 757–764.

[65] Kristine B Westergaard et al. "Glacial survival may matter after all: nunatak signatures in the rare European populations of two west-arctic species". In: *Molecular Ecology* 20.2 (2011), pp. 376–393. ISSN: 1365-294X.

[66] Sebastian Wetterich et al. "Palaeoenvironmental dynamics inferred from late Quaternary permafrost deposits on Kurungnakh Island, Lena Delta, Northeast Siberia, Russia". In: *Quaternary Science Reviews* 27.15–16 (2008), pp. 1523–1540. ISSN: 0277-3791. DOI: http://dx.doi.org/10.1016/j.quascirev.2008.04.007. URL: http://www.sciencedirect.com/science/article/pii/S0277379108001054.

[67] Anna V Williams et al. "Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia". In: *Molecular phylogenetics and evolution* 96 (2016), pp. 1–8.

[68] Matthew J Wooller et al. "The detailed palaeoecology of a mid-Wisconsinan interstadial (ca. 32 000 14C a BP) vegetation surface from interior Alaska". In: *Journal of Quaternary Science* 26.7 (2011), pp. 746–756. ISSN: 1099-1417.

[69] GD Zazula et al. "New spruce (Picea spp.) macrofossils from Yukon Territory: implications for late Pleistocene refugia in eastern Beringia". In: *Arctic* (2006), pp. 391–400. ISSN: 0004-0843.

[70] Grant D Zazula et al. "Early Wisconsinan (MIS 4) Arctic ground squirrel middens and a squirrel-eye-view of the mammoth-steppe". In: *Quaternary Science Reviews* 30.17 (2011), pp. 2220–2237. ISSN: 0277-3791.

Chapter 6

SimWreck

Simulating Ancient DNA

This chapter describes the program SimWreck and its application to the analyses described in the previous chapter. Chapter 5 section 5.2.5 provides essential background to the case study presented in section 6.4 of this chapter. The code is available in DS:Code:2.1 and online at www.github.com/mtrw/simwreck

---

6.1 Introduction

DNA damage can bias analyses and negatively impact the quality and yield of sequence data. When reconstructing an ancient sequence by mapping to a reference genome, short reads or reads with damage-induced mismatches are mapped with lower confidence. This may result in reduced coverage, and impair the discovery of informative variants [9, 12, 17]. In phylogenetic and demographic studies, damage may alter the lengths of branches and the placement of nodes in a phylogeny, or bias the timing of reconstructed demographic events [15–17]. Metagenomic studies of ancient DNA are similarly affected, being susceptible to loss of resolution and false positive results [4, 10]. Standard aDNA analysis pipelines account for the influence of DNA damage in several ways, for instance by characterising the degree of damage and modifying various quality scores accordingly [6, 8]. Nevertheless, aDNA studies require extensive verification that their results are not biased in any way. Simulations have proven invaluable in exploring important consequences of DNA damage [12, 15], yet the applicability of these results to real datasets is limited by differences between the simulated datasets and real ancient DNA samples. The distribution of read lengths differs significantly among ancient samples, and may also display various irregularities, multimodality, or periodicity [13]. Deamination and depurination also occur to differing degrees[2, 18]. In addition, novel analysis pipelines are regularly conceived and applied without formal or rigorous testing for damage-induced biases. As a result there is a

need for tools allowing the generation of simulated ancient sequence data, giving the user a fine-scale control over the damage profile and read length distribution.

SimWreck is a Perl script implementing a novel algorithm that aims to deliver these qualities. By default, SimWreck produces short pseudo-degraded reads based upon a given reference sequence. This mode is useful for investigating mapping and variant recovery, or creating simulated metagenomic datasets. In *Add Damage* mode, SimWreck applies deamination-like damage to user-provided reads. This is useful when exploring what effects damage is having upon the analysis of a particular dataset, since adding damage to reads can reveal in what ways and to what degree results are being skewed by deamination. In Uniform Damage mode, the program randomly makes deamination-type changes to the given sequences, changing Cs to Ts and Gs to As with a fixed probability. The program also features a convenient *Plot* function to aid in parameter selection by visualising requested read length distributions (FIG. 6.2).

6.2 Methods

6.2.1 Fragment Length and Depurination

The SimWreck algorithm allows users to simulate datasets with a broad range of bell-shaped length distributions. The beta distribution proves ideal for this purpose, as it can made to take many different bell-like shapes, as well as exponential and uniform shapes, with the adjustment of only two parameters.

The user may provide the shape and scale parameters of the distribution, and the minimum and maximum sequence lengths corresponding to the range $[0, 1]$ over which the distribution is defined. The challenge faced by the algorithm is to produce reads—effectively substrings of the reference sequence—whose lengths have not only the desired length distribution, but which also show the correct tendency to start after purines or end before pyrimidines, without introducing any unwarranted biases in coverage.

This proves a challenge. While read lengths can be randomly drawn from an appropriate distribution to simulate fragmentation, and while one end can be chosen to favour beginning after particular bases, the read length and the position of one end necessarily defines the position of the other end—and hence will not account for biases in end position on this second end. Two approaches were tried to meet this requirement. The first, which was eventually rejected being 10–1000 times slower in benchmark tests, involved choosing the starting point for each read, then characterising a probability distribution over all the possible endpoints. This endpoint distribution would be constructed newly for each read start point and would represent the relative probabilities of ending at each possible end point, accounting for both the shape of the desired length distribution and the identity of each base. As figure 6.1 shows, this results in an irregular distribution that had to be characterised and drawn from randomly for every read generated—and thus involves many operations.

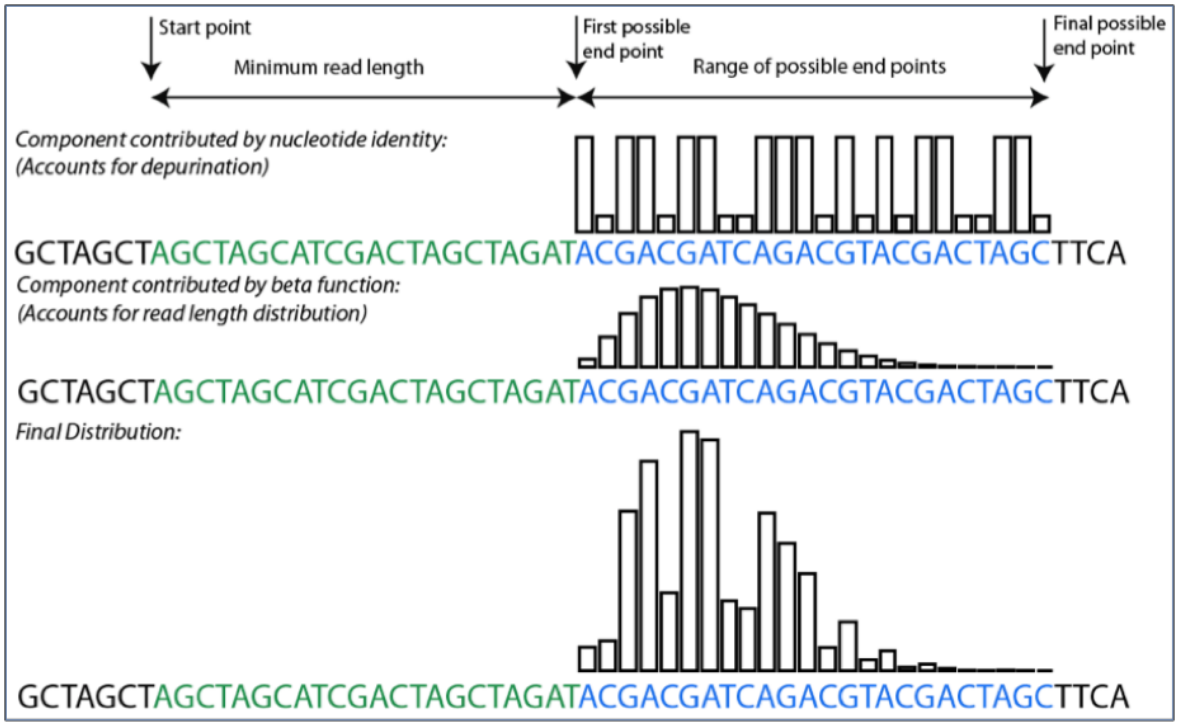

FIGURE 6.1: How the end points of reads are chosen using the (now depreciated) distribution method. The height of the bars at each nucleotide within the range of possible end points is proportional to the probability of the read ending immediately before the nucleotide.

The method SimWreck currently uses is perhaps less elegant, but far more efficient. Read start points and lengths are drawn randomly, and reads are then rejected or accepted based upon their endpoints with a probability calculated to produce the desired tendency. The parameters were designed to allow users to enter arguments that can be straightforwardly understood from a MapDamage-type profile (for an example see chapter 4, figure 4.2). In the case of depurination, the user is asked to input the approximate difference between the mean frequency of purines/pyrimidines in a sequence, and the frequency at the first position outside the reads (see argument -p in figure 2, panel D). A simple calculation produces the probability with which a read should be accepted should its end fall beside a purine (if it is the 3′ end) or a pyrimidine (5′ end). To illustrate, consider the 3′ ends of reads ending at either a purine or a pyrimidine (i.e. not N). Let the proportions of pyrimidines and purines in the reference sequence be denoted Y and U respectively. Assuming the sequence represents a random sequence of pyrimidines and purines in these proportions, then if read endpoints are placed on the reference sequence randomly, then the probability that the read end will fall at a purine rather than a pyrimidine is given by $\frac{U}{(U+Y)}$. But if reads ending at pyrimidines are only accepted with probability a, then the final proportion of purine-ending reads will be modified to $\frac{U}{(U+aY)}$. The difference between these gives the "boost" B afforded to purines, entered by the user:

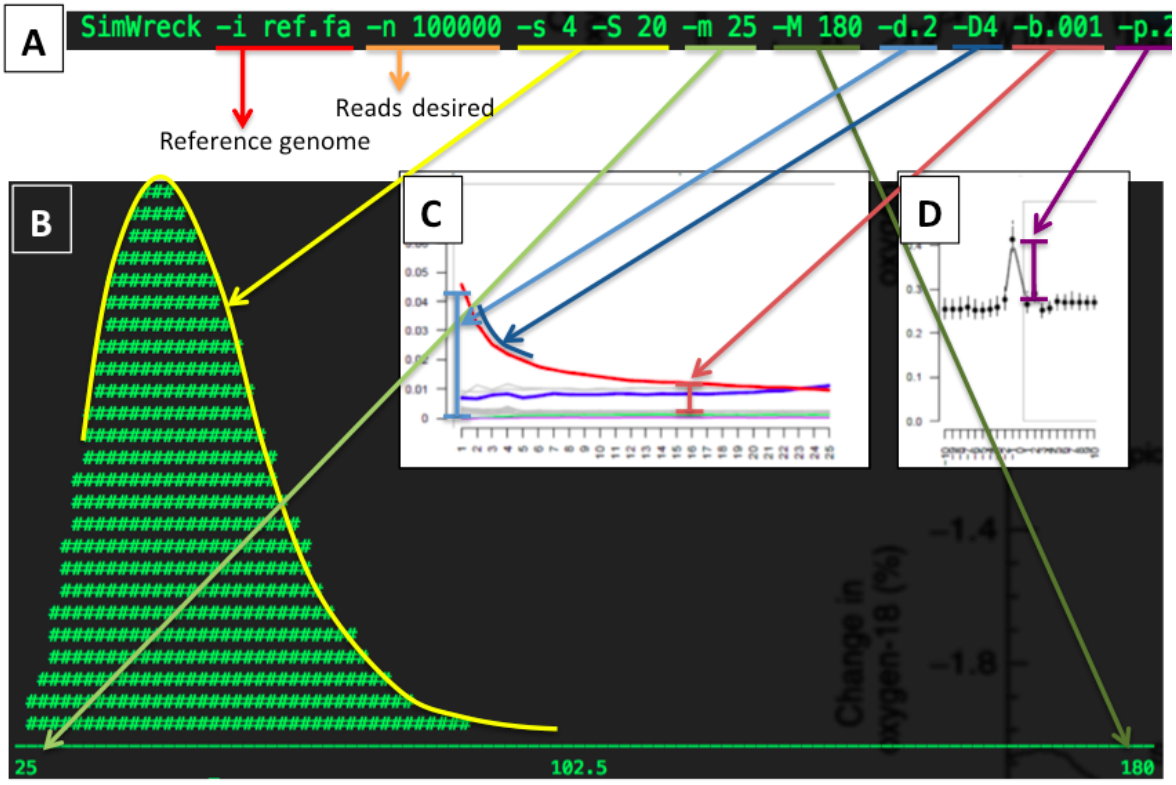

FIGURE 6.2: Running SimWreck. Argument flags in the run command in panel A are shown in relation to the aspects of output they control in panels B (length distribution displayed using SimWreck's Plot function), C (MapDamage substitution frequency plot, see figure 4.2), and D (MapDamage mapped base frequency plot, see figure 4.2). These profiles are taken from an ancient *Bistorta vivipara* sample, methods described in figure 4.2. **Red:** -i sets the user-provided reference sequence. **Orange:** -n sets the number of fragments the run will generate. **Yellow:** -s and -S set the shape and scale parameters that control the shape of the length distribution. Adding the -P flag sets plot mode, allowing the user to visualise the requested distribution, as in panel B. **Pale green:** -m sets the minimum limit of length distribution. **Dark green:** -M sets the maximum limit of length distribution. **Pale Blue:** -d sets the damage "weight". Intuitively represents the probability of a deamination at the terminal base in a fragment. **Dark blue:** -D sets the damage "decay". Controls the shape of the deamination frequency curve. **Salmon:** -b sets the damage "baseline" parameter. **Purple:** -p sets the depurination "boost" parameter.

$$\frac{U}{(U+aY)} - \frac{U}{(U+Y)} = B.$$

We can then solve for a to retrieve acceptance probability of a read that ends at a pyrimidine:

$$a = \frac{(U(B(U+Y)-Y))}{(Y(-B(U+Y)+U))}.$$

Since many sequences contain uncalled 'N' nucleotides, these must come with a probability of rejection, too. Not to do so would effectively give reads ending or beginning with N a

free pass, biasing the coverage towards N-rich regions (or sequence ends, to which Ns are added by the program allowing the creation of reads that partially overlap the ends). Given that the sequence is considered to be a random sequence of pyrimidines and purines, and that the Ns are considered "hidden" bases, they are rejected based on the probability that their "true" identities will merit the rejection test. At the 5′ end, where pyrimidines may be rejected, then the rejection probability for a read ending at N is:

$$\textit{Prob}(\textit{nucleotide is a pyrimidine}) \times \textit{Prob}(\textit{pyrimidine is rejected}) = \frac{aY}{(U+Y)}.$$

At present the program allows one value for B to be entered by the user, which is a compromise for simplicity, since it is known that adenine and guanine have slightly different depurination frequencies [11]. Both the above methods can be simply extended to account for this should it become necessary in the future.

6.2.2 Deamination

In keeping with the goal that users can enter arguments based upon a MapDamage profile, the frequency of deamination-induced C-to-T and G-to-A changes at the ends of sequences is described using an exponential decay curve, which can be fitted to many empirical deamination curves such as that shown in figure 4.2. The curve is defined by the following formula where d, D, and b represent the corresponding argument flags (see explanations in figure 6.2 text), and x represents the distance of a C or G from the 5′ or 3′ end of the fragment respectively:

$$\textit{Prob}(\textit{deamination-induced change}) = (d-b)e^{(-Dx)} + b.$$

Deamination events cannot normally cause substitutions to occur upstream (5′) of one another on opposing strands. Since 3′ overhangs are not retained, the above scenario would necessitate that both deamination events occurred in a double-stranded region. In order to copy the new uracils, nicks must occur upstream of each damaged site, on the opposing strands (see chapter 4). However, this places the nicks *downstream* of one another, and repairing these nicks would cause the fragment to break in two as the strand-displacing polymerase synthesising from one nick encounters a break in the template. Such fragments cannot be sequenced. SimWreck therefore discards fragments that break this rule. This causes a gradual decrease in deamination events below the "baseline" level (see figuure 6.2 text) moving further into the fragments (figure 6.4, panel C, red line), which can also be seen in empirical damage profiles (figure 4.2, right-hand panels).

6.3 Results

6.3.1 Using SimWreck

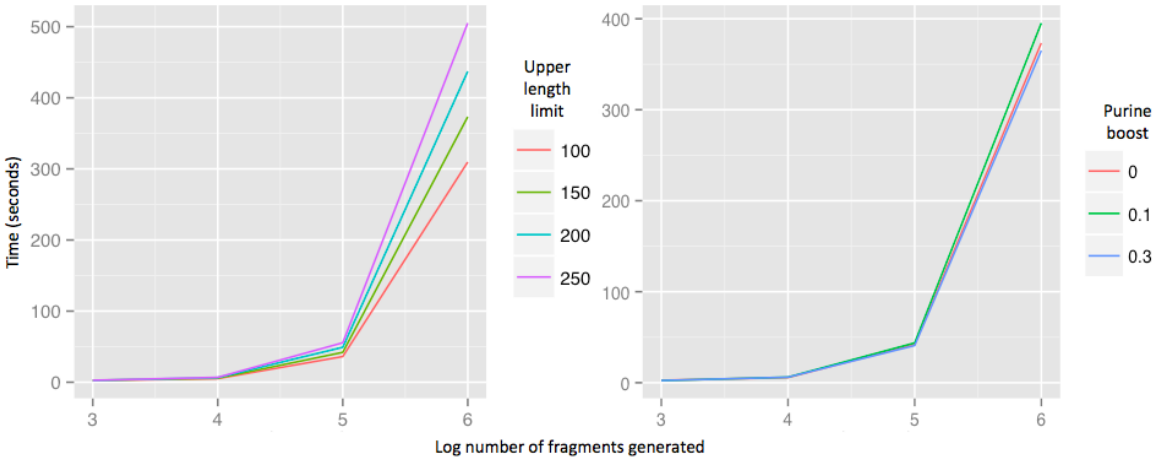

FIGURE 6.3: Benchmarking for SimWreck run on a MacBook Air, 1.7 GHz Intel Core i5, 4 GB 1333 MHz DDR3, Intel HD Graphics 3000 384 MB, using chromosome 1 of human genome hg19 as a reference sequence. **Left:** User runtime as the length distribution is changed to produce longer reads by increasing the arguments given via the -m flag. **Right:** User runtime as the depurination boost argument given via -p is increased. All other parameters left as default.

The SimWreck methods were implemented in Perl v5.18. The user interface make for easy parameter selection, as demonstrated in figure 6.2, which illustrates the process of setting parameters for a SimWreck run based upon the output of a MapDamage profile (panels C and D) and the *Plot* function (panel D). The speed of a SimWreck run (FIG. 6.3) scales with the number of reads required, and the average length of the fragments being simulated. The rejection of reads to achieve simulated depurination effects has a negligible impact on the user run time.

The program proves capable of producing simulated reads that, when mapped to an appropriate reference genome, produce mapDamage profiles that convincingly replicate the deamination, depurination, and fragmentation effects present in real aDNA damage profiles. Such a profile is shown in figure 6.4.

Empirical read length distributions do exhibit occasional irregularities that are not reproducible by SimWreck. In some cases it is possible to approximate an unusual length distribution very closely by combining reads generated from multiple SimWreck runs. The characteristics of HTS reads may in reality be influenced by nucleotide composition [3], epigenetic modifications [13], sequence motifs [19], the interaction of DNA with itself and other molecules such as histones [13], the behaviour of different laboratory enzymes [3], or the influence of target enrichment procedures [1]. However, many of these influences are poorly

characterised to the extent that explicitly modelling their effects on the signatures of DNA degradation may be premature at this stage.

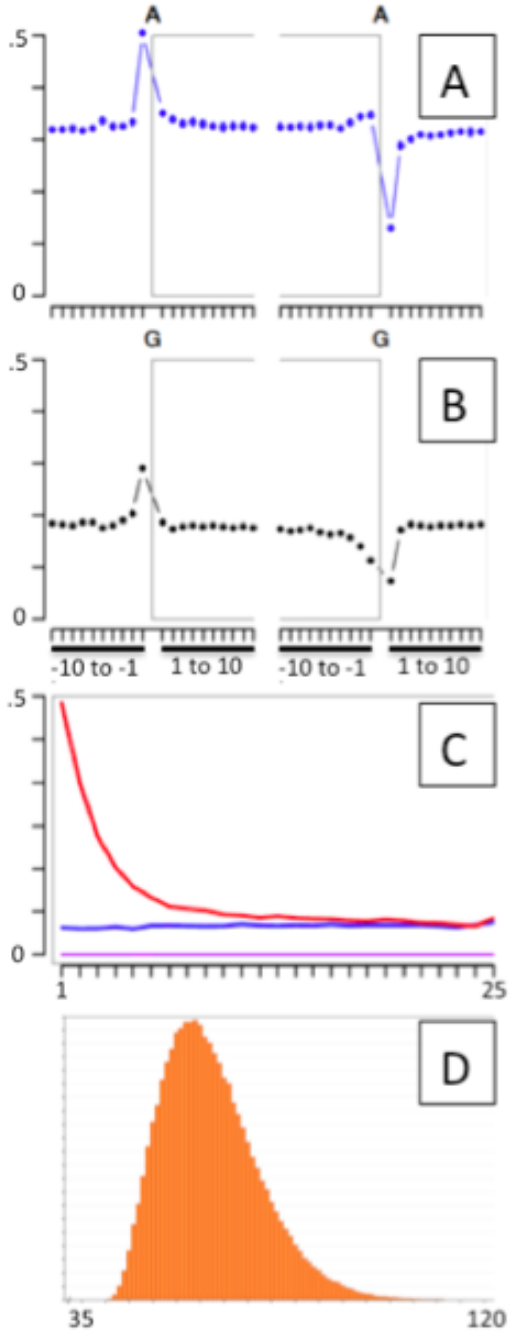

FIGURE 6.4: MapDamage profile of SimWreck-generated reads using human chromosome 1 from hg19 as a reference (generated using default parameters except –d.5, -p.15, -b.1). Damage profiles are explained in detail in sections 4.1, 4.2.3.3. and 4.3.2.1. **A and B:** Mapped base frequency plots for A and G, respectively. **C:** Substitution frequency plot for the 5′ ends of fragments. **D:** Length distribution of fragments shown as a frequency histogram (generated using Geneious 9[7]).

6.4 Case Study: Verification of Ancient Plant Phylogenetic Methods (Chapter 5)

The data used in chapter 5 originate from permafrost sediments dating in some cases to >50 ka., and from herbarium specimens, which have in recent studies been shown to accumulate typical DNA degradation signatures over time [20]. I took this opportunity to test the possible effects of damage on the analysis pipeline used, by imposing damage artificially upon the alignments used, then running the analysis pipeline described in that chapter, and investigating the effects of the damage upon the results. Damage was imposed with the command available in DS:Code:1.5.7.

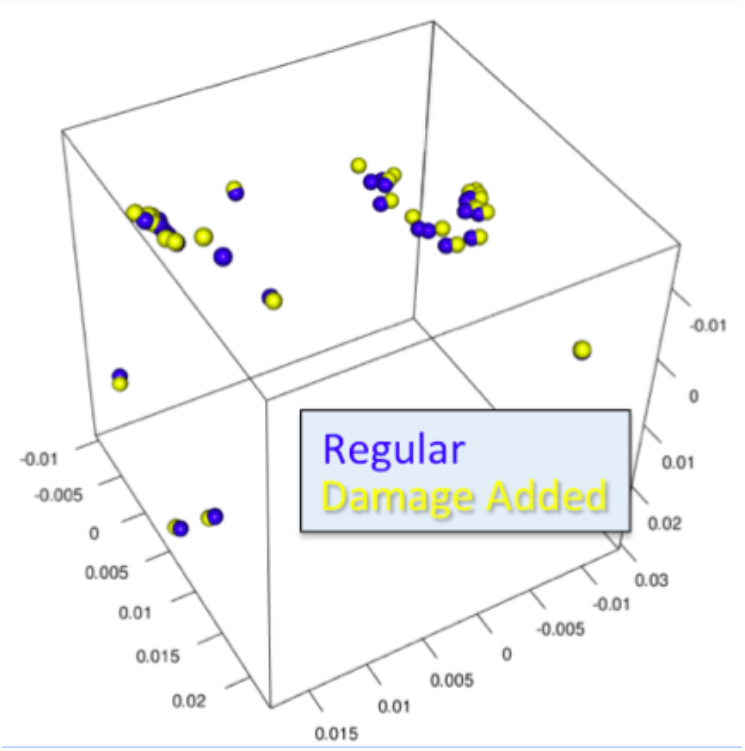

FIGURE 6.5: Exploring the influence of damage on the PCAs generated in chapter 5. Detailed explanations see section 6.4

The influence of low sequence completeness was also investigated as an alternative hypothesis for the factors driving uncertainty in the results. Grouping of closely-related species was displayed using a PCA, which is constructed using a distance matrix (Fig. 6.6). The PCA is therefore expected to retain its groupings if the relative distances between sequences are preserved even after damage is applied. This assumption was tested by comparing the distance between two sequences before and after damage was applied to them. In both cases, the full analysis pipeline was run, which aims to remove the influence of damage. Figure 6.6 strongly suggests the pipeline is apt to achieve this. Each point on the figure represents the distance between a pair of sequences, measured before (x-axis) and after (y-axis) damage was applied. The strong linear relationship is evidence that the relative differences are

indeed preserved. Points falling most distant from the central trend are almost invariably those where one of the samples had a very incomplete sequence, and it was in fact this observation that led to completeness being represented visually on the PCAs in chapter 5 as a proxy for uncertainty, and to the imposition of a lower cutoff for completeness.

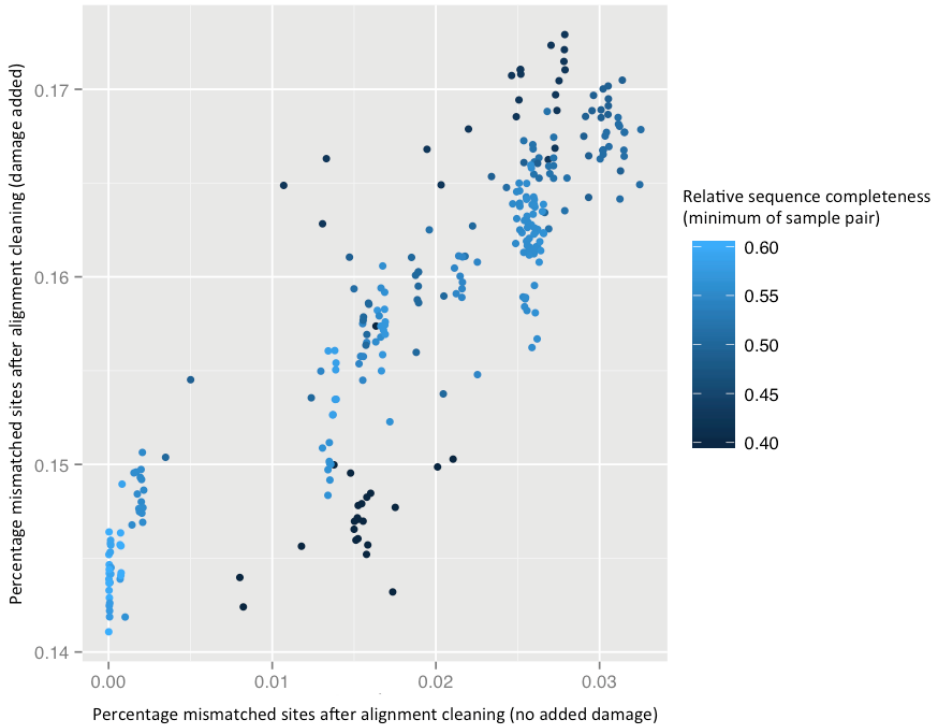

FIGURE 6.6: Pairwise distances for damaged and undamaged sequences. Sequence completeness is calculated as the proportion of A/G/C/T positions in the aligned sequence (c.f. uncalled bases, gaps, or ambiguous calls). Further explanations see text.

Further investigation aimed to help confirm that the influence of damage would not disrupt the distance matrices enough to invalidate the central findings of the PCA. To this end, the "cleaned" (see chapter 5) damaged and undamaged sequences were combined into a single distance alignment, the distance matrices recalculated, and the PCA redrawn with damaged and undamaged samples separated by colour. The "unclumped" datapoints in figure 6.5 clearly show that, even with high levels of damage, undamaged (blue) samples pair together with their damaged (yellow) counterparts, and that the separation between them is trivial compared to the broad separation between groups, increasing our confidence in the grouping of samples in these analyses.

6.5 Closing Remarks

The release version of SimWreck combines several new innovations in the simulation of ancient DNA—notably the use of a parameterised user-defined length distribution and the simulation of depurination via a probabilistic rejection scheme—into a user-friendly package. SimWreck's primary use will be to characterise the effects of DNA damage upon novel analysis methods. This is especially important for any analysis that relies upon individual read sequences, rather than consensus of many reads. Ancient metagenomics [14] is one example, as deamination is expected to artificially inflate the diversity of sequences in a sample. Another second example is the *de novo* assembly of genomes using degraded DNA [5], where deamination may be expected, under some circumstances, to create novel k-mers that introduce bubbles and short terminal paths into the De Bruijn graph. SimWreck's *AddDamage* and *UniformDamage* modes will be useful to rigorously quantify the actual relationship between certain study outcomes and the amount of deamination damage in the data: If adding progressively more damage to the data has minimal effect upon analysis outcomes, then this provides good evidence that whatever damage already exists in the data is also having little effect. This logic underlies the application of SimWreck to the analyses of chapter 5, described in section 6.4 above.

Future versions of the program will be implemented in a compiled language for greater speed. Features that may be added include the simulation of sequence damage and quality scores, the use of empirically-derived sequence properties that automatically mimic given libraries, the capacity to output different types of data such as simulated unmerged reads, and the simulated addition of contaminant DNA. Incorporating certain mapping or composition biases in the fragments produced may also be possible in future releases.

Chapter 6 Bibliography

[1] Maria C Avila-Arcos et al. "Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA". In: *Scientific reports* 1 (2011).

[2] Adrian W Briggs et al. "Patterns of damage in genomic DNA sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences* 104.37 (2007), pp. 14616–14621. ISSN: 0027-8424.

[3] J. Dabney and M. Meyer. "Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries". In: *Biotechniques* 52.2 (2012), pp. 87–94. ISSN: 1940-9818. DOI: 10.2144/000113809. URL: http://www.ncbi.nlm.nih.gov/pubmed/22313406.

[4] Gentile F Ficetola et al. "Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data". In: *Molecular ecology resources* 15.3 (2015), pp. 543–556. ISSN: 1755-0998.

[5] Chih-Ming Hung et al. "The de novo assembly of mitochondrial genomes of the extinct passenger pigeon (Ectopistes migratorius) with next generation sequencing". In: *PloS one* 8.2 (2013), e56301. ISSN: 1932-6203.

[6] Hakon Jonsson et al. "mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters". In: *Bioinformatics* (2013), btt193. ISSN: 1367-4803.

[7] Matthew Kearse et al. "Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data". In: *Bioinformatics* 28.12 (2012), pp. 1647–1649. ISSN: 1367-4803.

[8] Heng Li, Jue Ruan, and Richard Durbin. "Mapping short DNA sequencing reads and calling variants using mapping quality scores". In: *Genome research* 18.11 (2008), pp. 1851–1858. ISSN: 1088-9051.

[9] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 1367-4803.

[10] Daithi C Murray et al. "Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages". In: *Scientific reports* 3 (2013).

[11] Søren Overballe-Petersen, Ludovic Orlando, and Eske Willerslev. "Next-generation sequencing offers new insights into DNA degradation". In: *Trends in biotechnology* 30.7 (2012), pp. 364–368.

[12] Matthew Parks and David Lambert. "Impacts of low coverage depths and post-mortem DNA damage on variant calling: a simulation study". In: *BMC GENOMICS* 16 (2015). ISSN: 1471-2164.

[13] Jakob Skou Pedersen et al. "Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome". In: *Genome research* 24.3 (2014), pp. 454–466. ISSN: 1088-9051.

[14] Mikkel Winther Pedersen et al. "Ancient and modern environmental DNA". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370.1660 (2015), p. 20130383. ISSN: 0962-8436.

[15] Andrew Rambaut et al. “Accommodating the effect of ancient DNA damage on inferences of demographic histories”. In: *Molecular Biology and Evolution* 26.2 (2009), pp. 245–248. ISSN: 0737-4038.

[16] Andrew Rambaut et al. “Accommodating the effect of ancient DNA damage on inferences of demographic histories”. In: *Molecular Biology and Evolution* 26.2 (2009), pp. 245–248. ISSN: 0737-4038.

[17] Gabriel Renaud et al. “Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA”. In: *Genome biology* 16.1 (2015), p. 1. ISSN: 1474-760X.

[18] Martijn Staats et al. “DNA damage in plant herbarium tissue”. In: *PLoS One* 6.12 (2011), e28448. ISSN: 1932-6203.

[19] A. Valouev et al. “Determinants of nucleosome organization in primary human cells”. In: *Nature* 474.7352 (2011), pp. 516–20. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking). DOI: 10.1038/nature10002. URL: http://www.ncbi.nlm.nih.gov/pubmed/21602827.

[20] Clemens L Weiss et al. “Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens”. In: *bioRxiv* (2015), p. 023135.

Chapter 7

Conclusions

7.1 Innovations and Future Directions

A common theme throughout this work has been the development of new methods for exploring past climate. The first of these presented is in chapter 2: A method for inferring landscape moisture levels over time using nitrogen stable isotopes. This method lays out a framework that could equally be applied to other proxies that have similar behaviour to the isotopic data used, perhaps tree ring counts as a proxy for plant growth rates, stomatal density values as a proxy for atmospheric CO_2 levels, or various biometric data such as the dimensions of fossilised organisms. Several future possible future directions exist. The strength of the claims in the manuscript would be strengthened by a close examination of the dataset, with respect to possible technical biases introduced, for instance, by contamination. Some underlying assumptions could also be made more robust by investigating more thoroughly and formally the δ^{15}N–moisture relationship using data from modern systems. The RMAC algorithm could be optimised and written as an R package, and perhaps reparameterised for more intuitive use.

The methods developed for working with DNA from permafrost-preserved ancient fruits resulted in several useful methodological recommendations for future work on such material, and the development of a new alignment cleaning script proved effective in reconstructing robustness chloroplast-based phylogenies that included ancient samples. This robustness was confirmed using the new tool SimWreck, which will help to give ancient DNA researchers the ability to easily test the robustness of new analysis methods as they are developed in the future. The plant DNA work achieved the near-complete sequencing of the two oldest known draft chloroplast genomes to date, and future efforts using only data already generated in the project may be able to fully assemble these chloroplasts by applying more sophisticated assembly methods coupled with close manual manipulation.

7.2 Climate Adaptability And The Glacial Rangeland Biota

The methods developed are appied to investigations of the effects of past climate change on organisms in the northern holarctic, Europe, and Patagonia. In general terms, the type of climate shifts investigated involve a shift from drier/colder to warmer/wetter conditions. Of these regions, those whose near-term future climate prospects can be predicted with some confidence (western Europe, northern USA, and western Patagonia) are all expected to shift to towards hotter ands drier conditions, at least at the level of the soil [2] (figure 7.2), though in many areas the balance appears precarious, depending largely upon the balance of increased precipitation and counteracted by increased evaporation and runoff.

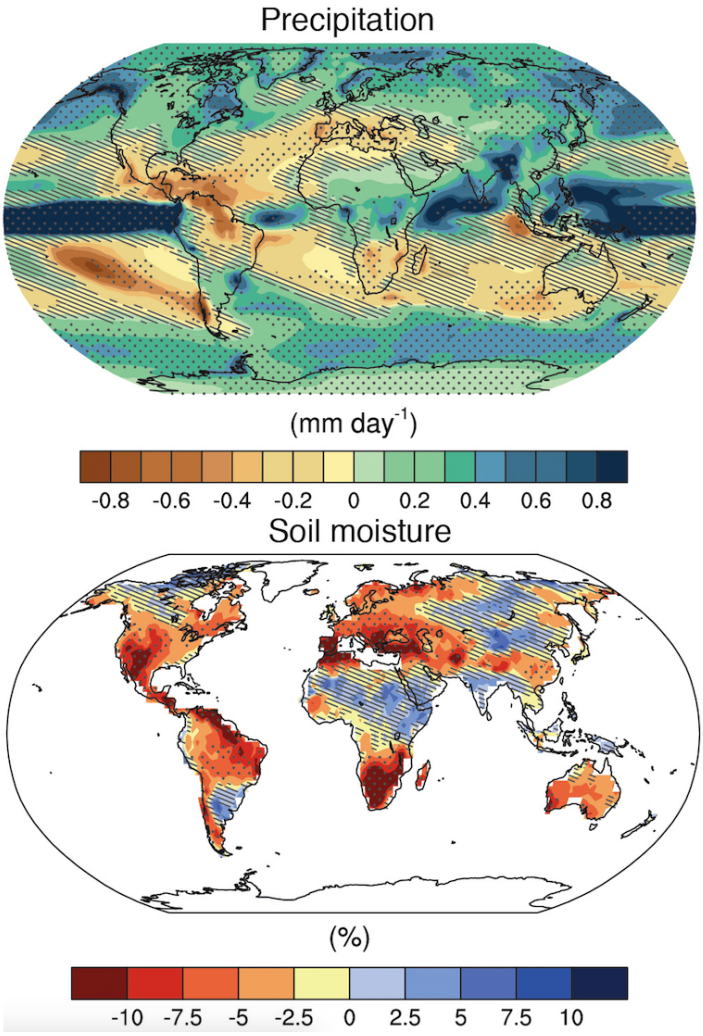

FIGURE 7.1: Projected global changes in precipitation (upper) and soil moisture (lower) between the period 1986–2005 and 2016–2035 (multi-model annual means from CMIP5 under scenario RCP4.5). Hatching indicate regions where the projected change is not significantly different from the inter-model variation (i.e. less than one standard deviation), and stippling indicates regions where the projected change is highly significant (i.e. more than two standard deviations greater than the inter-model variation).

The results support the view that the cryoxeric-adapted mammoth steppe flora are equipped to adapt efficiently to a range of arctic-alpine environments (see chapter 5), however this range is not broad enough—nor the plants' adaptive capability great enough—to allow these species to remain competitive in the face of climatic shifts that favour warmer-adapted organisms from the south. As the IPCC models confirm (figure 7.2), the post-LGM northward mass migration of mammoth steppe relicts noted in chapter 5 will continue, though at a greatly accelerated rate, and longitudinal studies on arctic-alpine boreal plant species' ranges are already beginning to reflect this development [3, 5].

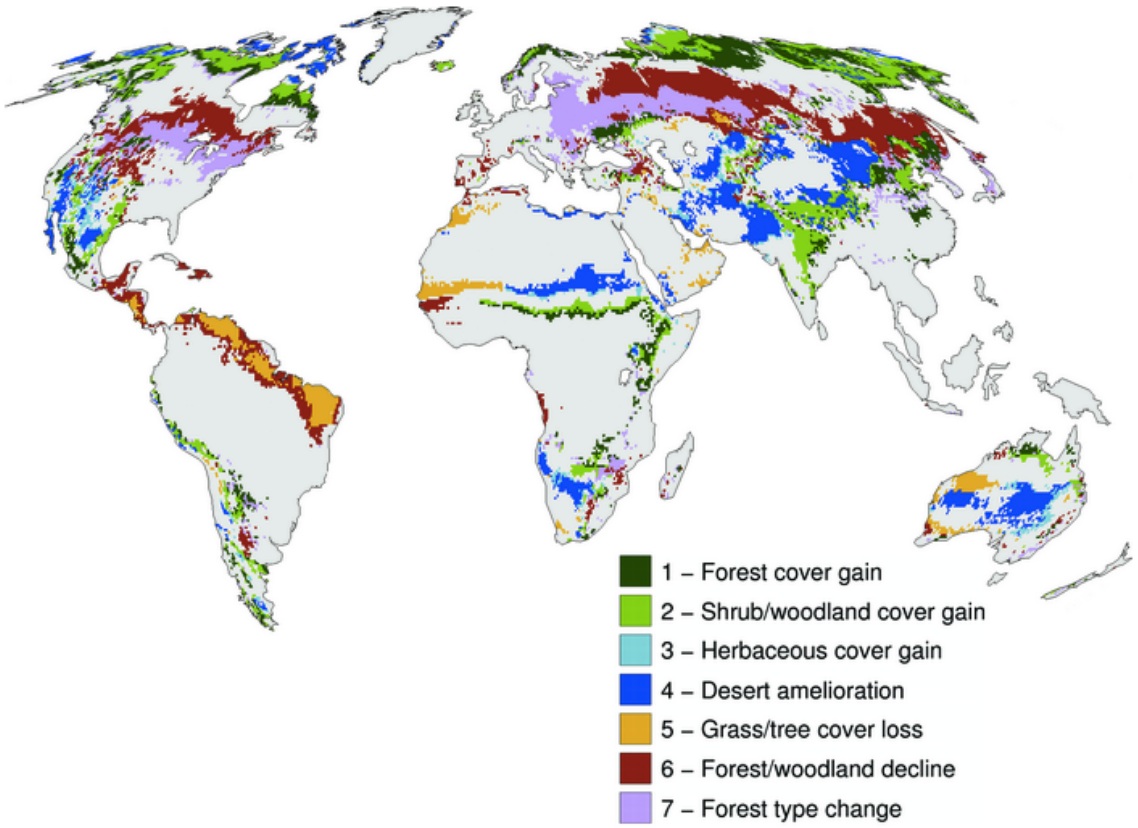

FIGURE 7.2: Projected ecosystem changes between 2000 and 2100, from the Intergovernmental Panel on Climate Change (IPCC) Annual Report 4, Working Group II [4], based upon the HadCM3 climate model. Changes are only shown where they exceed 20% of the area of a simulated grid cell (see full report for details).

Favourable conditions for adaptation may of course allow for the longer-term survival of some Glacial period rangeland taxa. While it is unlikely that mammoth steppe species like *Draba simmonsii* and *Ranunculus turneri* (see chapter 5) will retain competitive over their current distributions, one hope for adaptation is perhaps their tolerance of seasonality in more northern areas: if, at some latitudes and altitudes, even part of the season is too harsh for the survival of more mesic-adapted competitors, then this may limit invasions, allowing the mammoth steppe survivors to exploit their ability to survive harsh conditions as seeds, while continuing to rely on hybridisation and phenotypic/genomic plasticity to adapt to

whatever other changes to normal growing season conditions occur.

Optimistic scenarios aside, this work does bear a fairly ominous message about adaptability: Even where organisms seem to have undergone active selection for the capacity to adapt (a process Richard Dawkins first described as *the evolution of evolvability* by [1]), this fast evolving capacity is still limited to a range of phenotypes suited to a particuar suite of environments, and is still easily outpaced by changes in those environments. In the case of the centennial-scale changes currently threatening the boreal cryoxeric biomes, we should expect the local biota to be largely replaced by extant species migrating from other areas.

The isotopic work described in chapter 2 deals with the systemic interruption of a symbiotic relationship between large animals and their environments, driven by a climatic variable (in this case moisture) that underwent a millenial-length spike. While there are few regions where the same increased moisture can be predicted with any great confidence (the southern Sahara and western Europe being possibilities, according to the IPCC predictions shown in figure 7.2), the risks are undeniable. Chapter 2 focuses on the rangeland megafauna that depend upon graminoid-dominated biomes, which are sustained by a particular range of climatic variables. The predicted meridional migration of some climate zones may allow such ecosystems to shift, rather than undergo complete replacement. As figure 7.2 shows particularly well, in rangeland regions such as central-eastern Europe, north-central Africa, and east-central Australia, a longitudinal banding of biomes is likely to remain, but with the optimal range for rangelands shifting northwards. Nevertheless, this change is expected to be rapid, analagous to the global changes studied in chapter 2, and it is this rapidity which likely contributed to the pleistocene megafaunal demise. As such, the establishment of wildlife reserves, migration coridoors, and conservation efforts in general may play an essential part in preserving the biodiversity of rapidly-changing environments over the coming centuries.

More generally, however, this work reinforces a broader message on the fragility of ecosystems, particularly where long periods of coevolution has enforced a high level of interdependence between organisms. This certainly applies to the relationship bewteen megafauna and rangelands, which is still highly relevant in African and Australian ecosystems, for instance. Possessing such great biomass, and hence having the capacity to influence the surrounding environment, it is expected that megafauna especially will usually form a pivotal part of the biological interactions wherever they exist. This message therefore applies equally to both terrestrial *and* aquatic environments, where the lower trophic roles analogous to graminoid growth on the mammoth steppe—phytoplankton and corals, for instance—are certainly subject to changing environmental pressures under anthropogenic climate destabilisation.

7.3 Closing Remarks

"Life on earth is more like a verb. It repairs, maintains, re-creates, and outdoes itself."

Lynn Margulis, What is life?

"... for the difference is not great between fearing a danger, and feeling it; except that the evil one feels has some bounds, whereas one's apprehensions have none."

Pliny the Younger, Letter to Macrinus following a Vesuvian eruption.

Studying the effects of past climate change is particularly daunting living at a time in history when rising CO_2 levels exceed 400 ppb, and extreme weather records around the globe are broken annually. Lynn Margulis, who developed the endosymbiosis theory, reminds us that the threat of climate change is not to life itself, but that in the recreations it forces, individual species bear a very real risk of extinction. It could been argued that humans, being uniquely apt to engineer their environments and promote their own survival, may have more reason for optimism than some of the plants and animals studied in this thesis. In fact, as Margolis has often pointed out, symbiotic relationships occur at all levels of life: Human technology has itself evolved—to suit the stable climates humans have enjoyed for the past 10,000 years.

Fundamental technologies, such as those used in power generation, irrigation, agriculture, and the trading of resources, will find themselves under pressure to adapt and scale fast enough to keep food, water, and power supply networks broadly functional and minimise human suffering as the globe changes. The chapters of this thesis are motivated by the need for greater predictive abilities at the ecosystem level, so that these mitigating steps may be planned and applied.

Pliny the Younger's bleak comment on the relationship between suffering and foresight is highly apt to describe humans' approach to climate change: The human suffering that may occur as global ecosystems readjust to regional climatic changes is most unnerving because it is difficult to quantify. We can, however, be sure that this suffering will be reduced the better the future can be predicted, and the better the mechanisms of ecosystem collapse are understood, allowing us to channel effort into productive conservation and adaptation measures. It is my hope that the methods and findings presented here, and the pursuits that arise from them in the future, contribute some small amount to this understanding.

The work described in this thesis was performed in association with the Australian Centre for Ancient DNA (ACAD) and the Yukon Palaeontology Program (YPP) between 2013 and 2016. I am indebted to a large number of correspondents, collaborators, and others, who are listed in the acknowledgements. All the sequencing data, laboratory records, and samples, are available via ACAD and the YPP, and I would welcome opportunities to collaborate on future projects that might make further use of these resources.

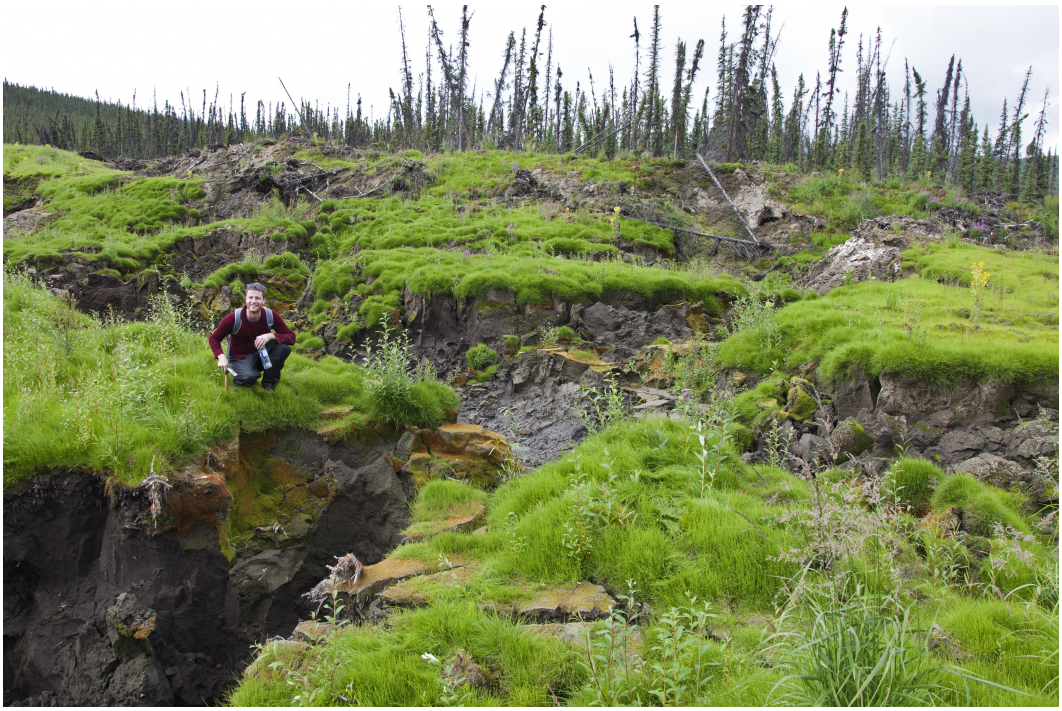

FIGURE 7.3: Surveying the permafrost exposed by a stream sampling trip at Laskey Creek in the vicinity of Dawson City, Yukon Territory, Canada. The 'drunken forest' of black spruce in the background is specialised to grow on permafrost with a shallow active layer. Driving over the ridges and depressions along Yukon highways, one observes a stark transition wherever the active layer deepens, with black spruce being replaced with white spruce, poplar, ash, and other trees. As global warming progressively deepens active layers and the permafrost retreats northward, black spruce forests risk being compressed against the northern tree line, and eventually exterminated. The lush green horsetails (*Equisetum sp.*) that cover most of the foreground, however, are a reminder that where conditions are appropriate, recolonisation and the succession of species will tend to re-establish the natural community from nearby source populations. This mine site was probably decomissioned only 1–2 years ago.

Chapter 7 Bibliography

[1] Richard Dawkins. "The evolution of evolvability". In: *On growth, form and computers*. Ed. by S. Kumar and P.J. Bentley. Elsevier Academic Press London, 1988. Chap. 13, pp. 239–255.

[2] B. Kirtman et al. "Near-term Climate Change: Projections and Predictability". In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by T.F. Stocker et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013. Chap. 11.

[3] Peter Lesica and Elizabeth E Crone. "Arctic and boreal plant species decline at their southern range limits in the Rocky Mountains". In: *Ecology Letters* (2016).

[4] ML Parry JP Palut and Osvaldo F Canziani. *Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*. 2007.

[5] Anibal Pauchard et al. "Non-native and native organisms moving into high elevation and high latitude ecosystems in an era of climate change: new challenges for ecology and conservation". In: *Biological invasions* 18.2 (2016), pp. 345–353.

(End of Document)