On the use of ancient DNA from plants and microbes for evolutionary inference

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Clemens Leonard Weiss
aus Sontra

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 05.07.2019

Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Oliver Kohlbacher

Zusammenfassung

Die kostengünstige Sequenzierung vollständiger Genome hat viele Wissenschaftsbereiche revolutioniert, von anwendungsorientierten Gebieten wie Medizin und Züchtung bis hin zur Evolutionsbiologie und Archäologie. Obwohl die Sequenzierung der Genome heute lebender Organismen zu signifikanten Durchbrüchen geführt hat, kann die Sequenzierung von DNA erhalten in historischen und antiken Relikten wie Skelettresten von Säugetieren, archäobotanischen Funden oder Museumsexemplaren in vielen Fällen zu anderweitig unerreichbaren Erkenntnissen führen. So hat das Erforschen solcher antiker DNA unser Verständnis der Menschheitsgeschichte wesentlich verbessert, die Rekonstruktion von Genomen ausgestorbener Arten ermöglicht und es uns erlaubt, genetische Veränderungen die mit Prozessen wie der Domestizierung von Pflanzen und Tieren einhergehen zu verfolgen. Um aussagekräftige Rückschlüsse aus den Daten der Sequenzierung antiker DNA ziehen zu können, muss man jedoch die besonderen Eigenschaften der antiken DNA-Moleküle aus Proben unterschiedlichen Ursprungs verstehen. Insbesondere der postmortale Zerfall von DNA-Molekülen ist eine Herausforderung für die Datengenerierung und -analyse, die es zu bewältigen gilt. Die vorliegende Arbeit trägt dazu bei, einige dieser Herausforderungen zu überwinden.

Um ein besseres Verständnis der Kinetik des DNA-Zerfalls zu erlangen und um die Planung von Versuchen zu unterstützen, wurden DNA-Zersetzungsprozesse mittels einer Zeitreihe von DNA aus Herbarbelegen untersucht. Dadurch konnten Muster altersbedingter DNA-Schäden identifiziert werden, die sich im Laufe der Zeit ansammeln. Untersucht wurden insbesondere solche Muster, die durch zufällige Fragmentierung des DNA-Gerüsts und durch die Desaminierung von Cytosinen verursacht werden, da diese direkt aus DNA-Sequenzdaten nachvollziehbar sind. Anschließend fokussierten wir uns auf die Auswirkungen dieser Schäden auf die weitere Analyse von Sequenzdaten antiker DNA. Eine Folge des DNA-Zerfalls ist ein erhöhtes Kontaminationsrisiko durch moderne DNA, die oft in einer viel höheren Konzentration vorliegt als die stark fragmentierte endogene DNA einer antiken oder historischen Probe. Dies erfordert die Authentifizierung von DNA-Sequenzen durch Nachweis ihres antiken Ursprungs. Aufgrund der ständigen Gefahr von exogener Kontamination in Studien antiker DNA untersuchten wir mehrere Ansätze zur Authentifizierung, von der Anwendung neuer Laborverfahren bis zur Entwicklung einer statistischen Methode zur Datenanalyse. Diese Methoden basieren in erster Linie auf dem Vorhandensein altersbedingter Schadensmuster, die wir und andere als allgegenwärtig in authentischen antiken DNA-Sequenzen nachgewiesen haben. Nachdem ein positiver Authentizitätsnachweis erbracht ist, besteht das Ziel oft darin, die genetische Variation innerhalb der gesammelten Proben, oder zwischen neu gewonnenen Proben und Referenzgruppen zu untersuchen. Um dies zu ermöglichen haben wir Methoden entwickelt, die die Analyse von Nukleotidvariation in Sequenzdaten mit für antike DNA

typischer niedriger Genomabdeckung erlauben. Darüber hinaus hilft eine weitere von uns entwickelte Methode bei der Analyse intra-spezifischer Ploidievariation direkt aus Sequenzdaten. All diese Methoden sind mit Fokus auf das Studium antiker DNA konzipiert, können aber auch in einem allgemeineren Kontext angewendet werden. Schließlich haben wir was wir über die Eigenschaften der antiken DNA gelernt haben, sowie die von uns entwickelten Methoden angewandt, um antike DNA-Sequenzen aus archäologischen Sedimenten zu untersuchen. Wir demonstrieren, wie spezielle experimentelle Verfahren und Analysemethoden auf solchen Sequenzen basierende aussagekräftige Rückschlüsse auf Evolutionsprozesse ermöglichen, mittels derer wir die Domestizierungsgeschichte der edlen Weinrebe, einer wichtigen Nutzpflanze, ergründen konnten.

Insgesamt trägt die vorliegende Arbeit zu unserem Verständnis vieler Aspekte der Handhabung von DNA aus antiken und historischen Proben bei und erschließt die hier dargelegten experimentellen und analytischen Ansätze für eine größere Vielfalt von Probentypen. Dies wird es ermöglichen, die Sequenzierung antiker DNA für eine wachsende Vielfalt von Organismen, insbesondere Pflanzen und Mikroben, zu nutzen, um die Inferenz von Evolutionsprozessen zu verbessern. Darüber hinaus versprechen wir uns von unseren Beiträgen eine kontinuierliche Verbesserung der Standards bei der Arbeit mit antiker DNA, insbesondere in Bezug auf die Authentizität von Sequenzen, auf denen nachfolgende Erkenntnisse beruhen.

Abstract

Our ability to cost-effectively sequence entire genomes has revolutionized many areas of science, from applied fields like medicine and breeding, to evolutionary biology and archaeology. Although the sequencing of extant genomes has led to significant breakthroughs, our understanding of many fields can be further enhanced by sequencing the genomes of historical and ancient specimens such as mammalian skeletal remains, archaeobotanical remains or museum specimens. Studying this ancient DNA has greatly improved our understanding of human history, allowed the reconstruction of genomes of extinct species, and has enabled us to track genetic changes during processes such as the domestication of plants and animals. To make meaningful inferences from ancient DNA sequencing data, however, requires understanding the unique characteristics of ancient DNA molecules extracted from different types of specimens. Particularly the post-mortem degradation of DNA molecules poses challenges for data generation and analysis that need to be addressed. This work contributes to overcoming several of these challenges.

To improve our understanding of the kinetics of DNA degradation and to aid experimental design, we studied degradation processes in a time-series dataset of herbarium specimens. This allowed us to identify patterns of age-associated DNA damage which accumulate over time, most notably those caused by random fragmentation of the DNA backbone and by the deamination of cytosines. Next, we focused on the implications of these damage patterns for ancient DNA data analysis. One consequence of DNA degradation is an increased risk of contamination with modern DNA, which is often at a much higher concentration than the highly fragmented endogenous DNA of an ancient or historical specimen. This necessitates the authentication of DNA sequences by providing proof of their ancient origin. Motivated by the constant danger of exogenous contamination in ancient DNA research, we investigated several approaches to aid authentication, from the application of novel laboratory procedures to the development of a statistical method for data analysis. These methods primarily rely on the presence of age-associated damage patterns, which we and others have shown to be present ubiquitously in authentic ancient DNA sequences. Once positive evidence of authenticity has been provided, the aim is often to study the genetic variation within a collected sample set, or between newly acquired samples and reference panels of genetic variation. To facilitate this, we present methods which are designed to allow the assessment of nucleotide variation from low-coverage sequencing data typical for ancient DNA. In addition, we developed a method to investigate intra-specific ploidy variation from sequencing data directly. All of these methods are designed with a focus on ancient DNA applications but can also be applied more broadly. Finally, we applied what we learned about the characteristics of ancient DNA, and the methods we developed, to

study ancient DNA sequences from archaeological sediments. We show how specialized experimental procedures and analytical methods permit meaningful evolutionary inference from such sequences, which allowed us to illuminate the domestication history of cultivated grape, an important fruit crop.

Altogether, the work we present contributes to our understanding of many aspects of working with DNA from ancient and historical specimens and opens up opportunities to apply the experimental and analytical procedures presented here to a larger variety of sample types. This will allow the use of ancient DNA sequencing for an increasing diversity of organisms, especially plants and microbes, to enhance evolutionary inference. In addition, we anticipate that our contributions add to the continuous improvement of the standards applied when working with ancient DNA, especially regarding the authenticity of sequences on which subsequent inferences are based.

Acknowledgements

First and foremost, I'd like to thank Hernán Burbano for his supervision, mentorship and collegiality throughout my PhD. It has been a wonderful time in your group, and I have learned a lot from you! Thank you also for the interactions you have facilitated both within the group and with others, which in many cases have led to friendships or fruitful collaborations.

I'd also like to thank Detlef Weigel for his help and support both during my PhD, as well as in my pursuit of the next chapter of my scientific life, and for his input during thesis committee meetings. I'd like to thank Oliver Kohlbacher for being my supervisor at the University of Tübingen, and his input during thesis committee meetings as well as during the writing of this thesis. I'd also like to thank Kay Nieselt and Oliver Bossdorf for agreeing to be part of my thesis defense committee, together with Oliver Kohlbacher and Detlef Weigel.

I'm grateful to Sophien Kamoun for our collaboration especially early on during my PhD, and for being part of my thesis advisory committee, together with Oliver Kohlbacher, Detlef Weigel and Hernán Burbano.

Among my many other collaborators, special thanks go to Matthias Meyer for a wonderful and fruitful working relationship as well as his help and support, and the entire "DNA Sequencing Techniques" group at the Department of Evolutionary Genetics at the MPI for Evolutionary Anthropology in Leipzig who have been incredibly helpful collaborators on several projects. Also in Leizig, I'd like to thank Viviane Slon for being a great collaboration partner, and Kay Prüfer and Michael Dannemann for their help and discussions.

Many colleagues in Tübingen have contributed to the work presented here, through discussions, scientific support, and especially through their friendship. I'd like to express special gratitude to Julian Regalado who was the best friend, office neighbor, football fan, beer connoisseur and enthusiast of Mexican food I could have hoped for; Patricia Lang for her friendship, help and support especially during the last months of my PhD - I am looking forward to going to California with you! ; Moi Exposito for his friendship and scientific support - I have learned a lot from you and look forward to being in California with you too! ; Sergio Latorre for his friendship, collegiality, and for the many good times; Rafal Gutaker for his friendship, a great working relationship, and for being a great colleague; Cristina Barragan, Effi Symeonidi and Max Collenberg for their friendship and the many good times; as well as all other members of the research group for Ancient Genomics and Evolution.

All of Weigelworld I'd like to thank for being great colleagues and facilitating a wonderful working environment. I'd also like to thank those that have kept the department and facilities running smoothly, especially Rebecca Schwab, Hülya Wicher, Christa Lanz

and Julia Hildebrandt.

I'd also like to thank Anna-Lena Van de Weyer and Felix Bemm, who made the move from Würzburg to Tübingen with me and who have been great friends and colleagues. Among the other colleagues during my Masters in Würzburg, I'd like to thank Jörg Schultz, Markus Ankenbrand, Thomas Hackl and Frank Förster, who were instrumental in inspiring me to pursue a PhD degree in Bioinformatics in the first place.

All other collaborators and co-authors I'd like to thank for their valuable scientific contributions and insightful discussions. My gratitude goes also to all archaeologists and museum curators, without whom many of the projects I have worked on would not be possible.

Last but certainly not least, I'd like to thank my family for their support and their interest in my work. The biggest 'Thank You' of all goes to Ana for always being there for me, unconditionally, and supporting me through everything. Thank you for all the wonderful times we've had over the last years, and for those to come.

General Remarks

- In accordance with the standard scientific protocol, the personal pronoun **'we'** will be used to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Prologue

From the standpoint of science today, it seems sometimes puzzling, how the study of inheritance, and many aspects of what we call "genetics" today predate any understanding of what a "gene" might be. In fact, when Mendel studied the inheritance of traits in the late 1850s, it took another ten years until the DNA molecule was even discovered by Friedrich Miescher, and another 100 years until it's structure was solved by Rosalind Franklin, James Watson and Francis Crick. By the mid 1900s, it was also shown that this molecule indeed contains the information that is passed on, as organisms reproduce and propagate.

However, the foundations of evolutionary biology and genetics were laid out before this, prompted by the work of Charles Darwin, and further developed in the first few decades of the 20th century. Crucially, the work of Thomas Hunt Morgan helped to establish the link between Mendelian genetics and how we understand inheritance today, still with limited knowledge of the molecular processes underlying his findings. The breakthroughs of this era were further formalized by the likes of Fisher, Haldane and Wright in what became known as the modern synthesis of evolutionary theory.

Still, understanding the structure and function of the DNA molecule lead to an explosion of research efforts in molecular biology and quickly to a major breakthrough in the ability to interrogate the exact sequence of DNA using a procedure called "Sanger sequencing", after its inventor Frederick Sanger who published it in 1977.

Yet again, the interest in studying evolution was ahead of these developments, founding the field of molecular evolution based on protein fingerprinting and allozyme variation. In fact, important developments such as the neutral theory of molecular evolution predate the advent of Sanger sequencing by almost 10 years. DNA sequencing of entire genomes, initially only from viruses, laid the foundation of what we now know as comparative genomics, which has become an important aspect of furthering our understanding of genome evolution.

But studying evolution means studying an inherently temporal process. Even though, many efforts in molecular evolution and comparative sequence analysis rely on contemporary data, which are snapshots in time from which historical states of genomes or molecules are inferred.

One way to tackle this limitation is by experimental evolution, where the evolutionary process can be surveyed through time. An example of this is Richard Lenski's long-term evolution experiment, which consists of populations of the bacterium *Escherichia coli* that have been propagated and monitored since 1988. However, only a limited number of organisms can be subjected to such experiments, and the time period that can be investigated is limited by the time of the experiment. Additionally, it does not allow the direct investigation of historical genetic diversity, which may have been lost through time.

In the mid 1980s, DNA sequences were generated for the first time from an ancient specimen long after the organisms death, which was made possible by cloning extracted DNA molecules into bacteria. This breakthrough would lead to developments that have revolutionized our understanding of human history, and prove of great use for many aspects of evolutionary- and population genetics.

Although DNA was first sequenced from ancient specimens already in the mid 1980s, it took at least two decades to put this procedure firmly on the radar of researchers studying different aspects of evolutionary biology. This is primarily due to the difficulties of working with such ancient DNA (aDNA), which are continuously being addressed by major advances in molecular biology and sequence analysis.

A motivation for this work is to extend the ancient DNA framework to a wider range of organisms and samples, and to facilitate the use of this great resource in studying their evolutionary history. For this, we will assess ancient DNA specific characteristics in a range of samples, and investigate necessary precautions that distinguish working with ancient DNA from other genetic resources. We will present computational methods to study genetic variation, which is tailored to ancient DNA specific characteristics, but applicable also to questions and datasets which do not involve historical specimens. Then, we will apply what we have learned about ancient DNA, and have developed for its analysis, in the first characterization of nuclear ancient DNA from archaeological sediments, to inform the evolutionary history of the domestication of grape vine.

The thesis is organized as follows. We will start with an introduction on topics relevant to the work presented, where we will touch on aspects of DNA sequencing in general and ancient DNA in particular, as well as evolutionary genetics and metagenomics. The introduction will conclude with a statement of the major objectives of this work. This is followed by five chapters on projects aiming to work towards these objectives. Each of the these five chapters starts with an introduction on topics relevant to its content, followed by a description of the methods developed or applied, and the contributions of the chapter to the objectives of this work. Each chapter will be discussed in isolation, before the work presented in this thesis will be concluded and discussed in a broader context.

1 Introduction

1.1 The early days of ancient DNA

The history of ancient DNA sequencing begins in the mid 1980s, when DNA molecules extracted from ancient specimens were successfully cloned into bacteria[1,2]. Since only low quantities of DNA are preserved in such tissue, cloning was required to amplify DNA molecules to a concentration high enough for sequencing. This process was successfully completed first in DNA from quagga remains[1], an extinct equid species, and then for DNA extracted from Egyptian mummies[2]. However, cloning of these DNA molecules proved difficult, primarily due to modifications to the DNA that occur by oxidative processes post mortem. Such modifications can hinder the integration of these molecules into living organisms, as they interfere with DNA replication and may not be able to be repaired by cellular DNA repair machinery[3]. Additionally, the successful cloning of an ancient DNA source molecule into a vector permanently removes this molecule from the DNA extract, which made it difficult to independently reproduce results through multiple experiments. This lack of replicability made it hard to assess the confidence in these early cloning-based experiments.

Around the same time of these early cloning successes, a method was developed for molecular biology that would have monumental impact on how we study DNA sequences, and the type of samples we can analyze. This method is the polymerase chain reaction (PCR) and was developed in the 1980s[4,5]. PCR allows the targeted amplification of DNA sequences, and greatly improves the effectiveness of retrieving minute amounts of DNA from a sample of interest. Also, it is an *in vitro* method, so DNA modifications such as blocking lesions found in ancient DNA molecules only interfere with the amplification success of specific, modified molecules instead of the viability of an organism which gets transformed with a cloning vector[6]. Since PCR makes copies from source molecules, it also allows better replication, as the same DNA extract can be used for many PCR reactions.

PCR was the breakthrough that allowed generating ancient DNA sequences from ancient and historical remains for a variety of animal species (about 50 by 2004), primarily from their mitochondrial genome[7]. It also allowed to continue on what had been started with the quagga, as sequences from 19 extinct animals had been sequenced by the turn of the millennium, including the Moa, the saber-toothed cat, the cave bear and the mammoth[8].

The era of PCR-based ancient DNA also led to important developments in establishing guidelines for best laboratory practices and authentication of ancient DNA[9]. When extracting DNA from ancient and historical specimens, and during experiments with these

extracts, certain precautions are necessary, primarily to overcome the risk of contamination[7]. As ancient DNA is present only at low concentrations, spurious contamination with free DNA can quickly overwhelm the endogenous DNA[10]. Additionally, ancient DNA comes in short fragments, which are less suitable templates for PCR reactions than fresh, long DNA[3]. To make matters worse, DNA modifications through oxidative processes, which are found in ancient DNA, slow down the polymerase during PCR, which poses an additional risk towards the amplification of endogenous ancient DNA over modern contamination[6].

One type of such modifications are crosslinks of DNA molecules with other macromolecules present in the sample, as the result of the Maillard reaction. The chemical reagent *N*-phenacylthiazolium bromide (PTB) breaks such crosslinks, and has been shown to greatly improve the amplification success of authentic aDNA molecules[11,12]. Because of this, PTB is sometimes included in aDNA extraction protocols.

Since PCR leads to the targeted amplification of molecules which are interrogated based on the sequences of the primers used, the reaction will amplify anything that matches these primers. This poses an additional difficulty to judging the provenance of such sequences, as animal mitochondria tend to have high sequence similarity. In most cases, hypervariable regions of the mitochondria were targeted[13]. This allows the post-sequencing comparison of these sequences with modern variation. As an example, if ancient DNA is extracted and sequenced from an ancient mammoth specimen, one would expect the hypervariable region to be more similar to elephants than to humans, and fall outside the variation of modern elephants as the mammoth represents a sister clade[6].

All these considerations led to the establishment of guidelines for experimentation on such samples, as well as the authentication of sequences generated by PCR[7,8]. Laboratories were set up in a way that the handling of modern samples was strictly separated from the handling of ancient samples, as DNA sequences generated e.g. from modern sister species of extinct organisms could easily be mistaken as authentic ancient DNA. The equipment and work place used for working with ancient DNA would be bleached and UV-irradiated regularly to destroy any free DNA at risk of contaminating ancient specimens.

Additionally, the ease of PCR over previous cloning methods allowed the continuous replication of amplification experiments. DNA extracts from ancient specimens were used for several amplifications, to prove the amplification product to be a reliable result. On top of that, several DNA extracts from the same specimen were used for amplification. Due to the fragmented nature of ancient DNA molecules, it is expected that the amplification success is inversely correlated with the length of the target[3]. This expectation was used as an additional authentication criterion. Also, in the presence of contamination with free DNA, DNA may be amplified without extraction. This mandated the inclusion of negative controls for all steps of an aDNA experiment. Last but not least, exceptional results were required to be able to be reproduced in a second laboratory from the same ancient specimen[14].

Some of these measures were additionally motivated by extraordinary claims made during the PCR era of ancient DNA research. Two examples were DNA sequences

claimed to be from insects enclosed in amber from over 100 million years ago[15], and DNA sequences from an 80 million year old dinosaur bone fragment[16]. These results were produced without the safety precautions described above, and proved either unreplicable, or were provably due to contamination[17]. Additionally, DNA from this time period, which was termed "antediluvian DNA", seems incompatible with our understanding of DNA degradation[18]. The oxidative modifications already detectable in specimens just a few hundred years of age are expected to accumulate and continue with time, making DNA preservation over millions of years highly unlikely. In fact, during this time, many in the ancient DNA field thought that the upper limit of successful ancient DNA extraction was about 100,000 years.

An extinct organism that luckily falls within this limit, and that quickly sparked the interest of ancient DNA research, is the Neanderthal. This extinct hominin populated Eurasia for many thousands of years, before anatomically modern humans migrated out of Africa. Neanderthals are thought to have gone extinct about 30,000 years ago, and remains exist well within the perceived preservation limit of ancient DNA[19]. However, the extraction and analysis of ancient hominin DNA poses additional challenges. Human DNA is the most prevalent source of contamination of ancient remains[14]. When working with DNA from animals, this source of contamination can often be distinguished from the organism of interest by its sequence alone. Ancient hominins however are much more similar to modern humans than any animals, so contaminants are more difficult to distinguish.

Nevertheless, in 1997 the first mitochondrial sequence was reported from the original Neanderthal type specimen, which again was investigated by PCR[13]. Stringent safety precautions were used to minimize contamination, and these efforts had paid off. The mitochondrial sequence generated from this specimen was close to modern human mitochondria, but had a sequence never before observed, which fell well outside of modern human diversity. In addition, its divergence to modern human mitochondria was intermediate to the divergence between humans and chimpanzees, as would be expected for a hominin sister clade. Mitochondrial sequences of other Neanderthal specimens soon followed, which corroborated these initial findings[12].

Even though they were undoubtedly monumental successes, the findings made in the PCR era of ancient DNA would soon be overshadowed by another breakthrough from molecular biology, which would start the next generation of DNA sequencing.

1.2 2nd generation (of aDNA) sequencing

The introduction of second generation DNA sequencing technologies revolutionized how we generate and interact with all forms of genomic sequences, including but certainly not limited to ancient DNA sequences from ancient and historical specimens. The first draft of the human genome heralded the era of genomics in 2001[20], but was generated using immense amounts of resources, mostly due to the cost and effort required by Sanger sequencing. One might argue, that the real revolution soon followed with the introduction of sequencing methods with higher throughput and increased automation.

The length of most DNA molecules, such as for example human chromosomes, greatly exceeds the length of sequence data that can be generated by any method. Advanced Sanger sequencing machines would produce sequences several hundred base pairs (bp) long, at very high quality. To sequence entire genomes then, one would employ a method known as shotgun sequencing[21,22,23]. Here, DNA molecules are fragmented into sizes readable by the sequencing technology. These fragmented sequences have to be pieced together after sequencing to recover the sequence of the entire source molecule, a highly complex process known as sequence assembly.

In contrast to the targeted amplification of certain sequence motifs by PCR, shotgun sequencing uses so-called sequence libraries. Originally, a shotgun sequencing library represented a collection of DNA molecules present in a sample, which were cloned into bacterial vectors. This made them suitable for DNA sequencing, and enabled amplification and propagation of the cloned DNA fragment.

Ancient DNA sequencing is perfectly suited for such an approach, as the molecules extracted from ancient specimens generally are already fragmented to sizes below 100bp. These short fragments are difficult to interrogate by PCR, as the template is short and they are easily out-competed by modern contaminant molecules[3,6]. A library-based approach instead allows the amplification and propagation of DNA fragments through clones. Since all DNA fragments present in a DNA extract are sequenced, it additionally allows to assess the taxonomic composition of DNA sequences generated from such a library[24]. While this was successfully done for ancient DNA libraries, it also suffers from similar shortcomings as the early cloning-based ventures into ancient DNA, with complications mostly due to the difficulty of amplifying ancient DNA molecules *in-vivo* due to age-associated modifications[3].

The second generation of sequencing technologies primarily had two major innovations over this approach. First, the generation of sequencing libraries suitable for this technology does not require cloning of DNA fragments into bacterial vectors. Instead, short DNA molecules of known sequence, so called adapters, are ligated to the ends of all DNA molecules present in a sample[25]. These adapters are used to prime the sequencing reaction, but also immortalize the DNA fragment they are ligated to, as they allow the almost indefinite amplification by PCR. This is especially important when working with ancient DNA, where each molecule extracted from a unique ancient specimen can be rare and valuable. Also, this type of library preparation combines all the advantages of PCR-based approaches such as the lack of *in-vivo* amplification, with the advantages of shotgun-based approaches, which allow the untargeted interrogation of all DNA molecules present in a sample[26,27].

The second major innovation of these technologies is the vastly increased sequencing throughput (which is why they are sometimes referred to as "high-throughput sequencing" (HTS) technologies)[25]. As shotgun sequencing requires the sequencing of many short fragments, these technologies parallelize the concurrent sequencing of such fragments, so that hundreds of thousands, or even millions of sequencing reactions are carried out per experiment. In most second generation technologies, this increased throughput is achieved in large parts by replacing the electrophoresis required for sequence detection

in Sanger sequencing with imaging-based detection of incorporated nucleotides[28].

The first such technology was pyro-sequencing, commercialized by 454 technologies. Here, the incorporation of a new base is detected by using the pyrophosphate released upon base incorporation to trigger a reaction cascade which results in the emission of light[25]. The development of a commercial product using this technology through 454 technologies and led by Jonathan Rothberg was, in fact, tightly linked with its application in ancient DNA[27]. As described above, the approach perfectly lends itself to be used for ancient DNA sequencing, due to its highly fragmented nature, and because it overcomes many shortcomings of cloning- or PCR-based technologies. In 2006, 454's pyro-sequencing facilitated the first report of nuclear Neanderthal sequence data[27], although the report was published together with work describing a similar dataset generated using cloning-based shotgun sequencing[29]. However, the first ancient DNA application of this technology was published earlier the same year in a study investigating the taxonomic composition of DNA extracted from mammoth remains[26]. In addition to nuclear sequences, pyro-sequencing also led to the sequencing of the entire Neanderthal mitochondrial genome in 2008, largely supporting previous findings based on PCR of hypervariable regions[30].

Another high-throughput sequencing technology which would soon surpass 454's pyro-sequencing was initially developed by Solexa, and later acquired by Illumina[31,32,33,34]. This technology also uses imaging-based detection, but directly recognizes the incorporation of fluorescently labelled nucleotides. Identical DNA molecules are organized in clusters, which are generated through bridge-PCR from an original source molecule. When synthesizing complementary strands in these clusters in a synchronized manner, the light intensity is enough to be detected by a camera. This technology is both cheaper and easier to parallelize than pyro-sequencing, which is why it would quickly become the market leader[35,36]. Additionally, it was less error prone when sequencing homopolymers[37]. In its first implementations, the major disadvantage of Illumina sequencing would be the very short sequences it was able to generate. For ancient DNA however, where the read length quickly exceeds the size of the DNA fragment, the increased throughput of Illumina sequencing would quickly make it the sequencing method of choice[35].

A great advantage of library-based sequencing, as provided by the 454 or Illumina platforms, is that it allows the complete reconstruction of DNA molecules present in the extracts. As previously discussed, ancient DNA molecules have been subject to different degradation processes which leave signatures such as modifications of bases due to oxidation. Library-based sequencing allows the direct observation of some of these signatures, which was not possible through previous sequencing methods[38]. This was a very important development, as the direct observation of age-associated degradation patterns could be used to aid authentication of ancient DNA sequences directly[39]. This is contrast to proxies used during the targeted amplification of aDNA sequences using PCR, such as the relationship of amplification success and the length of the targeted sequence[7,8].

Arguably the most important age-associated degradation pattern observable in align-

ments of short ancient DNA sequences to modern reference sequences (such as e.g. the draft of the human genome) is an excess of C-to-T and G-to-A conversions. This pattern was observed already in sequences generated by PCR amplification[40], and was especially prevalent in the first aDNA sequences generated using library-based shotgun sequencing[27]. This pattern is due to the deamination of cytosine bases to uracil, which is read as thymine by most DNA polymerases. While this pattern has been incredibly helpful in the authentication of ancient DNA, it became evident quickly that it would interfere with the reliable assessment of genetic variation from ancient DNA sequences[27,38,39].

To tackle this problem, another breakthrough from molecular biology was necessary. After the detection of such substitution patterns in sequences generated through targeted amplification, it had also been shown that the treatment of DNA molecules with an enzyme called Uracil-DNA-glycosylase (UDG) greatly reduced the presence of such substitutions[40]. This enzyme removes uracils from DNA molecules, and leaves abasic sites[41]. Once these patterns were observed prevalently in sequences generated by library-based approaches, this enzyme was included in the library preparation protocol in a non-destructive way. It was found, that an additional DNA repair step after UDG treatment removed the overwhelming majority of such substitution patterns[42]. This paved the way for more reliable sequence variation data gathered from ancient DNA. Still, it became common practice that non-UDG treated sequences would be generated for authentication.

Another common problem when working with ancient DNA is the often very low fraction of DNA sequences from the organism of interest (endogenous DNA). This became evident already in the first shotgun-based approaches to sequencing DNA from ancient specimens[24,26,27,29]. Often times, the DNA of interest is at such low frequency, that generating sequence data from a specimen by brute force becomes unfeasible. While library-based sequencing allows the characterization of all DNA molecules present in a sample, it is sometimes desirable to enrich sequences from a certain organism over DNA from other sources. A method of choice to achieve this has become targeted hybridization capture, were probes of a specific sequence are designed to be complementary to sequences of interest[29,43,44,45]. These probes can then be used to enrich DNA libraries for molecules that show complementarity to targeted sequences.

Library-based high-throughput shotgun sequencing, as well as other technological advances outlined above, finally allowed the investigation of the entire nuclear Neanderthal genome in 2010[46]. This was a major breakthrough, as it represented the first full ancient genome from an extinct organism, but also because it showed for the first time definitive evidence for interbreeding between modern humans and extinct hominins. Later in the same year, ancient DNA from a hominin fossil lead to another groundbreaking finding. A finger bone fragment found in Denisova Cave in the Altai Mountains in Siberia turned out to belong to a sister clade to all Neanderthal sequences generated up to this point[47]. Thus, a genetically distinct population of ancient hominins was found - a discovery made possible only by the advances in ancient DNA sequencing. These findings show the immense power and potential of studying DNA sequences from ancient specimens by library-based sequencing.

All fossils discussed up to this point fall within the previously mentioned age range of less than 100,000 years. This barrier was broken most prominently in 2013, with what is still the oldest specimen to have yielded reliable nuclear sequence data. In a study investigating the evolutionary history of domestic horses, a horse bone recovered from permafrost which was dated to be about 700,000 years old yielded enough sequencing data to cover almost the entire horse genome[48]. The fragmentation of DNA molecules extracted from this sample suggested, that DNA might indeed survive up to one million years in permafrost. Still, very strong signatures of deamination can be observed, again showing that this pattern is a viable characteristic to authenticate ancient DNA sequences.

The oldest non-permafrost sample yielding authentic ancient DNA sequences was a hominin fossil found at the Sima de los Huesos site in Spain, which was dated to be about 400,000 years old[49,50]. The successful sequencing of this specimen was made possible by specialized DNA extraction and library preparation methods developed for ancient DNA applications[51,52]. Sequencing libraries were prepared from the highly fragmented DNA molecules using a single-stranded DNA library preparation method, where each DNA single strand can be used as a template for a library molecule[52]. This is in contrast to the standard double stranded protocols, were intact DNA double strands are required to create functional library molecules[53]. Additionally, biotinylated adapters are used to immobilize DNA molecules throughout the library preparation procedure. These measures greatly increase DNA recovery from ancient specimens, especially if DNA is heavily fragmented[54,55].

Developments in ancient DNA sequencing have primarily been driven by efforts to sequence DNA from mammalian remains, especially humans and extinct hominins. Still, from the early days of aDNA sequencing on, there have been endeavours to investigate the feasibility of DNA extraction from ancient and historical plant remains. These endeavours predate library-based sequencing methods, and were based on PCR. Efforts focused primarily on early remains from domesticated crop species such as wheat[56,57] and maize[58], but also already included the targeted investigation of archaeological sediments in search for plant plastid genomes[59].

The breakthrough of second generation shotgun sequencing was applied to plant ancient DNA with less urgency, and found its first application in studying the role of transposable elements in the genome evolution of cotton in 2012[60]. Studies on the domestication of bottle gourds[61], as well as the evolutionary history of maize[62] and barley[63] cultivars followed in the years from 2013 to 2016, but overall, there have been only few large scale surveys of ancient DNA derived from plants. This is in part due to the sparsity of remains suitable for DNA sequencing[64], as seed remains are often charred, which greatly diminishes their viability for DNA preservation[65]. In addition, fossilized wood contains only very little endogenous DNA, especially compared to mammalian bones[66]. Apart from archaeobotanical remains such as corn cobs[62,67], one of the most promising sources of plant ancient DNA are herbarium specimens[68,69].

Herbaria are archives of biodiversity in the form of dried, desiccated plant specimens, going back to the 16th century. Compared with other sources of ancient DNA such as

mammalian bones, these resources are vast, with over 300 million specimens estimated to be stored in herbaria globally[69,70]. While DNA from herbarium specimens had been interrogated by PCR[71], the first library-based study to utilize this resource did not, in fact, study the genomes of the plants that DNA was extracted from. Instead, the genome of interest was from a plant pathogen called *Phytophthora infestans*, the causal agent of the Irish potato famine in the 1840s[72]. Herbarium specimens of this era were identified, some of which showing lesions compatible with infection of *P. infestans*. DNA extracted from these leaves indeed showed to be a mixture of plant and pathogen DNA, reinforcing the power of using historical collections for genetic analyses. However, especially samples from herbaria remain under-utilized given the wealth of past biodiversity they contain.

Of course, the revolution of high-throughput shotgun sequencing was by no means limited to ancient DNA applications. After the initial draft of the human genome was published[20], the interest quickly turned to studying human genetic variation[73]. This means, that variable sites in the genome are assessed in as many people as possible, primarily with the aim to identify variants of medical relevance. Apart from humans, understanding patterns of genetic variation is of high importance also in domestic animals and cultivated plants, as much can be learned about traits relevant for agriculture, and how to improve them more efficiently through precision breeding[74]. However, the release of the human genome, along with the genomes of some model species, predated the high-throughput sequencing revolution. This meant, that genetic variation was initially studied using genotyping arrays, informed by what had been learned from the genome sequence. Initially, a little over 1 million sites were surveyed[75], out of over 3 billion sites that constitute human genomes. However, the extent to which these variants were associated with phenotypes of interest led to some frustration, in what became known as the problem of "missing heritability"[76]. There are multiple factors that contribute to this problem, but it became clear, that the genotyping strategy was biased towards common variants, which are more easily captured by the initial efforts to decide which variants to genotype[77]. Variants at lower frequencies, which might have major phenotypic contributions, are easily missed by such targeted genotyping strategies, as they have, by definition, a lower chance to be identified in any one individual. Each additional variant to be genotyped was costly however, so this problem was difficult to circumvent.

Highly parallel sequencing, especially as provided by the Illumina platform, led to a drop in sequencing cost that made it feasible to sequence many entire genomes to assess genetic variation genome-wide[78]. This promise spawned projects such as the 1,000 genomes project in humans[79] and the 1,001 genomes project in the model plant *Arabidopsis thaliana*[80,81], and has revolutionized genomics for many other species at a smaller scale[82].

1.3 2nd generation sequencing workflows

Analyzing the data produced by high-throughput shotgun sequencing is different in many ways from the approaches to interrogate genetic sequences which came before it. During PCR, the fragments which get amplified and sequenced are defined by the primer sequences that flank the targeted region. This targeted approach (ideally) leads to sequences only from a specified region of interest. Similarly, in arrays used to genotype individuals at a specific set of single nucleotide polymorphisms (SNPs), the set of sites which are genotyped is predefined and interrogated in a targeted manner.

In contrast, the result of shotgun sequencing is a vast collection of short sequence fragments, on the order of tens to hundreds of base pairs[83]. Their analysis is challenging not only because of the amount of data generated, but also because the collection of sequences from one individual has to be made comparable to the collection of sequences from another individual to be able to study genetic variation between them[84]. This is exactly why the creation of reference genomes is of such importance. A reference genome represents an example of a more or less contiguous genome sequence of a species, ideally representing all chromosomes in full. Still, it is important to remember that usually there is nothing special about the sequence of the reference genome which was chosen for sequencing, and that large amounts of species variation will not be represented within it[85].

What the reference genome provides however, is a common coordinate system for the afore mentioned collections of short sequence fragments from different individuals. These datasets, also called re-sequencing data, are not suitable to assemble entire genomes from scratch (known as de novo assembly). Instead, a reference genome is used as a template after which to order sequences from other individuals. Once these short sequences from many individuals are aligned along a reference genome, the positions covered by such sequences are comparable between individuals and can be interrogated for intra-species variation[86].

This process, known as read mapping, is the first step in such re-sequencing studies, after some initial quality control on the sequence data itself[84]. It is a very computationally intensive task, as many millions of short sequences are queried for sequence matches with the reference genome, while allowing for substitutions and gaps. Algorithmic tools such as using the Burrows-Wheeler transform for indexing of the reference genome, and using seed-and-extend alignment methods have made this problem computationally tractable[87,88].

A great achievement from the beginnings of this process, primarily driven by the 1,000 genomes project in humans, has been the standardization of file formats. While there are many formats in which to represent alignments between sequences, the SAM format ("Sequence Alignment/Map") is the de-facto standard to represent read mappings of short HTS data to a reference genome[89]. The popularity of the SAM format is largely due to its use by the 1,000 genomes project, but its adoption in other communities has been driven by a variety of factors. First, the SAM format is a human- and machine readable plain text format, which makes parsing and interacting with it easy for researchers

with different levels of experience. Storing read mappings in SAM format would however require immense amounts of disk space. Because of this, the format is accompanied by the Binary Alignment/Map format (BAM), which is a compressed representation of SAM. One of the biggest reasons for the adoption of the SAM/BAM infrastructure is the tooling provided for it. The `samtools` program provides many ways of interacting with these files, among which is the seamless conversion between BAM and SAM. For software developers, the backend of `samtools` is available in the `htslib` library. Apart from interacting with files of short read alignments, this library implements the backend for interacting with many aspects of the high-throughput sequencing infrastructure, and is a big reason for the successful standardization of tools and file formats in this field.

Once all sequences from all individuals are mapped to the coordinate system defined by the reference genome, it becomes possible to assess genome wide variation in this set of samples[90]. This can include the discovery of variable sites in a population of samples, as well as the genotyping of all samples at these sites. Alternatively, new samples can be genotyped at positions of known variation, foregoing the de novo identification of variable sites.

An important aspect of studying genetic variation from HTS data is the concept of sequencing coverage or "depth". All HTS technologies have some rate of sequencing error, which is generally higher than the error rate of Sanger sequencing[91]. Yet, library-based sequencing allows, in theory, to sequence all DNA molecules present in a DNA extract. Most of the time, these extracts contain many more than one copy of the genome, unless DNA is present at extremely low concentration. The presence of multiple copies of the genome in the sequencing library allows to sequence each genomic position multiple times in replicate, to a certain "coverage" or "depth". If sequencing error is random, each base in each fragment will have an equal chance of being erroneous. In contrast, real genomic variation will, if homozygous, be present in all DNA fragments covering a genomic position. Thus, while sequencing error and variation are indistinguishable at one fold coverage, they can be easily teased apart at higher coverage. The coverage required for a re-sequencing study then depends on the heterozygosity and ploidy of the target organism, as heterozygous variation is more difficult to distinguish from error, especially at higher ploidy levels.

Variable positions and genotypes are obtained from read mappings in a process called variant calling[92,93,94,95]. The aim of this process is to use the coordinate system provided by the reference genome, as well as reads aligned to it, to assess genetic variation across sequenced samples. The file format commonly produced by this process is the Variant Call Format (VCF), which is also standardized by `htslib` and supported by a well maintained infrastructure of tools[96]. This format includes the positional information defined by the reference genome, and the sequence information supporting variation in different samples. A common approach of variant calling tools is to first capture as much variation as possible, which is subsequently subjected to filtering based on different quality criteria[95,97]. These can include coverage filters, base quality filters, mapping quality filters (the uniqueness of a positional assignment of a short read to the reference), allele frequency filters within samples and within populations, allele balance filters, and

many more. If available, external high quality data of known variation and known genotypes can also be incorporated as training data for variant quality recalibration, as implemented in the popular Genome Analysis ToolKit (GATK)[94,97]. A final dataset of genomic variation in the population can then be used as input for evolutionary inference and population genetics software. Unfortunately, these tools generally do not conform to standardized file formats, and considerable effort is put towards file format conversion as many tools invent their own formats[98].

1.4 Evolutionary genetics and ancient DNA

Genetic variation data, once acquired for example by high-throughput sequencing data mapped to a suitable reference genome, is useful for a variety of purposes. Most commercially relevant are uses in human medical genetics[79], as well as for assisted breeding in domesticated plants and animals[74]. Additionally, the rise of consumer genetics has found large success, primarily by offering ancestry inference from genetic data[99]. In medical genetics and breeding, genetic data is primarily used for its association with phenotypes of interest. For precision medicine, this can be useful to identify risk variants for certain diseases, or to stratify a population of patients by their quantitative risk for a more complex phenotype[100,101]. In breeding, genetic data is useful to select parental individuals to breed elite hybrids.

The field of population genetics is concerned not only with such phenotypes, but also with population histories. This includes effective sizes of populations, relative split times of divergent populations or species, as well as ancient migration events[102]. Additionally, it is of great interest to assess to what extent different populations are structured or are exchanging migrants and genetic material[103,104].

For example, understanding the structure of human populations is of great interest in medical genetics. As most inferences are based on genotypic associations to certain phenotypes, it is paramount to exclude the possibility that genotypic associations stem from structured populations rather than from the phenotype itself[105,106,107].

In addition, these questions are of interest from a historical perspective, since the genetic signatures left behind by movement of genetic material across the world can inform the migration history of human populations, as well as the origins and dispersal of domesticated plants and animals[108,109].

In this context, population genetics also focuses on the inference of selective pressures that have acted, or are acting, on genetic variation. In natural populations, this can be of interest to understand local adaptation in a reverse ecology framework[110]. Here, acting selective pressures are used as a source of information on the ecological niche occupied by an organism. Inferring artificial selection pressures such as the ones imposed during domestication can inform us on the traits that were bred for, and the major transitions undergone during this process[111].

Most commonly, population genetic inferences are based on contemporary, extant populations, such as the 1,000 genomes project of humans[79], the maize HapMap project[112]

or the 1,001 genomes project of *Arabidopsis thaliana*[81]. These populations are used to infer historical population parameters, such as their effective sizes or split times, as well as the extent of population structure today. Much of the power of these inferences is due to recombination, which provides an intrinsic source of replication within each analyzed genome. Since the genome of each individual is effectively composed from contributions of its many genetic ancestors, it carries information about past demographic events far beyond just the genomes of its parents[102]. This is especially evident with approaches like the pairwise sequential Markovian coalescent (PSMC), which allows to infer population size histories from one diploid genome alone[113].

Still, there can be many historical scenarios leading to the same composition of extant population, and they often can not be easily teased apart from data sampled at a single time point[114]. This is especially true if intermediate states get lost due to extinction, fixation or replacement.

The power of supplementing population- and evolutionary genetics with ancient DNA from historical specimens is that it allows the sampling of such species or populations which carry intermediate genetic states. Examples are structured populations which have since been lost or replaced[115], early domesticates with intermediate genotypes[58,67], or entire extinct species which supplement inter-species comparisons with better time resolution[46].

As described before, this has been done for many, primarily mammalian, species, most prominently in the sequencing of the Neanderthal genome[46]. Only whole genome data from Neanderthal individuals allowed to infer, with high confidence, that modern humans and ancient hominins indeed interbred and exchanged genetic material. This endeavour was a major driver in the development of statistical and computational tools that would allow to infer introgression from genomic sequencing data. Indeed, the publication introducing the Neanderthal genome draft, as well as conclusive evidence for interbreeding with Neanderthals, also introduced the now well known ABBA-BABA test, also known as the D-statistic (or Patterson's D, to avoid confusion with other such statistics). The ABBA-BABA test aims to quantify shared ancestry between individuals, as observed from genetic drift shared between their genomes.

The power of the approach is the use of sites which are distributed across the entire genome. This provides semi-independent observations of ancestry in each genomic segment, which may have been contributed through different recombination events. Given a tree connecting four individuals, with one being an outgroup to the other three, the majority of genealogies along the genome will be concordant with this tree. These genealogies can be observed, at single sites, by the site patterns alone. If the fourth (outgroup) individual carries the ancestral state (A), and the third (most diverged ingroup) individual the derived state (B), the only concordant site patterns are BBBA, and AABA (Figure 1.1, top). The two discordant site patterns are ABBA and BABA. If drift is the only force acting on these genomes, these discordant trees can only be produced by incomplete linage sorting, which is expected to produce these discordant configurations in equal amounts (Figure 1.1). A significant deviation from the balance of discordant trees distributed across the entire genome can not be explained by drift alone. Instead, it is evidence for the exchange of genetic material between the two individuals

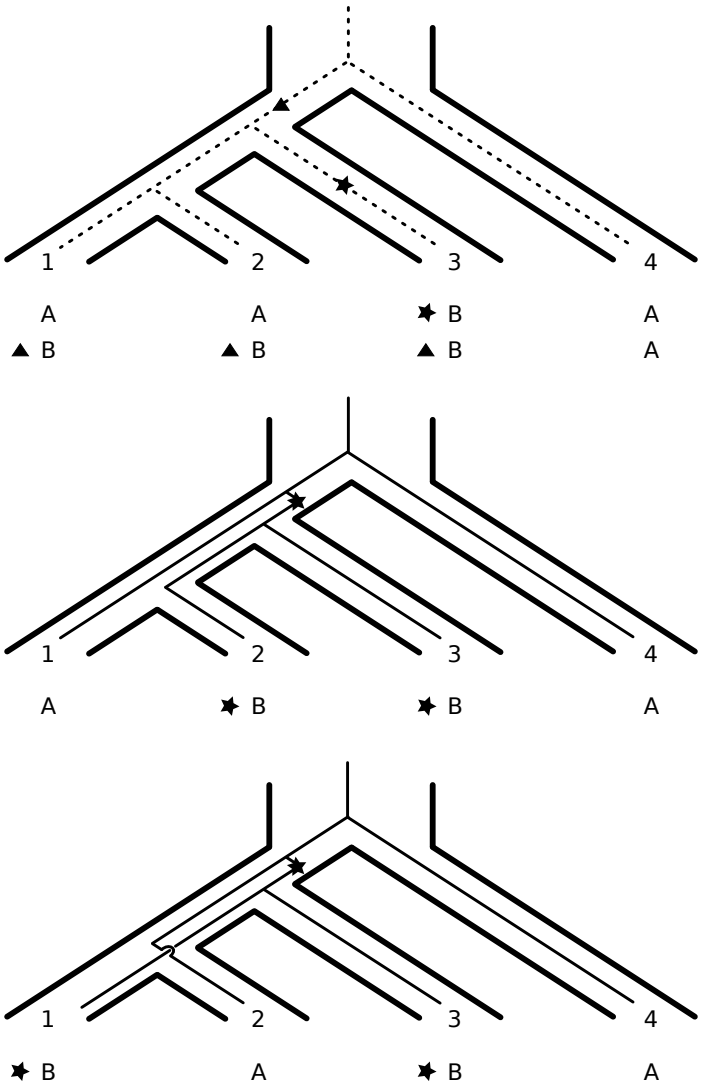

Figure 1.1: Schematic overview of concordant and discordant site patterns. There are two possible concordant site patterns in a four-population tree with an outgroup (pop. 4) carrying an ancestral allele ('A'), and an early diverged population (pop. 3) carrying the derived allele ('B'), which are shown on top. In the absence of gene flow, incomplete lineage sorting can lead to discordant site patterns (middle and bottom). Mutations from the ancestral to the derived allele are indicated by a star or a triangle.

carrying the derived state (B). Since genealogies are assessed by site patterns alone, this approach is very efficient computationally. Also, as genome wide patterns are observed, it allows statistical assessment of genome-wide significance using the blocked jackknife, which makes the statistic very robust[46].

A variation of the ABBA-BABA framework also allowed the estimation of the extent to which Neanderthals contributed genetic material to modern humans. This showed, that modern human genomes outside of Africa contain on average about 2% of Neanderthal ancestry[46].

In addition to the contribution of genetic material from Neanderthals to modern humans outside of Africa, this approach also helped to characterize a newly found hominin lineage called the Denisovan. This lineage is special as it is the only distinct hominin lineage initially discovered solely due to its genetics[47]. With the help of genomic sequencing data, and the statistical tools to analyze them, it was found that this lineage also contributed to the genetic material of modern humans outside of Africa. In contrast to the Neanderthal however, the contributions of the Denisovan are far less uniform, and most prominently found in Melanesians. The distinct patterns of ancient hominin contributions from these lineages in modern human populations quickly lead to the hypothesis of at least two events of interbreeding from either lineage, leading to the asymmetrical distribution of ancient hominin ancestry observed today[54,116].

The knowledge of this ancient hominin contribution to modern human genetic diversity has lead to a great interest in the distribution of these archaic contributions in human genomes, as well as their phenotypic effect. Again, this interest drove advances in method development, especially regarding the inference of introgression maps along human genomes[117]. These approaches show, that large parts of the Neanderthal genome are represented in modern human genomes, although none of the individual genomes carry more than four percent of Neanderthal sequence[118,119]. There are, however, large parts of the human genome devoid of Neanderthal introgression. These are thought to be associated with selection against the Neanderthal sequence, for example due to incompatibilities in human-Neanderthal hybrids[117]. As more data on human phenotypes and gene expression accumulated, there have been repeated efforts of assessing archaic contributions to human genome function. In contrast to regions under purifying selection against the Neanderthal sequence, there are also Neanderthal haplotypes at high frequency in modern humans, which are thought to have risen in frequency due to positive selection[120,121]. These sequences are associated with functions in immunity, metabolism, and responses to environmental queues such as temperature, light and altitute. For example, Neanderthal haplotypes have been shown to be associated in pigmentation and the regulation of the circadian rhythm[122].

In addition, ancient DNA has continued to paint a clearer picture on the interbreeding of early human and ancient hominin lineages, as more sequencing data from ancient specimens is being generated. For example, in 2015 the DNA from a 40,000 year old modern human specimen found in Romania was shown to carry substantially more Neanderthal DNA than any other modern human found before[123]. Interestingly, the size of these segments of Neanderthal ancestry suggested that the individual could have had a Neanderthal ancestor as recently as four to six generations in the past. More recently,

the DNA of the first true hominin hybrid was sequenced, when a specimen from Denisova cave was shown to have a Neanderthal mother and a Denisovan father[124].

The assessment of ancestry through the quantification of shared drift at sites distributed along the genome has continued to be a major contributor to advances in population genetics. A similar approach to the ABBA-BABA test introduced above has become known as the f-statistics family[125]. They allow the inclusion of population-scale data through the extension of the site-pattern framework to use allele frequencies, and provide a powerful toolbox to assess population differentiation, as well as admixture and introgression. Although informally introduced in studying the population history of India from modern sequencing data[126], this toolbox has found its major use in studying population histories of modern humans from ancient DNA sequencing data as well.

Over the last few years, an incredible wealth of genomes from ancient DNA of modern human remains has been generated. This is driven by lower prices of high-throughput sequencing, as well as advances in the technology to generate sequencing libraries from ancient mammalian remains. Primarily though, it has illuminated a vast history of migration, replacement and admixture that characterizes the history of our species, and that would have remained hidden from us without the ability to query the genomes of ancient remains[115].

An example is the population history of western Europe. Efforts of relating patterns of genetic diversity across Europe with potential routes of migrations from extant data alone proved problematic[127], as these proposed routes were highly correlated with patterns explainable by geographic structure of relatedness alone[128]. Only the interrogation of ancient DNA from samples widely distributed across space and time was able to disentangle the complex ancestry of European populations today. These have shown, that western Europeans are largely a mixture of three historical ancestry components. These are attributed to Hunter-Gatherer populations in the Mesolithic, Anatolian farmer populations from the Levant, which arrived in western Europe about 8,000 years ago, and a major migration from the Eurasian steppe about 5,000 years ago[109,115].

Interestingly, these migrations from the east are also strongly correlated with the domestication of plants and animals, and the spread of agriculture into western Europe[114]. For pigs, cattle and horses for example, the center of domestication is thought to be the near east. From there, domesticates spread in conjunction with human movements, including migrations into western Europe[108]. This is thought to be similar for many cultivated plants such as wheat and barley, though there has been less success in the recovery of ancient genetic material from archeobotanical remains than from mammalian bones[63,129].

As mentioned before, one of the great advantages of interrogating DNA sequences from ancient and historical specimens directly is that it allows to observe intermediate states. This is especially useful when looking at selection processes, which are characterized by the rise in frequency of certain alleles or haplotypes. Ancient DNA allows to identify such alleles through time, and discern when they arose with higher certainty than from extant genomes alone[130,131]. Denser sampling in time can even allow to to track allele frequencies as a population evolves. Some of the earliest work of using ancient DNA to identify selected alleles investigated processes of artificial selection during domestication and

cultivation. These studies included the characterization of the domestication syndrome in ancient maize[58], and the variation of coat color in domesticated horses[132]. One of the most prominent early examples in human populations is the rise in frequency of lactase persistence in western European populations, which has been shown to have risen in frequency only following the many migrations into the region described above[133,134].

In domesticated plants, the study of maize has benefited the most from ancient DNA. As mentioned, it was the first species for which ancient DNA was used to look at selection on genes associated with cultivation[58]. There has also been great interest in its population history and the relative contributions of different wild progenitor species into early cultivars, land-races and elite material[62]. In addition, genotypes from ancient maize specimens have successfully been used in phenotypic prediction using an experimental mapping population, which has greatly contributed to our understanding of its adaptation to temperate climate[67]. This approach is extremely powerful, as it allows the characterization of extinct phenotypes from genetic data recoverable by the techniques of ancient DNA sequencing.

More recent historical plant specimens, such as those preserved in herbaria worldwide, have helped for example to shed light on the population history of the model plant *Arabidopsis thaliana*. Due to the wealth of genomic data already available for this species, it has become a research interest to better understand how the population structure we observe today arose. Herbarium specimens have helped to show, that African *A. thaliana* individuals are native to the continent, and are the most deeply diverged lineages of the species[135]. In addition, its propagation through selfing has allowed to investigate the accumulation of mutations in a North American clonal lineage, using the demographic history of this population as a natural evolutionary experiment[136].

In all these studies presented here, ancient DNA was sequenced from remains of a species of interest. However, DNA extracted from these specimens is inherently a mixture of molecules from different sources. The research opportunity that this enables will be discussed in the following section.

1.5 Metagenomics and ancient DNA

From the beginnings of sequencing DNA extracted from ancient specimens, it became clear that the content of endogenous DNA, meaning DNA from the source organism of the specimen, varied between samples[2,7,8]. In contrast to early targeted approaches, library-based sequencing allowed to assess the composition of DNA molecules from their sequence, first by a cloning-based approach[24], and quickly thereafter using second generation sequencing[26]. These analyses showed, that DNA extracted from ancient specimens is an intrinsically metagenomic mixture of molecules originating from the genomes of a collection of organisms. These can be post-mortem colonizers of the tissue, or represent the genomes of organisms already present ante-mortem.

Thus, the taxonomic characterization of such a mixture of DNA sequences, as well as the authentication of putatively ancient components of it, plays a big role in ancient DNA sequence analysis. Here again, aDNA was a driver of methodological innovation,

as the analysis of the first ancient metagenome from a mammoth fossil[26] was conducted using an unpublished version of what would become one of the most popular software packages for the analysis of metagenomes, MEGAN[137].

Similarly to the alignment of sequences to the reference genome of an organism of interest, such taxonomic characterization is generally based on alignments. The major difference is that alignments are conducted to a collection of reference sequences, to be able to identify the presence of a variety of taxa.

The identification of DNA from organisms associated with ancient and historical specimens has been especially fruitful in studying ancient pathogens. Using samples contemporary to the outbreak of disease pandemics allows the characterization of DNA sequences which may otherwise have been lost to time, especially in quickly evolving pathogens. This opens for example the possibility of detecting virulence genes associated with past pandemics. Examples of such studies are the previously mentioned sequencing of *Phytophthora infestans* genomes contemporary to the outbreak which lead to the Irish potato famine, which was facilitated by herbarium samples of potato and tomato with visible lesions of infection[72]. Also, ancient DNA sequencing has facilitated the investigation of the genomes of *Yersinia pestis* from victims of the Black Death in the middle ages[138], as well as from Bronze Age samples[139].

In all these cases, care has to be taken to authenticate the historical origin of these genomes. As mentioned before, a major component of this process is the presence of deamination patterns characteristic for ancient DNA sequences[72,139].

Apart from identifying specific taxa such as pathogens in ancient DNA mixtures from plant and animal samples, there has also been an interest in characterizing the taxonomic composition of ancient DNA more broadly. An example of this is the characterization of oral microbial communities using ancient dental calculus. This source of microbial ancient DNA has become of special interest, since the taxonomic composition of sequences derived from it resemble to some extent the composition of contemporary oral microbiomes. This is in contrast to the microbial composition in ancient teeth and bones, which most closely resembles the profiles found in soil samples, most likely due to secondary colonization of these specimens[140]. Sequences from dental calculus have been used to associate the microbial composition with pathogens and disease[140], as well as with host diet[141,142]. However, the authentication of sequences in compositional analysis is often difficult, and special care has to be taken not to over-interpret spurious data[143,144].

Another potential source of DNA of historical origin which is not limited by the availability of suitable samples and specimens is sometimes referred to as environmental DNA. This term summarizes DNA extracted from sources such as lake sediment cores, cave sediments or permafrost ice cores. These types of samples have been shown early on to contain DNA from plants and animals[59], which can help to reconstruct the ecology of environments now inaccessible to us[145]. In sediments, DNA evidence can also be used to supplement evidence from other sources such as pollen or macro-remains to achieve a more complete, credible reconstruction of the past[146,147]. Still, just as with all other sources of ancient DNA, it is imperative to provide positive evidence for the

authenticity of DNA from taxa on which important conclusions rest. In DNA from these more complex sources, this is often complicated by the low abundance of DNA molecules, and the taxonomic complexity of the DNA mixture.

1.6 Objectives of this work

As described in this chapter, using high-throughput sequencing to interrogate DNA from ancient and historical specimens is a very powerful resource for many aspects of evolutionary biology and population genetics. It has illuminated the history of interbreeding between modern humans and ancient hominins, has allowed the detection of a previously unknown ancient hominin lineage, and continues to unravel the complex migratory history of human populations around the globe.

To date, the use of ancient DNA is primarily focused on mammalian remains, especially from ancient hominins and modern humans. A promising but underutilized resource are plant collections stored in herbaria, which represent an unprecedented wealth of samples from many different species. The last 500 years which are spanned by herbarium samples are an era of broad interest with regards to the discovery of the new world, the impact of global change, and the industrialized cultivation of crops[69]. The density of herbarium samples in space, time, and species distribution makes this resource valuable for many aspects of evolutionary genetics, for example through the ability to track allele frequencies through time in response to such events. However, to be able to effectively utilize this resource, it is important to understand the characteristics of ancient DNA extracted from this type of samples.

One objective of this work is to characterize patterns of ancient DNA damage and decay from DNA sequences of a time-series of herbarium specimens. The processes which lead to DNA fragmentation and chemical damages are strongly affected by environmental factors such as pH and humidity. This has made the detection of temporal relationships of these patterns difficult. The aim is to use herbarium specimens, which have been exposed to a more stable environment in comparison to remains buried in soil, to investigate the kinetics of these damage patterns.

In addition to herbarium samples, there have been recent advances in characterizing ancient plant DNA from other sources, including sediments. With ancient DNA from such sources, it is especially important to assess both the confidence of taxonomic assignments, as well as the authenticity of DNA from identified taxa. A second objective of this work is to critically assess measures to provide evidence for the authenticity of ancient DNA from these complex mixtures. Building on the insight gained into the temporal patterns of ancient DNA damage, these measures primarily rely on the presence of age-associated damage. These types of damage can be either selectively enriched using specialized library preparation protocols, or detected directly from HTS data.

However, sequencing data generated from ancient DNA is often inherently of low coverage, either due to the complexity of the DNA extract, the low relative abundance of the taxon of interest, or the short fragment size. This makes it important to devise approaches which can extract signal from such data. The third objective is to provide

computational methods to aid authentication of ancient DNA sequences, as well as to extract signals of genetic variation from low-quality sequencing data. Even though these methods are targeted at the ancient DNA research community, it is still important to provide user-friendly interaction with the method to a larger variety of users. Because of this, a special focus has been to use file formats which are commonly used in the respective research fields, and to implement stable, efficient procedures to interact with these data.

Lastly, what has been learned about ancient DNA from sources which are difficult to handle, and the methods developed to analyze low coverage, low quality data, shall be applied to further our understanding of the domestication and cultivation history of an important crop plant, in the first successful effort to recover nuclear sequence data directly from archeological sediments.

2 Ancient DNA decay

Contributions

Parts of the content of this chapter have also been published in the article "Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens"[148]. The following people have contributed to the work presented in this chapter: Hernán Burbano and myself conceived and designed the project, and wrote the article with help from all authors. I analyzed the data and generated all figures. Verena Schuenemann (VS), Ella Reiter (ER) and Billie Gould performed DNA extractions. VS and ER prepared sequencing libraries. Johannes Krause and John Stinchcombe coordinated laboratory experiments.

2.1 Ancient DNA decay in herbarium samples

DNA sequences generated from ancient and historical specimens have already revolutionized our understanding of human evolution[115], and have contributed to advances in many other areas of evolutionary biology such as the study of plant and animal domestication[68,108]. To utilize such sequences for evolutionary genetics analyses however, it is necessary to overcome certain difficulties which arise from the biochemical characteristics of ancient DNA molecules, primarily driven by post-mortem degradation[149]. These characteristics are the distinguishing features of ancient DNA (aDNA), rather than a strict temporal definition based on the age of a sample[150].

To effectively utilize this resource, it is important to understand these degradation processes, the signatures they left behind, as well as their relationship with the age of a sample. A better understanding of these patterns can inform experimental design[42,51], as well as the procedures used to analyze sequencing data[151].

Studying temporal patterns of age-associated degradation is however complicated by large variability in the environment and preservation, especially of mammalian remains[152]. As environmental factors such as temperature, pH and humidity vary for example between soils of different archaeological sites, so does the preservation of specimens from which DNA is extracted[153]. Nevertheless, characteristic biochemical patterns of DNA degradation have been found in ancient DNA from many different sources, including permafrost soil[48].

Herbaria archive a record of changes of worldwide plant biodiversity harbouring millions of specimens that contain DNA suitable for genome sequencing[69,70,72]. These collections of pressed, dried plant specimens have the advantage that samples are prepared and

stored using standardized procedures, which could potentially reduce the effect of environmental variation between samples[154]. Additionally, most herbarium samples contain leaf samples, making the tissue of origin more comparable than for example mammalian remains, where often times only few bones from different parts of the organism are preserved. This makes herbarium samples an ideal system to study temporal patterns of degradation processes in ancient DNA.

The sample set analyzed here consists of 71 samples from three different plant species spanning approximately 250 years (Table 2.1, Figure 2.1, Table S1). The samples were sequenced using library-based high-throughput sequencing, which allows the assessment of a set of damage patterns relevant for the analysis of genomic sequencing data. Additionally, a subset of these samples where infected with visible lesions of a plant pathogen, facilitating the analysis of correlated patterns of ancient DNA decay between host and pathogen.

Table 2.1: Species, number, and age range of herbarium samples. A subset of samples showed lesions compatible with infection by *Phytophthora infestans*.

Species	# samples	collection year (range)	# infected
Arabidopsis thaliana	54	1863-1993	-
Solanum tuberosum	12	1845-1896	12
Solanum lycopersicum	5	1737-1876	2
total	71	1737-1993	14

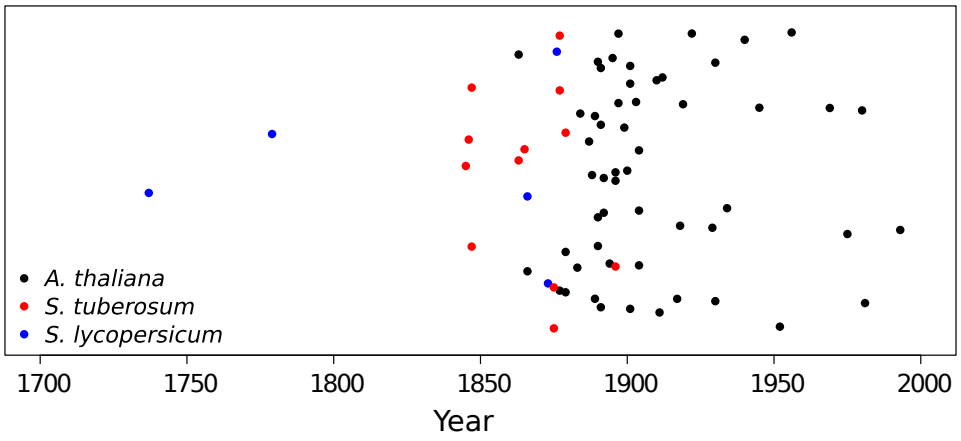

Figure 2.1: Temporal distribution of samples from *Arabidopsis thaliana*, *Solanum tuberosum*, and *Solanum lycopersicum*.

2.2 Results

2.2.1 Patterns of aDNA damage

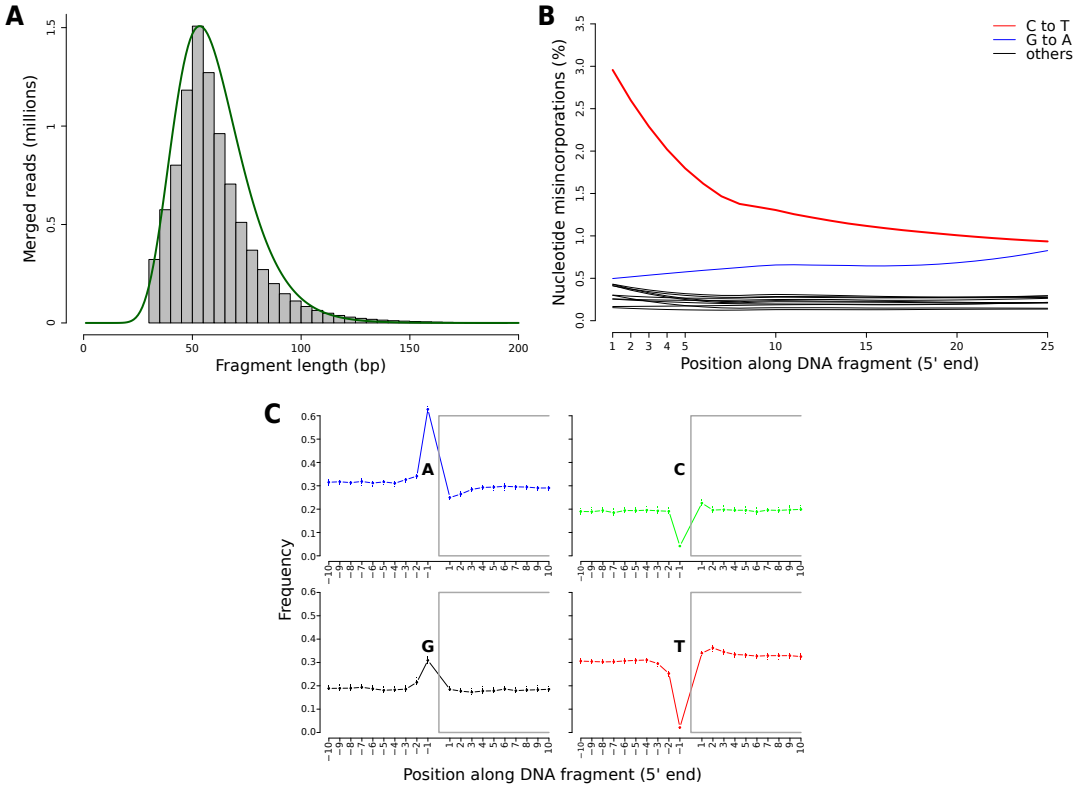

Figure 2.2: Patterns of aDNA damage detectable from short read sequencing in *A. thaliana* sample NY1365354. **A.** Distribution of fragment lengths of merged reads. The green line shows the fit between the empirical and the log-normal distribution. **B.** Nucleotide misincorporation profile at the 5'-end of reads. **C.** Base composition of the first 10 nucleotides of reads and their upstream genomic context, visually separated by grey boxes, and split by the base present at the -1 position of the 5' end.

When investigating ancient DNA sequences from a specific organism, there are primarily three patterns of age-associated degradation detectable from mappings of short read sequencing data to a suitable reference genome (Figure 2.2)[38].

First, we investigated the fragmentation of DNA into shorter and shorter molecules by assessing the distribution of sequence lengths. This is possible, since paired-end library-based sequencing allows to reconstruct the full aDNA molecule, as long as the molecule is shorter than the total length of the forward and reverse reads. This reconstruction is achieved by merging the forward and reverse reads into a single sequence, based on the overlapping sequence at their ends. In the herbarium dataset presented here, 96% of read pairs could be merged on average (83% - 99%), requiring at least a 10bp overlap.

The subset of merged reads which map to the reference genome of the organism of interest allows to investigate the size distribution of DNA fragments likely originating from that organism (Figure 2.2A). This empirical distribution of fragment sizes could be described by a log-normal distribution, which can be summarized using the median (following from $median = e^{log-mean}$).

Mapping merged aDNA sequences to a reference genome also allows to query their genetic context, since the sequences represent the original DNA molecule from 5'-end to 3'-end[38]. This allowed us to investigate one of the processes by which the DNA backbone breaks over time. The process is called depurination[155], and leads to breaks in the DNA following purine bases (A and G). From read mappings, the "-1" position of each mapped sequence can be queried, meaning the base immediately upstream of the 5'-end. A signature of ancient DNA is the enrichment of purine bases at this "-1" position, driven by fragmentation through depurination (Figure 2.2C).

The third signature is only tangentially related to the fragmentation of ancient DNA, and is caused by the deamination of cytosine bases to uracil[40,156]. These uracils are read as thymines during sequencing-by-synthesis, and are detectable as C-to-T substitutions in alignments of sequencing data to a suitable reference. The rate of deamination is much higher in single-stranded DNA than in double-stranded DNA[157]. Thus, cytosines predominantly deaminate at the ends of fragments, where they have single-stranded overhangs. When summarizing the magnitute of C-to-T substitutions across all fragments, this manifests itself as a characteristic pattern, where the frequency of misincorporations is highest at the 5'-end of fragments, and decays towards the center (Figure 2.2B).

In library preparation protocols which include end-repair and fill-in steps, the excess of C-to-T substitutions at the 5'-end is mirrored by an excess of G-to-A at the 3'-end, due to the complementary nature of the fill-in repair[158]. In contrast, library preparation protocols which use a single strand of DNA as their template show an excess of C-to-T at both ends[52,54].

2.2.2 Kinetics of aDNA damage

2.2.2.1 DNA fragmentation

A big advantage of ancient DNA from herbarium samples is the temporal density of samples available, which allows studying the dynamics and kinetics of age-associated degradation and how damage accumulates over time.

First, we investigated the relationship between median fragment length of a sequencing library, and the collection year of the specimen the DNA was extracted from. When regressing the natural logarithm of the median fragment size (or log-mean) on the collection year, we saw a significantly positive linear correlation (Figure 2.3). Thus, age and fragment size are exponentially related, with younger samples having longer fragments.

An exponential relationship can also be seen in the fragment size distribution of a single library alone (Figure 2.4A). This is the result of random fragmentation of DNA molecules[159], and means, that part of the fragment size distribution follows an exponential decay:

$$F(L) = F_0 \cdot e^{-\lambda L}$$

where L is the fragment length, $F(L)$ is the frequency of fragments with length L, F_0 is the frequency intersect at length 0, and λ is the decay constant. The decay constant can be understood in this context as the frequency of breaks in the DNA backbone, per site.

The decay constant λ can be inferred from the linear relationship that follows from logarithmic transformation of the exponential decay and has the slope $-\lambda$ (Figure 2.4B):

$$\log F(L) = \log F_0 - \lambda L$$

The DNA decay rate per base per year, k, is calculated as:

$$k = \frac{\lambda}{age}$$

Following this, once λ is inferred from the fragment size distribution of every library, k can again be inferred from the slope of a linear regression of λ as a function of sample age (Figure 2.4C):

$$\lambda = k \cdot age$$

The decay rate of nuclear DNA from the herbarium specimens analyzed here was $k = 1.66 \times 10^{-4}$.

One process that contributes to the post-mortem fragmentation of DNA is depurination[38,160], which can be inferred from sequencing data by the excess of purine bases at the position directly upstream of the 5'-end of each fragment. To investigate the temporal dynamics of this process, we assessed the relationship of purine enrichment with the age of each sample for both purine bases (Figure 2.5). The enrichment of purine bases was assessed by the relative enrichment of either base at the "-1" position over the "-5" position, where no enrichment is found (Figure 2.2C). An enrichment in purines at the "-1" position was found in all ancient DNA libraries analyzed here, but this enrichment was constant over the time period spanned by our samples.

2.2.2.2 Nucleotide misincorporation

Another signature of ancient DNA damage is due to the deamination of cytosines to uracils[40], which manifests in sequencing data as an excess of C-to-T substitutions at the 5'-end of sequences[38,156]. Since this pattern is always highest at the 5'-end, its magnitute there can be used as a proxy for the magnitute of this type of damage. The regression of the percentage of C-to-T substitutions at the 5'-end with the year of collection of samples showed a significant negative correlation, where older samples had more damage due to deamination of cytosines (Figure 2.6).

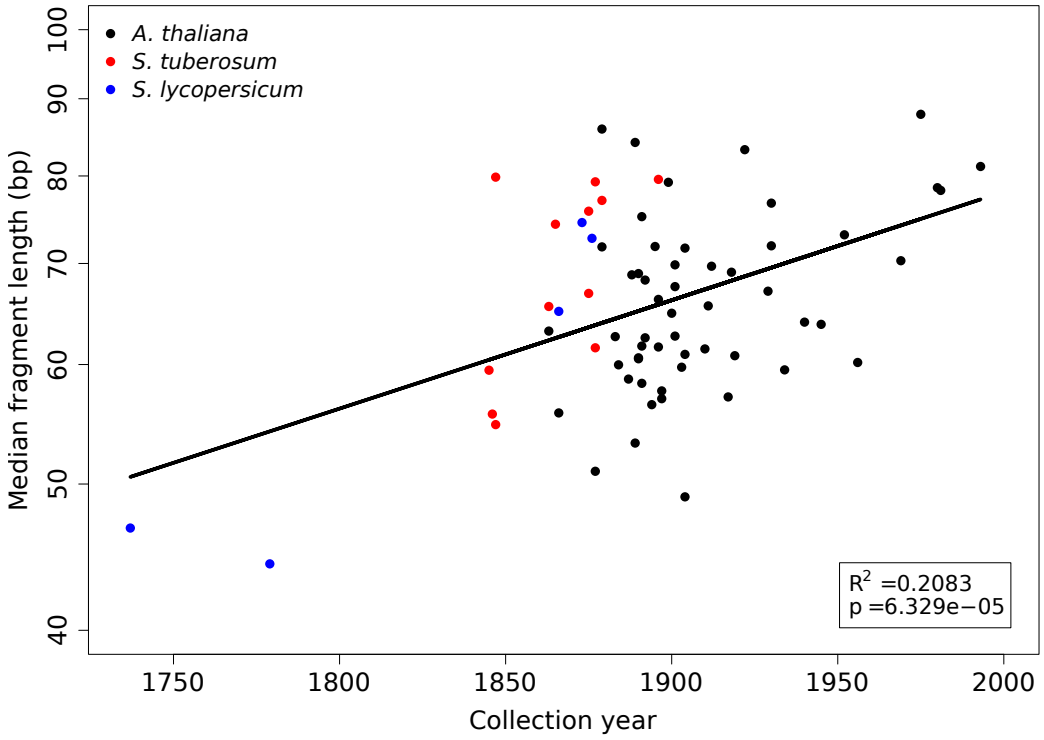

Figure 2.3: Median length of merged reads as a function of collection year (N=71). The line indicates the linear regression. The inset shows the regression statistics between the natural logarithm of median length and collection year. The y-axis is log-scaled and shows, therefore, that the correlation is exponential.

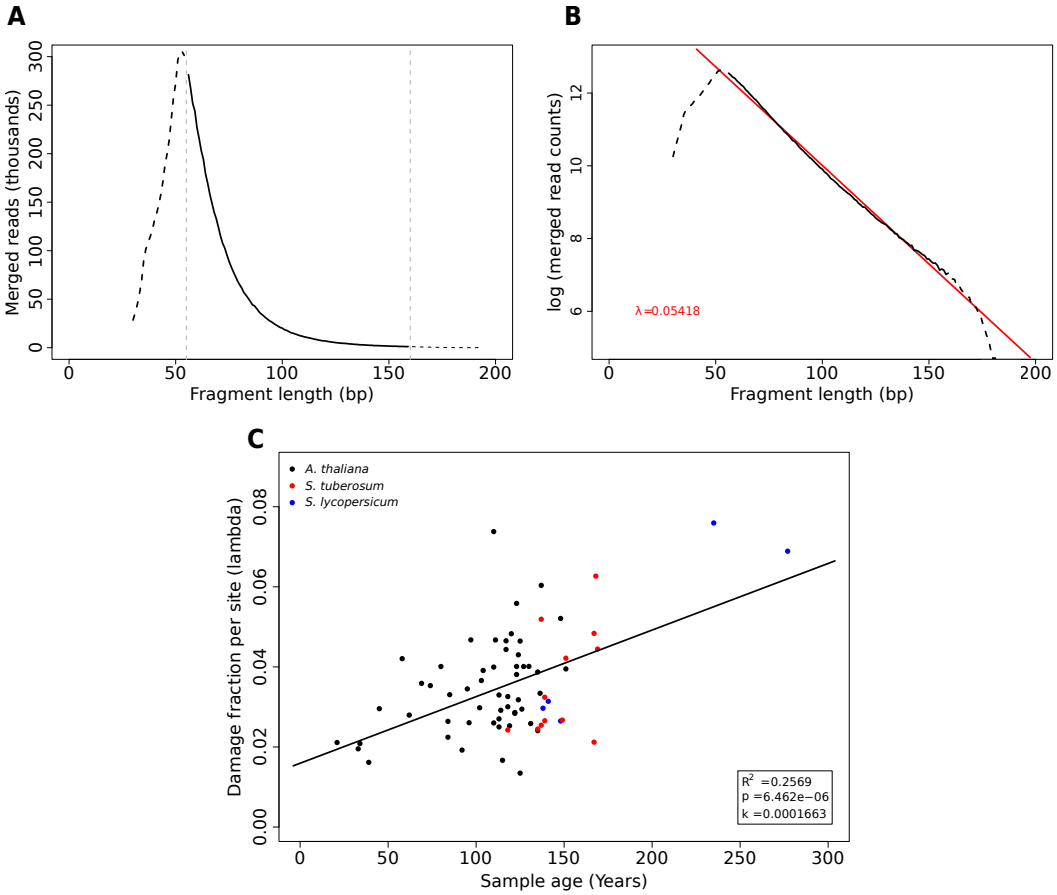

Figure 2.4: DNA fragmentation and decay rate. A. Distribution of fragment lengths of merged reads from *A. thaliana* sample NY1365354. The solid line, which is surrounded by horizontal dotted lines, shows the part of the distribution that follows an exponential decline. **B.** Distribution of fragments length for the same library using a y-axis with a logarithmic scale. The slope of the exponential part of the distribution (red line) corresponds to the damage fraction per site (λ). **C.** Damage fraction per site (λ) as a function of sample age (N=71). The slope of the regression corresponds to the DNA decay rate (k) following the formula: $\lambda = k \cdot age$.

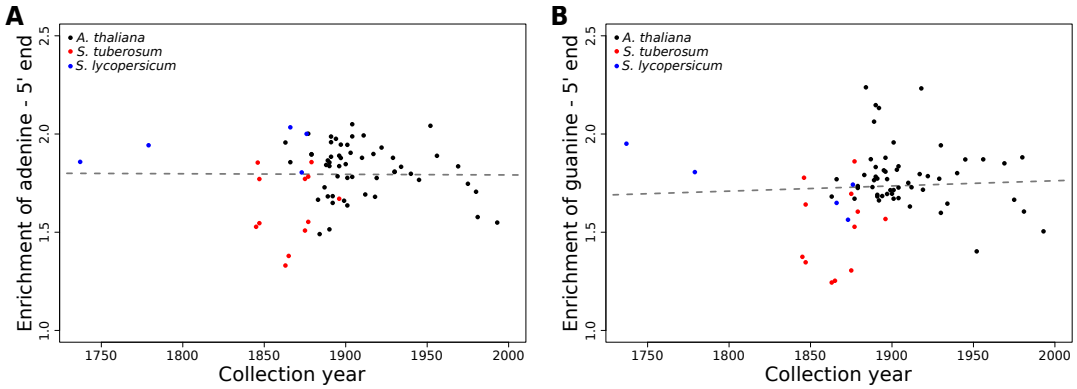

Figure 2.5: Signatures of depurination as a function of collection year. Relative enrichment of adenine (**A**) and guanine (**B**) at the 5' end (position -1 compared with position -5). The dotted lines show the linear regression.

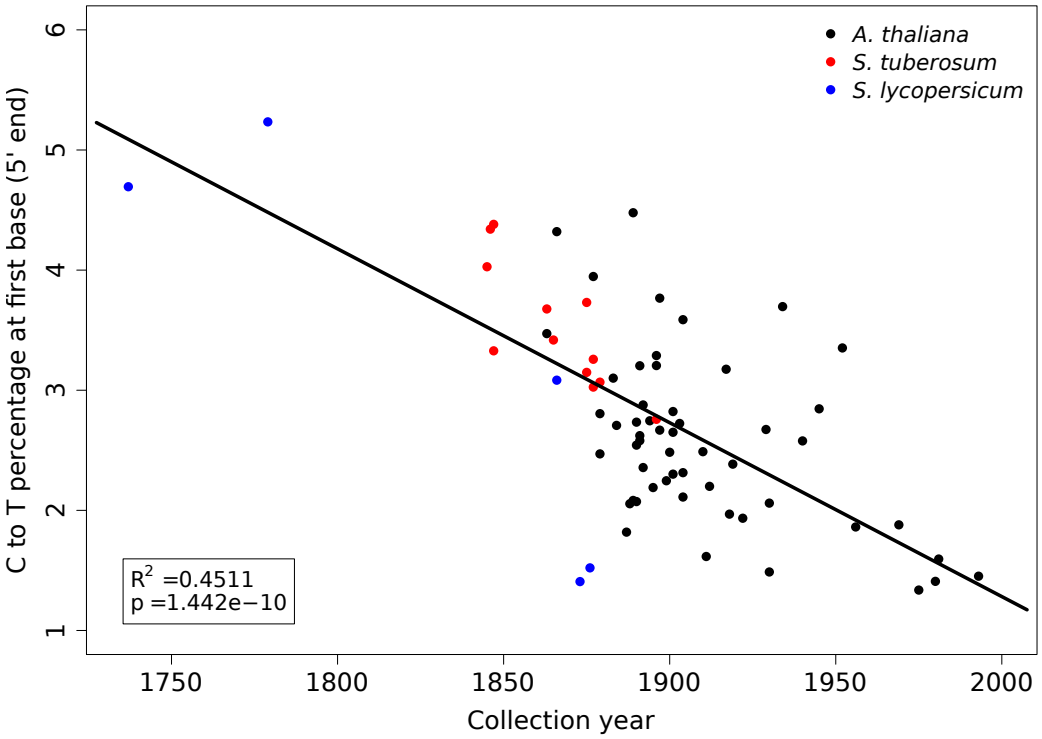

Figure 2.6: Signature of deamination as a function of collection year. The extend of deamination is summarized by the C-to-T percentage at the 5'-end of sequences from each sample (N=71). The inset shows the regression statistics for a linear relationship of C-to-T percentage and collection year.

2.2.2.3 Host - pathogen dynamics of ancient DNA damage

A subset of the *S. lycopersicum* samples, as well as all analyzed *S. tuberosum* samples showed lesions compatible with infections by the oomycete *Phytophthora infestans*. The presence of two eukaryotic genomes in the same specimen presented the unique opportunity to assess the extend of correlated damage kinetics between the nuclear genome of the host and the nuclear genome of the pathogen.

The relationship between C-to-T substitutions at the 5'-end of fragments and the year of sample collection showed that the magnitute of this damage pattern increased with sample age in the nuclear genomes of both host and pathogen (Figure 2.7A). While the absolute magnitude of deamination was different between the two genomes, there was no significant difference in the slope of its relationship with sample age (i.e. no significant interaction term Pr(Sample age:DNA origin) = 0.722).

We compared the extend of DNA fragmentation in both genomes using λ, which was calculated for sequences from the pathogen just as described for the host genome alone. This showed a strong correlation in the dynamics of DNA fragmentation between the different organisms (Figure 2.7B).

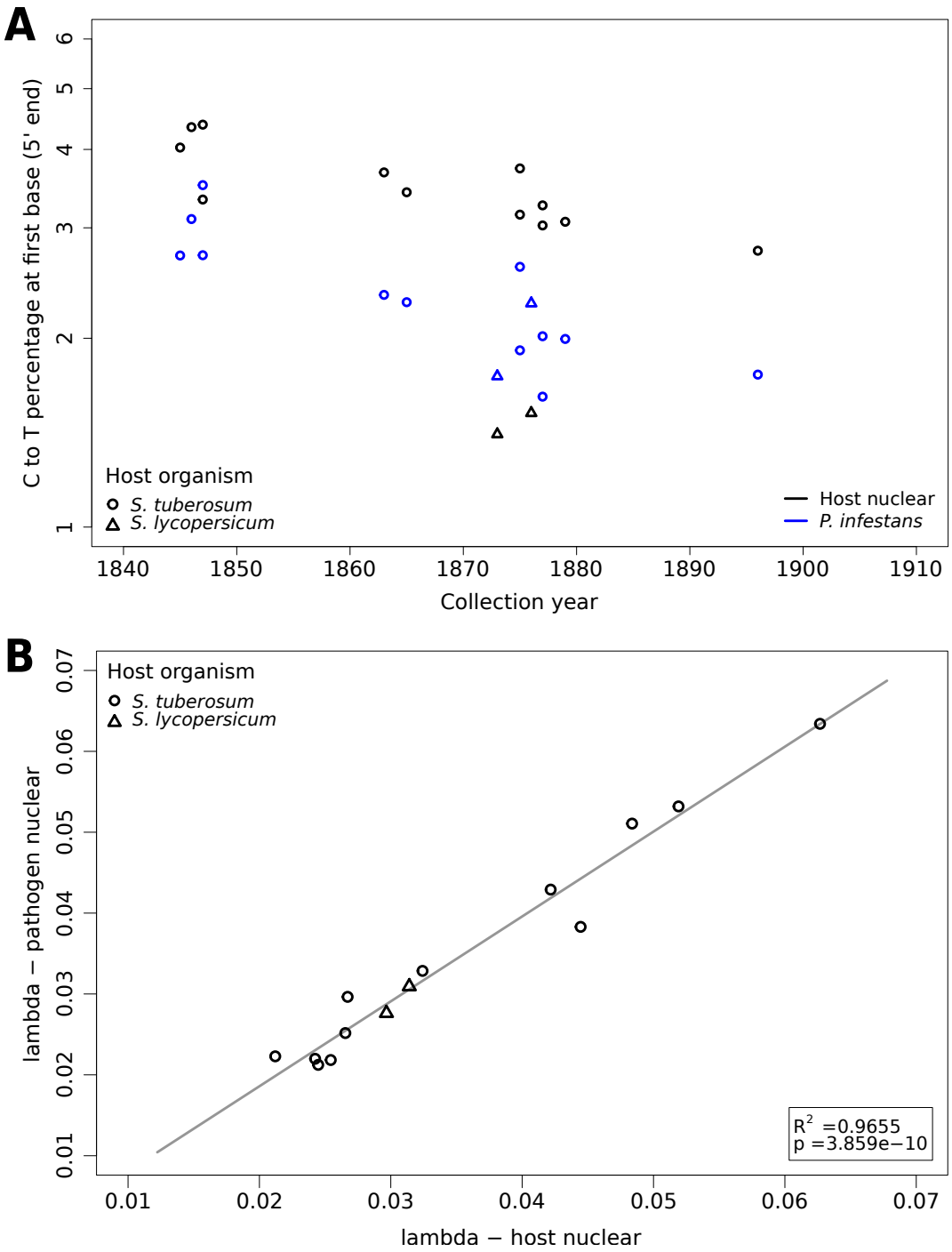

Figure 2.7: Correlation of damage patterns between host and pathogen. **A.** C-to-T percentage at first base (5'-end) as a function of collection year, in a subset of samples infected with *Phytophthora infestans* (N=14), split into sequences aligned to the host plant and the pathogen genome. **B.** Correlation of the damage fraction per site (λ) between the nuclear genome of the host, and the nuclear genome of *Phytophthora infestans* (N=14). The line and inset show the linear regression.

2.3 Discussion

The DNA molecules from herbarium specimens investigated here show high levels of fragmentation, with sizes comparable to DNA extracted from animal remains which are up to several thousand years old. In contrast to data from animal remains[152], we found a significant exponential relation between fragment length and collection year, where more recent samples have longer DNA fragments. Being able to detect this relation could be caused by a higher signal-to-noise ratio due to lower levels of environmental variation experienced by herbarium samples.

DNA decay and degradation can be understood as a two-step process, with a first rapid phase where the damage is caused mainly by nucleases and digestion by microorganisms, and a second phase where the damage is driven by hydrolytic and oxidative reactions that occur at a much lower rate than the first phase[161]. The correlation between age and fragmentation might be the result of a process occurring in the second phase that can only be detected in samples that have experienced very similar environments, as is the case for herbarium samples.

The DNA decay rate in herbarium samples calculated here is about six times faster than the rate established in bones by a similar procedure[162]. A possible explanation are differences in the characteristics of the types of tissue analyzed. In bone, DNA is adsorbed to hydroxyapatite, which decreases the rate of depurination compared with free DNA[160]. Additionally hydroxyapatite binds nucleases[163], which further prevents DNA degradation, especially in the first rapid phase of DNA degradation. DNA in desiccated leaf tissue might be less protected and more exposed to enzymatic and chemical damages. Furthermore, the vast majority of herbarium samples are not mounted on acid-free paper. Acid-free paper was introduced only in the mid-20th century[164], which could have contributed to DNA degradation, as an acidic pH increases the rate of depurination in vitro[155].

The extend of depurination can also be directly assessed from library-based sequencing data, by querying the base in the "-1" position directly upstream of the 5'-end of fragments. Both purine bases were over-represented at this "-1" position in all libraries, but no correlation was found between this enrichment and the age of samples. This implies that the contribution of depurination to the fragmentation of the DNA backbone does not change through time.

The highly fragmented nature of DNA molecules extracted from herbarium tissue has implications for sequencing very old herbarium specimens going back to the 16th and 17th century. A better understanding of the kinetics of the fragmentation process is necessary for designing sequencing experiments of such specimens as there are methods optimized for retrieving ultrashort DNA molecules[165]. Additionally, the extend of DNA fragmentation requires care in the choice of sequencing strategy, since PCR-based approaches[68] and reduced representation methods based on DNA restriction[166] have shown inferior performance to library-based shotgun sequencing and targeted capture.

All historical libraries showed clear signs of deamination, detected by the excess of C-

to-T substitutions at the ends of molecules. The magnitude of this pattern was strongly correlated with the age of the samples, as has been found using animal remains[152]. Since the characteristic pattern of C-to-T substitutions has been found in ancient DNA extracted from many different tissues from very different climatic environments, including samples from permafrost[48], it has been proposed to use the presence of this pattern as an authentication criterion[39]. Authentication of ancient DNA is necessary since molecules are usually present at low concentrations and easily contaminated with surrounding modern DNA. The presence of the characteristic deamination pattern can then be used to distinguish modern DNA molecules from historical ones[167]. This is especially important when metagenomic taxa are discovered from ancient DNA, for example pathogens of the organism from which the DNA was extracted. If DNA from the taxon discovered is of historical origin, its damage patterns are expected to be highly correlated with those in DNA from the host species, as shown here for herbarium samples infected with the plant pathogen *Pythophthora infestans*.

2.4 Methods

2.4.1 Data availability

Previously published sequences derived from *Solanum tuberosum* and *Solanum lycopersicum* infected by *P. infestans* are deposited in the European Nucleotide Archive, with accession number PRJEB1877.

New DNA sequences are deposited in the European Nucleotide Archive, with accession number PRJEB9878.

2.4.2 DNA extraction, library preparation and sequencing

DNA extractions from historical herbarium samples were carried out in clean room facilities in all cases. The majority of the samples were extracted following the PTB extraction protocol[168] as previously described[72]. Samples from the Cornell Bailey Hortorium were extracted using the CTAB extraction protocol[168].

Illumina double indexed sequencing libraries[54,169] were prepared from each sample as previously described[72]. The excess of C-to-T substitutions associated with DNA damage and caused by deamination of cytosines[40] was not repaired in order to quantify the amount of damage present in samples of different ages.

Libraries were paired-end sequenced on the Illumina HiSeq 2000, HiSeq 2500 or MiSeq instruments.

For further details on these samples, see also Weiß et al. [148].

2.4.3 Read processing and mapping

Reads were assigned to each sample based on their indices. Adapters were trimmed using the program Skewer (v. 0.1.120) with default settings and the natively imple-

mented Illumina TruSeq adapter sequences[170]. Forward and reverse reads were merged using the program Flash (v. 1.2.11) with default settings, except for an elevated maximum overlap (100-150 bp depending on read length) to allow a more accurate scoring of highly overlapping read pairs[171]. Merged reads were mapped as single-end reads to their respective reference genomes: *Arabidopsis thaliana*[172,173], *S. tuberosum*[174], *S. lycopersicum*[175], *P. infestans*[176]. The mapping was performed using BWA-MEM (v. 0.7.10) with default settings[177]. PCR-duplicates were identified after mapping based on start and end coordinates and for every cluster of duplicate reads a consensus sequence was generated[151].

2.4.4 Analysis of DNA damage patterns

We analyzed the fragment length distributions of merged reads. We fitted a lognormal distribution to the empirical fragment length distributions using the `fitdistr` function from the package `MASS`[178] using `R`[179]. Since in a lognormal distribution the logarithm of a variable is normally distributed, we used the mean of this distribution (log-mean) to summarize the fragment length distribution. The regression on the log-mean of fragment lengths and the sample collection year was carried out using the `lm` function in `R`. For visualization, we used the fragment length median on a log-scaled y-axis, since the median is more intuitive to understand than the log-mean value. The relationship between log-mean and median follows the formula: $median = e^{log-mean}$.

To analyze the genomic context around DNA break points, we used the software mapDamage 2.0 (v. 2.0.2–12)[180]. mapDamage calculates the genomic base frequencies around mapped reads and within reads. We calculated the relative enrichment of either adenine or guanine at the 5'-end (position "-1" compared with position "-5"). The frequencies of both adenine and guanine were extracted from the output file dnacomp.txt produced by mapDamage. The regression on the relative enrichment of purines (either adenine or guanine) and the sample collection year was carried out using the `lm` function in `R`.

All types of nucleotide misincorporations relative to the reference genome were calculated per library using mapDamage 2.0 (v. 2.0.2–12)[180]. The percentage of C-to-T substitutions at first base was extracted from the output file 5pCtoT_freq.txt produced by mapDamage. The regression on the percentage of C-to-T substitutions at first base (5'-end) and the sample collection year was carried out using the `lm` function in `R`. For the regression we used the percentage of deamination at first base. The whole procedure was carried out also for sequences aligned to the nuclear genome of the pathogen *P. infestans* in the subset of infected samples.

2.4.5 Analysis of covariance

To test if the regressions between host- and pathogen-derived reads were significantly different, we performed an analysis of covariance. We used the `aov` function in `R` to test

models where the sample age was the covariate and the DNA origin (host- or pathogen-derived) was the factor. In the first step, a model of type "$y \sim covariate \times factor$" was used to include a possible interaction between covariate and factor, which would mean that there is a difference in the slope of the regression depending on the factor. If no significant interaction was detected, the `anova` command in `R` was used to test this model against a model of type "$y \sim covariate + factor$". This last model does not include the interaction, therefore we can test whether the removal of the interaction has an effect on the fit of the model. If not, the second model was accepted with the conclusion that the regressions do not differ in slope, but possibly in their intersects (if there is a significant effect of the factor on the dependent variable y).

3 Ancient DNA metagenomics

Contributions

Parts of the content of this chapter have also been published in the articles "Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data"[181], and "Mining ancient microbiomes using selective enrichment of damaged DNA molecules"[182].

The following collaborators from the first article have contributed to the work presented in this chapter: Hernán Burbano (HB), Kay Prüfer and myself conceived and designed the project, and wrote the article with help from all authors. HB, Michael Dannemann and myself developed the statistical methods. I analyzed the data, implemented the statistical methods and generated all figures.

The following collaborators from the second article have contributed to the work presented in this chapter: HB, Matthias Meyer and myself conceived and designed the project. HB and myself wrote the article with help from all authors. I analyzed the data and generated all figures. Marie-Theres Gansauge and Ayinuer Aximu-Petri generated the sequencing data.

3.1 Authentication of aDNA

In addition to degradation processes of endogenous DNA from ancient and historical specimens, a major difficulty of working with ancient DNA is the danger of contamination with exogenous DNA. Therefore, special sample preparation procedures have been developed to reduce DNA contamination[8,9]. Nevertheless, it remains difficult to estimate how well preventive measures work. If contamination is a possible explanation for a result, it is crucial to exclude this possibility by giving positive evidence for the authenticity of aDNA[39,183].

One of the most typical features of aDNA is the presence of uracils (Us) that originate from post-mortem deamination of cytosines (Cs), especially in single-stranded overhangs at molecule ends[38]. Uracils are read as thymines (Ts) by most DNA polymerases, which generates a characteristic increase in C-to-T substitutions at the end of aDNA sequences ([38], Figure 3.1A). The presence of such C-to-T substitutions can be used as evidence for the authenticity of DNA sequences retrieved from historical material[39,183].

DNA retrieved from historical or ancient samples is an often complex mixture of molecules that contains DNA from organisms that were present ante-mortem or colonized the tissue post-mortem[26]. Therefore, all ancient DNA shotgun sequencing projects are metagenomic in nature. While sequencing endogenous ancient DNA from plant and

animal remains is a major focus of the field[68,184], a growing number of studies are now interested in the identification and characterization of ancient pathogens and microbiomes[144]. Genome sequences from ancient microbes permit the replacement of indirect inferences about the past with direct observations of microbial genomes through time. In the pathogen field, it has been possible to identify causal and/or associated agents of historical plant and animal disease outbreaks, as well as their spreading patterns through space and time[72,138].

In addition to studying individual aDNA genomes from pathogens and other microbes, it is also possible to reconstruct ancient microbiomes from ancient DNA extracts. For example, dental calculus from human and hominin remains allows the reconstruction of ancient oral microbiomes[140,142], while DNA extracted from lake sediments can inform the reconstruction of paleo-environments[145].

Whether the object of study is the endogenous DNA of a historical specimen, an associated microbe, or an ancient microbiome, a crucial step in the analysis of ancient DNA from any source is the distinction between bona fide ancient sequences and sequences of recent origin. The burden of proof of authenticity lies with the researcher, as inferences from ancient DNA generally rely on the temporal context of the sample that DNA was extracted from. It is therefore imperative that positive evidence of authenticity is provided.

Most commonly, this is done by investigating signatures of age-associated degradation, such as short fragment sizes and characteristic substitution patterns due to deamination[39]. Since library-based sequencing allows to reconstruct sequences of thousands of extracted DNA molecules[38,185], this type of evidence is often straight forward to provide. However, increasingly complex mixtures of DNA molecules may complicate this process, as DNA molecules of interest get highly diluted[27]. If target sequences are known, shotgun sequencing can be complemented by targeted hybridization capture to enrich fragments of interest[44,45,138]. When taxa are discovered de novo, it can be more challenging to provide evidence of an authentic historical origin. In this chapter, we present two case studies on how evidence of authenticity can be provided in such difficult cases.

3.2 Taxonomic binning of aDNA sequencing data

One of the advantages of library-based shotgun sequencing over targeted sequencing approaches is a faithful representation of the composition of DNA extracted from a sample[26]. This makes it perfectly suited for meta-taxonomics - the taxonomic characterization of a collection of genomes from different organisms, which is also known as the metagenome of a sample[186].

To be able to associate single sequences to taxonomic groups, one commonly depends on reference sequences linked to a taxonomic tree. The metagenomic sequences are then queried for their similarity to these reference sequences, and the classification of sequences to a taxonomic group is based on this similarity[137].

Sequence similarity can be assessed by pairwise sequence alignments[187,188,189]. For taxonomic classification of entire metagenomic datasets, this is very computationally

expensive, with datasets of more than hundreds of millions of sequences and reference databases on the order of tens of millions of sequences (e.g. NCBI BLAST nt database).

To reduce the computational burden of taxonomic classification, alignment-free methods have been developed as well, for example those based on shared k-mers[190]. Any sequence of length l consists, at most, of $L - k + 1$ overlapping words of size k. Because there are 4^k possible DNA sequences of size k, it is unlikely even for moderate sizes of k that two unrelated sequences would share a k-mer simply by chance. k-mers can be easily counted through techniques such as hashing[191], and counts can be easily compared. This makes it a great method to assess sequence similarity without the requirement for sequence alignment.

However, the comparison of k-mer counts relies on exact matches of k-mers. A single substitution can affect the counts of multiple k-mers, depending on its position in the sequence. This makes the method less suited for applications in ancient DNA, where age-associated degradation patterns such as deamination affect the k-mer profile of query sequences.

Another characteristic of ancient DNA are highly fragmented DNA molecules, leading to short sequences. This also has an effect on k-mer based methods, as each sequence contains less k-mers on which a k-mer sharing profile can be build. It does however also affect certain alignment-based methods intended to reduce computational cost, for example methods based on blastx[188].

Blastx is a method for aligning nucleotide sequences, such as full metagenomes, to protein databases by translating the nucleotide sequences in-silico. Certain characteristics of aligning protein sequences allow significant speed-ups of this process[192]. However, for ancient DNA the in-silico translation means that already short nucleotide sequences become even smaller protein sequences (as codons of three nucleotides code for one amino acid). In addition, protein databases cover by definition only protein coding regions of genomes. As ancient DNA libraries are often of low complexity and low coverage, this reduces their probability of classification even further.

With these drawbacks, the specific characteristics of ancient DNA make nucleotide sequence alignments the best method by which to classify ancient DNA metagenomes[193,194]. Blastn greatly reduces the computational cost of aligning sequences to large databases of reference genomes. Still, it is not practical for aligning many millions of short sequences. More recent implementations, which load entire databases into main memory to avoid disk access, now allow taxonomic classification of large metagenomes within hours[195,196,197].

Once all alignments of a query sequence to a set of reference sequences (above a certain quality threshold) are recorded, the sequence has to be classified taxonomically. The sequences present in the reference database are generally classified to at least the species level (or below, e.g. strain, subspecies etc). This means, that reference data is available only for the leaves of the taxonomy. If a query sequence has alignments exclusively to one such leaf node, its classification is trivial and the sequence is assigned to this node. However, often times a sequence will align to several reference sequences. One way this might happen is due to a taxonomic group being highly represented in the database. If, for instance, many genomes of bacterial subspecies of a given species are

in the database, they will often be highly similar across large parts of their genomes. A query sequence from such a region will align to several subspecies genomes with similar quality.

A common way to address this problem is by a lowest common ancestor (LCA) algorithm[137]. Here, the taxonomic assignment of a sequence is based on the lowest node in the tree connecting all nodes that the sequence aligns to. In the example above, this might be the node of the bacterial species represented by many subspecies.

Another way by which a sequence might align to several, even more distant nodes in the taxonomic tree is if it originates from a highly conserved sequence shared by many genomes in the reference database. In this case, the query sequence has little information content for the taxonomic profile of the metagenome, and will be assigned to a higher taxonomic level by the LCA.

Any method reliant on reference databases will of course be heavily dependant on the completeness of such databases, and affected by biases like the uneven representation of taxonomic groups[144,193,194]. However, even with incomplete, biased databases, the taxonomic classification of the inherently metagenomic mixture of DNA sequences present in extracts from ancient and historical samples can inform further analysis. This is important for example in the discovery of ancient microbes and infections[72,138,139], as well as in the investigation of ancient metagenomes from sources such as archaeological sediments[145,198]. Still, it is important that shotgun sequencing is complemented by other, more targeted methods, and that care is taken in the interpretation of potentially spurious alignments as well as in the authentication of a samples' historical origin[139,145,194].

3.3 Case Study 1: DNA authenticity at low coverage

3.3.1 Introduction

Although a vast proportion of flora and fauna does not fossilize, traces of their DNA may be preserved in sediments allowing the characterization of past biodiversity[146,198].

In 2015, Smith et al. [199] presented DNA sequences generated from sedimentary DNA extracts from Bouldnor Cliff, a submerged archeological site in the United Kingdom, which suggested the presence of domesticated wheat 8,000 years ago. This would push back the arrival of wheat to the British Isles by about 2,000 years compared to expectations based on archeological remains, and is 400 years earlier than in nearby European sites (reviewed in Smith et al. [199]). Since Smith et al. [199] did not find wheat pollen or archeological remains associated with wheat cultivation, they conclude that the wheat presence in Bouldnor Cliff was the result of trading.

In total they produced ~72 million Illumina reads, of which they robustly assigned 152 sequences to wheat (genus *Triticum*), with dozens more (160 reads) to higher taxonomic ranks that include wheat. Smith et al. [199] took state-of-the art preventive measures to avoid contamination and exercised great effort to ensure the accuracy and robustness of their phylogenetic assignments. The authors attempted to authenticate the aDNA molecules based on the expected excess of C-to-T substitutions, but because of the very small number of reads assigned to wheat, they failed to do so using standard approaches. As a result of that, the authors did not present any positive evidence for the ancient origin of recovered DNA molecules. Here we present an approach that compares the pattern of C-to-T substitutions in a set of test reads with the distributions of C-to-T substitutions in reads from known ancient- and modern-DNA and apply this approach to the sedimentary DNA from Smith et al. [199].

3.3.2 Methods

3.3.2.1 Read processing for ancient and modern DNA samples

Sequencing data from most samples was downloaded from the European Nucleotide Archive (Table 3.1), with the exception of the *Gorilla gorilla* sequences which were provided directly by the authors[152]. Adapters were trimmed for both paired- and single-end runs using the program Skewer (version 0.1.120) using default parameters[170]. For paired-end runs (Table 3.1) forward and reverse reads were merged requiring a minimum overlap of 10 base pairs (bp) using the program Flash (version 1.2.11)[171]. Merged or single-end reads were mapped as single-end reads against their respective nuclear or organellar genomes: *Solanum tuberosum* nuclear genome[174], Solanum lycopersicum nuclear genome[175], *Triticum aestivum* nuclear genome[200], *Gorilla gorilla* mitochondrial genome[201], *Homo sapiens* nuclear genome[202]. The mapping was carried out using BWA-MEM (version 0.7.10) with default parameters, which include a minimum read length of 30 bp[177]. PCR duplicates were removed after mapping using bam-rmdup (available at `https://github.com/udo-stenzel/biohazard`), which computes a consensus sequence for each cluster of duplicated sequences. Alignments were stored in BAM format[89].

3.3.2.2 Read processing for sedimentary DNA from Smith et al. [199]

We used two different approaches to process the reads from sedimentary DNA[199].

Curated reads: we used a set of 152 reads assigned to tribe Triticeae and to genus *Triticum* by Smith et al. [199] after phylogenetic curation. However, we consider the complete sequence and do not exclude the initial 10 nucleotides as was done in the original processing[199]. Reads were then aligned to the wheat genome as described above.

All sedimentary DNA reads: we aligned reads from all four layers sequenced by Smith et al. [199] to the *T. aestivum* nuclear genome[200]. Duplicates were removed and only alignments with mapping quality greater or equal than 30 were used for further analysis. Additionally, we include a sequence complexity filter based on entropy, which removed low complexity reads with entropy less or equal to 50. The entropy filtering was carried out with prinseq-lite (version 0.20.4)[203].

3.3.2.3 Exponential function fitting and calculation of goodness-of-fit p-value

For each set of aligned reads (complete libraries or subsamples) the C-to-T substitution patterns along the 5'-end of reads were assessed using the program PMDtools[167]. We fitted an exponential function to the frequency of C-to-T substitutions for the first 20 nucleotides at the 5'-end. The fitting was performed in `R` using the `nls` function, which determines the nonlinear least squares estimate of the parameters in a nonlinear model. The fitted exponential follows: $y \sim N \cdot e^{-rate \cdot x}$. From the `nls` fitting we obtained the t-value and degrees of freedom for the *rate* parameter and then calculated a goodness-of-fit p-value by using a one-sided t-test.

3.3.2.4 Generation of empirical distributions of goodness-of-fit p-values

Subsets of different alignment numbers were randomly sampled (with replacement) 10,000 times from alignments stored in BAM format. The random sampling was performed using samtools[89]. For every subset of alignments we assessed the fraction of C-to-T substitutions, fitted an exponential function and calculated a goodness-of-fit p-value as explained above.

3.3.2.5 Calculation of test empirical p-value

Curated reads: we compared the goodness-of-fit p-value of the test set of 152 sedimentary DNA reads with distributions of goodness-of-fit p-values generated from bona fide modern and ancient DNA. For the distribution of goodness-of-fit p-values from aDNA, we count how many of them are equal or greater than the sedimentary DNA goodness-of-fit p-value. To calculate the empirical p-value of the test we subsequently divided this number by the total number of values in the empirical distribution. With this approach we test the null hypothesis that the test set of reads contains a signal of ancient DNA damage that is comparable or even more pronounced than the signal in the aDNA library used to generate the empirical distribution of goodness-of-fit p-values.

For the distribution of goodness-of-fit p-values from modern DNA, we count how many of them were smaller than or equal to the sedimentary DNA goodness-of-fit p-value. We calculate the empirical p-value of the test by dividing this number by the total number of p-values in the empirical distributions. With this approach we test the null hypothesis that the test set of reads matches the absence of ancient DNA damage patterns seen in reads of modern origin.

All sedimentary DNA reads: We tested alignments from each of the layers sequenced by Smith et al. [199] using a bona fide aDNA sample for the generation of the distribution of goodness-of-fit p-values. For each layer we tested 10 sets of different numbers of reads (from 100 to 1,000 reads, with increments of 100 reads). For each layer and for each number of reads in the test set we repeated the test and calculated the empirical p-value 1,000 times as described above.

Other ancient DNA and modern DNA libraries were tested using the same procedure.

3.3.2.6 Stand-alone implementation of the sampling test

In addition to the workflow based on third party tools presented above, we also implemented this test in a single program to improve usability and reduce dependencies. The tool is implemented in the `C` programming language, and uses the library `htslib` for reading BAM files, as well as the GNU Scientific Library (`GSL`). It only requires a BAM file with sequence alignments to a suitable reference as input. From this BAM file, C-to-T substitution patterns in the first 20 bases from the 5'-end are extracted. Optionally, a specified fraction or number of sequences is subsampled from the BAM file, and substitution patterns are extracted only from the sampled fraction.

We used these C-to-T frequencies along the first 20 bases, to fit an exponential decay function following the formula used also with `nls` in `R`: $y \sim N \cdot e^{-rate \cdot x}$. The non-linear least-squares minimization is performed by `GSL`, using the Levenberg-Marquardt algorithm as implemented in the `gsl_multifit_fdfsolver_lmsder` solver.

This returns the best-fit parameter vector $\hat{\beta}$ of the estimates for N and $rate$. From the final estimates, we assess the goodness-of-fit using the t-value of the $rate$ parameter estimate $\widehat{rate}$. This t-value is calculated by dividing the estimate by its standard error:

$$t_{\widehat{rate}} = \frac{\widehat{rate}}{s.e.(\widehat{rate})}$$

Assessing the standard errors of the parameter estimates $\hat{\beta}$ requires scaling the variance of the estimate itself by the variance of the residuals about the best fit. The variance of the estimates can be retrieved from the covariance matrix of the best-fit parameters, which is calculated from the Jacobian matrix J:

$$covar = (J^T J)^{-1}$$

where the diagonal elements of the $covar$ matrix represent the variance of the parameter estimates. In this calculation, $J^T J$ estimates the Hessian matrix at the parameters

providing the best fit (ignoring higher order terms)[204], the inverse of which is the covariance matrix[205].

The residual variance is calculated by dividing the sum of squared residuals by the degrees of freedom:

$$\sigma^2_{res} = \frac{\sum(y_i - Y(x_i, \hat{\beta}))^2}{dof}$$

where $dof = n - p$, i.e. the number of estimated parameters p subtracted from the number of data points n. The quantity $Y(x_i, \hat{\beta})$ represents the prediction for the dependent variable from its corresponding regressor x_i and the best-fit parameters $\hat{\beta}$.

The final standard errors of the parameter estimates are then calculated as:

$$s.e.(\hat{\beta}) = \sqrt{\sigma^2_{res} diag(covar)}$$

In our test, the t-value of $\widehat{rate}$ is used in a one-sided t-test to get the goodness-of-fit p-value. The program and its source code are openly available at `https://github.com/clwgg/ugat`.

3.3.3 Results and Discussion

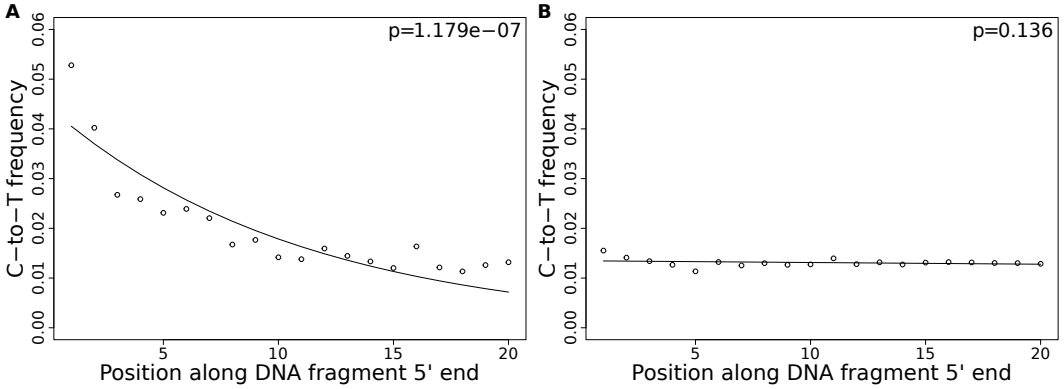

Figure 3.1: Patterns of C-to-T substitutions at the 5'-end of ancient and modern DNA. Shown are the frequency of cytosine to thymine (C-to-T) substitutions at the first 20 bases, the fit of an exponential distribution, and the goodness-of-fit p-value of an exponential decay (assessed from the decay rate parameter). **A.** Data from a historical *Solanum tuberosum* herbarium specimen (ancient DNA). **B.** Data from a modern *Triticum aestivum* sample (modern DNA).

Although the excess of C-to-T substitutions at the 5'-end occurs at different magnitudes in samples of different ages, the exponential increase of substitutions towards the end is a ubiquitous pattern in aDNA studies[152]. To capture the presence of this pattern in various datasets, we fitted an exponential function and evaluated the goodness of fit

by using a one-sided t-test on the rate parameter to test for significant exponential decay. As expected, true aDNA libraries show significant goodness-of-fit p-values (Figure 3.1A), whereas non-significant goodness-of-fit p-values, neither decay nor growth, are observed in libraries derived from modern DNA (Figure 3.1B). A given C-to-T damage pattern plot can thus be summarized by its goodness-of-fit p-value that, when significant, indicates exponential decay of C-to-T substitutions at the 5'-end (Figure 3.1A).

We resampled (with replacement) 10,000 sets of 150 sequences from a library of historical *Solanum tuberosum* collected in 1846[72]. The number was selected to be comparable to the 152 reads that Smith et al. [199] assigned to wheat. An empirical distribution of goodness-of-fit p-values was generated by performing the goodness-of-fit test for each subsample (Figure 3.2). When we evaluate the goodness-of-fit p-value of the sedimentary derived sequences assigned to wheat, we find that it falls within the upper 3% of subsamples with the least good fit. We can therefore reject the null hypothesis that the sequences assigned to wheat are as deaminated as the historical *S. tuberosum* library. We repeated the procedure using a modern wheat library to generate the distribution of goodness-of-fit p-values (Figure 3.2) and find a better match ($p = 0.83$). Thus, we cannot reject the hypothesis that the sequences assigned to wheat are of modern origin.

We sought to investigate how the test behaves when the empirical distribution of goodness-of-fit p-values is generated from different aDNA libraries. For this purpose we used a set of samples from animal[152] and plant remains[72] with an age range of 85-170 years before present, and scored the sedimentary wheat sequences against distributions generated from these libraries (again with subsamples of 150 sequences). We observed that the goodness-of-fit p-value for these libraries is positively correlated with the empirical p-value of the sedimentary wheat sequences tested against them (Figure 3.3). Using a significance level of 0.05, we rejected the hypothesis that the wheat sequences are of ancient origin with 7 out of 13 libraries used in our test (Figure 3.3). Thus, the purportedly 8,000-year old wheat sequences show a less pronounced deamination pattern than many plant and animal samples less 200 years old.

Next, we took a less conservative approach and scored the sedimentary wheat sequences against a distribution of goodness-of-fit p-values (subsamples of 150 reads) generated from a 7,000-year-old human Mesolithic sample from la Braña in Northern Iberia[206]. La Braña is a site with cold environment and stable thermal conditions that has yielded exceptionally well conserved human fossils with ~50% of human endogenous DNA that reach a C-to-T substitution rate of about 15% at the 5'-end[206]. We could reject the null hypothesis that the 152 sedimentary wheat sequences are as deaminated as the sample from la Braña ($p = 0.0014$), a sample that is closer in time to the allegedly 8,000-year-old wheat reads (Figure 3.4). It is worth pointing out that almost all 10,000 subsamples from la Braña had a very low (close to 0) goodness-of-fit p-value, even though we subsample only 150 reads.

To test the number of sequences needed to reliably assess the presence of deamination patterns, we tested both an aDNA (Figure 3.5A) and a modern DNA library (Figure

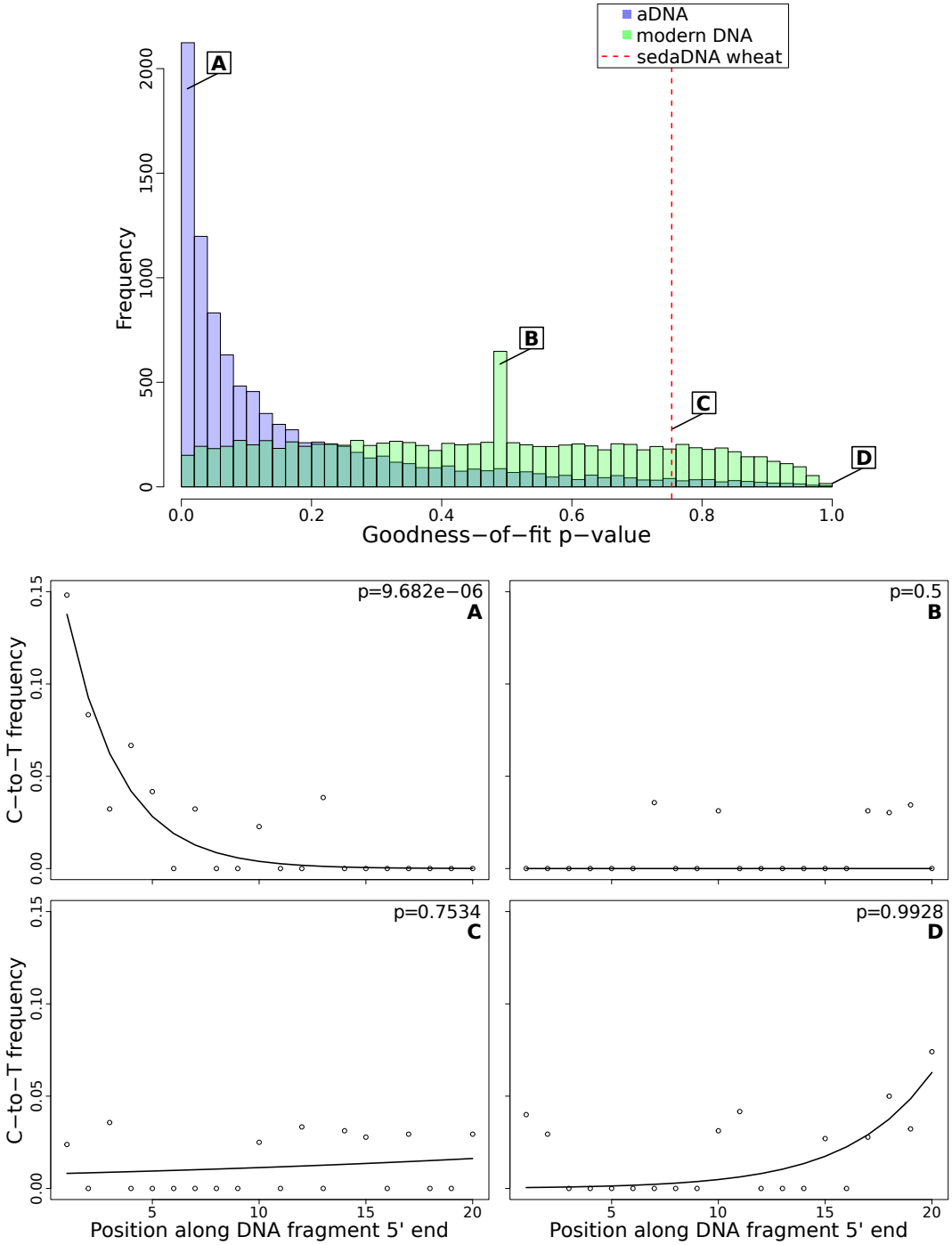

Figure 3.2: Authenticity test of DNA sequences assigned to Triticum by Smith et al. [199]. The histograms on top show the empirical distributions of goodness-of-fit p-values of subsamples of 150 reads from ancient and modern DNA (same libraries as in Figure 3.1). The dotted red line indicates the location of the goodness-of-fit p-value from reads assigned to wheat in sedimentary ancient DNA[199]. The panels **A**-**D** show cytosine to thymine (C-to-T) substitutions at the 5'-end at different points of the goodness-of-fit p-value distributions, and from the reads assigned to wheat in sedimentary ancient DNA.

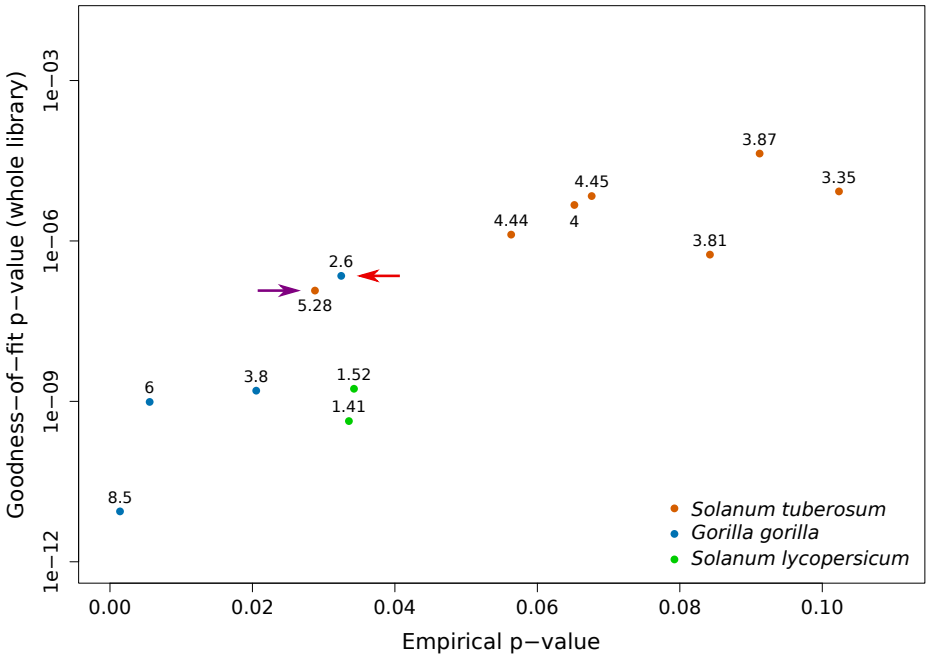

Figure 3.3: Assessment of empirical p-values across samples. Variation of the empirical p-value of the test depending on the goodness-of-fit p-value of the whole library used to generate the empirical distribution. Numbers adjacent to the points indicate the percentage of C-to-T substitutions at first base. Red arrow indicates the aDNA library used as test in Figure 3.5A. Purple arrow indicates the library used to generate the empirical distribution of goodness-of-fit p-values in Figure 3.5A-C.

3.5B) against a distribution built from a bona fide aDNA library, while varying the number of sampled sequences. While p-values testing the hypothesis that a true aDNA library is ancient were distributed uniformly (Figure 3.5A), the hypothesis that a modern library has ancient origin could be rejected when sufficient number of sequences were sampled (in tests with more than 300 reads the median empirical p-value was always below 0.05) (Figure 3.5B).

Finally, we skipped the phylogenetic curation step applied by Smith et al. [199] and mapped all sedimentary sequences to the wheat genome. After stringent filtering of these mappings, we repeated our test varying the subsample size from 100 to 1,000 sequences. The empirical p-value was dependent on the number of reads tested, and declined with an increasing number of tested reads for all layers of sediments sequenced by Smith et al. [199] (Figure 3.5C). This pattern resembled the one obtained from a modern DNA library (Figure 3.5B). As for the phylogenetically curated 152 sequences, we were able to reject the hypothesis that the mapped reads are as deaminated as the test library (mean p-value < 0.05 for all tests with more than 400 reads for layers 1–2 and 4, and 800 reads for layer 3). This analysis also shows that the 152 sequences after phylogenetic cu-

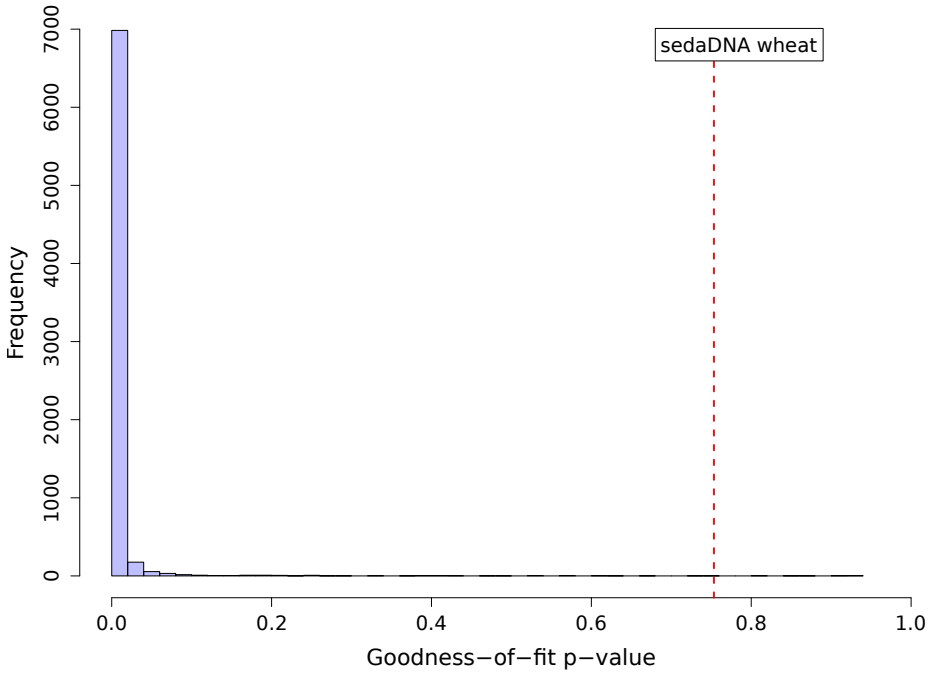

Figure 3.4: Authenticity test of DNA sequences assigned to Triticum by Smith et al. [199]. The histogram shows the empirical distribution of goodness-of-fit p-values of subsamples of 150 reads from an ancient DNA library. In contrast to Figure 3.2, the library used to construct the empirical distribution is of a similar age as the proposed wheat DNA. The dotted red line indicates the location of the goodness-of-fit p-value from reads assigned to wheat in sedimentary ancient DNA[199].

ration are not a biased subsample from the distribution of all wheat-matching sequences.

In summary, we were able to reject the hypothesis that the sequences assigned to wheat by Smith et al. [199] are as deaminated as a wide range of known ancient DNA libraries. This is true even when we compared the putative 8,000 year old sequences with only century old samples that show low deamination signatures. This means that a scenario in which wheat was transported to the Bouldnor Cliff site 8,000 years ago is unwarranted. Our approach for authentication of aDNA can be used even with a very small number of sequences, and we hope that it will proof useful to test for positive evidence of authenticity for ancient DNA studies whose conclusions rely heavily on the ancient origin of analyzed sequences.

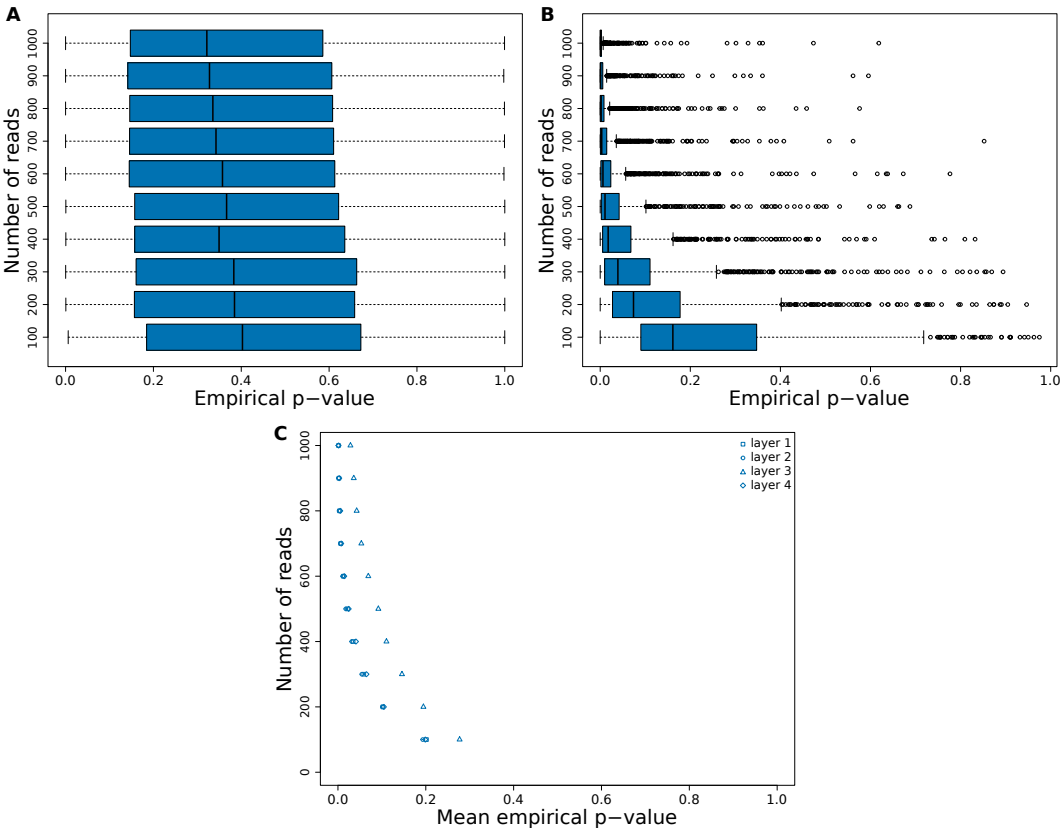

Figure 3.5: Evaluation of test performance. Box-plots are based on 1,000 tests. **A.** Variation of the empirical p-value of the test depending on the number of reads sampled from an ancient DNA library (indicated with red arrow in Figure 3.3). **B.** Variation of the empirical p-value of the test depending on the number of reads sampled from a modern *Triticum aestivum* library (same library used to generate the distribution of empirical goodness-of-fit p-values in Figure 3.2). **C.** Variation of the empirical p-value of the test depending on the number of reads sampled from sedimentary ancient DNA mapped directly to the *T. aestivum* genome. Layers as reported in Smith et al. [199].

Table 3.1: Datasets used in Section 3.3.

Species	Type of DNA	Age	Ref.	Study ID	Sample ID	Library
Metagenomics	Sedimentary	8030-7908[1]	[199]	PRJEB6766[2]	ERR567364[2]	Single end
Metagenomics	Sedimentary	8030-7908[1]	[199]	PRJEB6766[2]	ERR567365[2]	Single end
Metagenomics	Sedimentary	8030-7908[1]	[199]	PRJEB6766[2]	ERR567366[2]	Single end
Metagenomics	Sedimentary	8030-7908[1]	[199]	PRJEB6766[2]	ERR567367[2]	Single end
Metagenomics	Sedimentary	8030-7908[1]	[199]	PRJEB6766[2]	ERR732642[2]	Single end
T. aestivum	Modern	NA	[207]	PRJNA250383[2]	SRR1170664[2]	Single end
S. tuberosum	Ancient	135[3]	[72]	PRJEB1877[2]	ERR267886[2]	Paired end
S. tuberosum	Ancient	137[3]	[72]	PRJEB1877[2]	ERR267882[2]	Paired end
S. tuberosum	Ancient	149[3]	[72]	PRJEB1877[2]	ERR330058[2]	Paired end
S. tuberosum	Ancient	165[3]	[72]	PRJEB1877[2]	ERR267872[2]	Paired end
S. tuberosum	Ancient	166[3]	[72]	PRJEB1877[2]	ERR267868[2]	Paired end
S. tuberosum	Ancient	166[3]	[72]	PRJEB1877[2]	ERR957324[2]	Paired end
S. tuberosum	Ancient	167[3]	[72]	PRJEB1877[2]	ERR267868[2]	Paired end
S. lycopersicum	Ancient	136[3]	[72]	PRJEB1877[2]	ERR267884[2]	Paired end
S. lycopersicum	Ancient	139[3]	[72]	PRJEB1877[2]	ERR267878[2]	Paired end
G. gorilla	Ancient	83[3]	[152]	NA	107[4]	Paired end
G. gorilla	Ancient	100[3]	[152]	NA	109[4]	Paired end
G. gorilla	Ancient	100[3]	[152]	NA	110[4]	Paired end
G. gorilla	Ancient	103[3]	[152]	NA	114[4]	Paired end
H. sapiens	Ancient	7000[1]	[206]	PRJNA230689[2]	SRR1045127[2]	Single end

[1]B.P. (before present years).
[2]IDs from the European Nucleotide Archive.
[3]Calculated from collection date (in years).
[4]IDs from Sawyer et al. [152].

3.4 Case Study 2: Uracil enrichment for taxon discovery

3.4.1 Introduction

In cases where it is difficult to assess the authenticity of generated sequences, it is also possible to utilize experimental procedures which enrich the DNA mixture in molecules of ancient or historical origin.

Recently, a single-stranded library preparation method (U-selection) was developed, which allows physical separation of uracil-containing molecules from non-deaminated ones[208]. During U-selection, all library molecules are initially immobilized on streptavidin beads, to which molecules without uracils remain attached (U-depleted fraction), while uracil-containing molecules (originally deaminated) are released into solution (U-enriched fraction). U-selection was originally developed with the aim of increasing the amount of ancient hominin DNA (e.g. Neanderthals) from a background of present-day human and microbial DNA[208]. However, the method seems to be especially suited to study ancient microbiomes, due to the inherent difficulty to authenticate their ancient origin.

This complication arises from microbes colonizing tissues at different times, which results in different levels of deamination of microbial DNA in historical samples. Although sequences that carry terminal C-to-T substitutions can be selected in silico[49,209], there are two factors that could hinder this approach. First, low levels of deamination will reduce the number of molecules suitable for selection in silico. Second, high sequence divergence between samples and reference genomes can mask age-associated deamination signals thereby hindering authentication. Consequently, enriching for deaminated molecules during library preparation can be fundamental to tackling these problems. As a proof-of-principle experiment, we used U-selection in combination with taxonomic binning of sequences to characterize the microbiomes of Neanderthal bones (~39,000 years old), herbarium specimens (between 41 and 279 years old) and plant archaeological remains (~2,000 years old) (Table 3.2).

3.4.2 Methods

3.4.2.1 Sequencing libraries from Neanderthal remains

We used sequencing libraries from Neanderthal samples (Table 3.2) prepared by[208], which were sequenced deeper for this study.

3.4.2.2 DNA extracts from historical plant samples

Previously published DNA extracts from four different plant species were used for this study. These extracts were derived either from herbarium specimens (*Arabidopsis thaliana*[136], *Solanum tuberosum*[72], *Solanum lycopersicum*[72]), or from archaeobotanical remains (Zea mays[67]). Ages ranged from 41 to 279 years for herbarium samples, and 1852 to 1881 years for Zea mays samples (Table 3.2). The sequencing data for these samples is available on the European Nucleotide Archive under study number PRJEB30666.

3.4.2.3 Sequencing libraries from historical plant samples

Three sequencing libraries were produced for each plant DNA extracts, one using a double-stranded library preparation[53,169] without enzymatic removal of uracils[42], and one for each fraction resulting from the single-stranded uracil enrichment protocol[208] .

3.4.2.4 Sequencing and initial data processing

Since aDNA molecules are in most cases shorter than the read length of the sequencing platform, it is possible that parts of the aDNA molecule is sequenced by both the forward and reverse read, and also that parts of adaptors are sequenced[151]. Therefore, it is recommended to merge sequences based on the overlapping fraction sequenced by both forward and reverse reads[151]. We removed adaptors and merged sequences using the software leeHom with the "--ancientdna" option[210]. Putative chimeric sequences were flagged as low quality.

3.4.2.5 Mapping of sequenced reads to their host genome

Merged reads were mapped as single-ended reads to their respective host genome: *Zea mays*[211], *Arabidopsis thaliana*[172,173], *Solanum tuberosum*[174], *Solanum lycopersicum*[175], *Homo sapiens*[202]. The mapping was performed using BWA-MEM (version 0.7.10) with default parameters, which includes a minimum length cutoff of 30 bp[177].

3.4.2.6 Metagenomics assignment of sequenced reads

Reads were aligned to the full non-redundant NCBI nucleotide collection (nt) database (downloaded January 2015) using MALT (version 0.0.12,[195]) in BlastN mode. The resulting RMA files were analyzed using MEGAN (version 5.11.3,[137]). The reads were assigned to the NCBI taxonomy using a lowest common ancestor algorithm[137].

3.4.2.7 Mapping of sequenced reads to microbial genomes

Libraries were mapped to microbial reference genomes of interest, after the presence of certain taxa was detected during metagenomic assignment. Specifically, the references of *Streptosporangium roseum*[212], *Pseudomonas syringae pv. syringae B728a*[213], *Pseudomonas rhizosphaerae*[214] and *Pantoea vagans*[215] were used. Since mapping metagenomic libraries to bacterial reference genomes is very prone to false alignments, we used a different mapping strategy for these genomes. The mappings were performed with bowtie2 (version 2.2.4,[216]), with the settings "--score-min 'L,-0.3,-0.3' --sensitive --end-to-end" to increase stringency.

3.4.2.8 Assessment of nucleotide substitution patterns

All types of nucleotide substitutions relative to the reference genome were calculated per library using mapDamage 2.0 (v. 2.0.2–12,[180]). The percentages of C-to-T substi-

tutions at the 5'-end were extracted from the output file 5pCtoT_freq.txt produced by mapDamage.

3.4.2.9 *Pantoea vagans* genomic variation

In order to reduce the effect of aDNA-associated C-to-T substitutions on variant discovery, we used exclusively the U-depleted fraction of libraries where *P. vagans* was detected in the metagenomic screening. The libraries were mapped to the *P. vagans* reference genome using BWA-MEM, to reduce reference bias and increase SNP discovery. False alignments from the metagenomic libraries posed a lesser problem here, as variants were ascertained based on modern material. Variants for historical samples were called for both libraries together using the bcftools (version 1.8,[217]) utilities mpileup ("bcftools mpileup -q 1 -I -Ou -f $REF $IN1 $IN2") and call ("bcftools call --ploidy 1 -m -O -z"). Additionally, 11 assemblies of different contiguity were downloaded from NCBI (`https://www.ncbi.nlm.nih.gov/genome/genomes/2707`). These assemblies were aligned to the reference genome using minimap2 (version 2.10-r764,[218]) and its "asm20" parameter preset. Only strains with at least 80% reference coverage were kept for subsequent analysis (9/11, average reference coverage: 91%). The paftools utility, which is distributed with minimap2, was used to call variants from these alignments, with the parameter set "-l 2000 -L 5000". All resulting VCF files from modern samples were merged using bcftools' merge utility with the parameter "--missing-to-ref", assuming that those positions not called by paftools in any one sample were indeed reference calls. The merged VCF from modern material was then merged with the VCF from the two historical samples using bcftools, and filtered to include only full information, biallelic SNPs. This approach discovers sites, which are segregating in modern material, and have read data (be it reference, alternative or segregating sites) in both historical samples. The resulting VCF file was loaded into `R` using vcfR (version 1.7.0,[219]), and a PCA was produced by converting the information into a genlight object using adegenet (version 2.0.1,[220]) in `R` (version 3.3.3,[179]).

3.4.2.10 *Pseudomonas* spp. assembly and evaluation

To evaluate the presence of *Pseudomonas* spp. strains in a *Solanum tuberosum* historical herbarium sample, we extracted from this library all reads that were taxonomically assigned to the *Pseudomonas* genus or to lower taxonomic levels within it. These reads were then assembled using SPAdes (version 3.5.0) with default parameters[221]. The resulting contigs were filtered for a minimum length of 2Kb, which yielded 3,314 contigs with a total length of 16Mb. We used the lastz (version 1.03.66,[222]) and Circos (version 0.64,[223]) interface of AliTV[224] to align these contigs to either the *P. syringae* or *P. rhizosphaerae* reference genome. We were able to align 72% of contigs to either one or both of these reference genomes in alignments of at least 1Kb. We then extracted all contigs which had alignments of at least 10Kb in length and were unique to one of the reference genomes. These sets of contigs were again aligned to their corresponding reference using AliTV as described above. Additionally, we used these uniquely aligning contigs to

assess their average kmer coverage during the assembly, as reported by SPAdes.

3.4.3 Results and Discussion

Our experiments were motivated by the previous observation that in some Neanderthal samples, e.g. from El Sidrón, Spain, the proportion of Neandertal DNA fragments remains unchanged in both the U-depleted and U-enriched fractions, whereas in others, e.g. from Vindija Cave, Croatia, this proportion increased in the Uracil-enriched fraction[208]. It was hypothesized that the latter effect could have been due to differences in deamination, and hence in age, between Neanderthal- and microbial-derived DNA fragments. To explore this effect further, we re-analyzed the previously generated Neanderthal sequence data from both sites by performing taxonomic binning of reads derived from the U-depleted and U-enriched fractions, instead of aligning them only to the human reference genome.

Reads aligning to the two most abundant bacterial phyla (Actinobacteria and Proteobacteria) from the Vindija Neandertals were enriched in the U-depleted fraction, while hominin reads were enriched in the U-enriched fraction (Figure 3.6A). This is in accordance with a previous study that reported absence of DNA damage in Actinobacteria derived from a Neanderthal bone from Vindija cave[225]. In contrast, in reads obtained from the El Sidrón Neanderthals, we found enrichment of both hominin and Actinobacteria reads in the U-enriched fraction, whereas Proteobacteria reads were enriched in the U-depleted fraction (Figure 3.6B). Overall, bacteria-derived reads were dominated by the Actinobacterium *Streptosporangium roseum* (Figure 3.6C), which showed almost 50% deamination at the first base in the U-enriched fraction (Figure 3.6D), suggesting an ancient origin. The analysis of reads derived from Neanderthal bones illustrates how U-selection permits distinguishing between ancient bacteria enriched in the U-enriched fraction and more recent colonizers enriched in the U-depleted fraction.

In order to further evaluate the performance of U-selection in characterizing microbial communities, we selected a set of plant samples (both herbarium specimens and archaeological remains) with low levels of deamination. We extracted DNA from plant samples and generated libraries using both a regular double-stranded (ds) approach[53], and U-selection[208]. Sequences from the dsDNA libraries were used as a baseline to evaluate depletion and enrichment of uracil-containing molecules (Figure 3.7A). U-selection successfully enriched for deaminated molecules in all plant samples, as shown by much higher levels of deamination present in the U-enriched fraction compared with the dsDNA libraries and the U-depleted fraction (Figure 3.7A-B). The plant samples showed substantial variation in the content of endogenous DNA (2.8-91%), which was very similar between the U-depleted and U-enriched fractions, indicating similar levels of deamination between host- and microbe derived reads (Figure S2A). Assuming that plant- and microbial-derived DNA deaminate at a similar rate[148], this observation indicates that microbes found in plant tissue were present at the time of collection or colonized the tissue shortly thereafter. The percentage of reads (including host-derived reads) that could be taxonomically binned varied depending on the sample (Figure S2B) and,

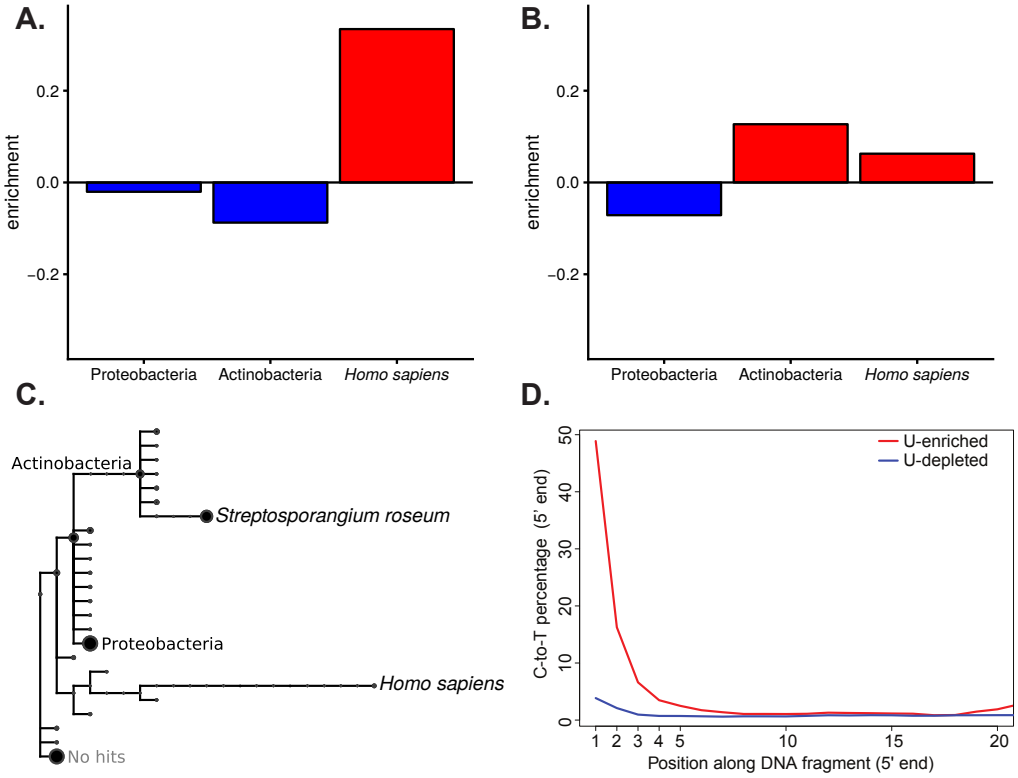

Figure 3.6: Analysis of Neandertal-derived U-selected libraries. A. Relative enrichment (number of reads) in the U-enriched relative to the U-depleted fraction from Vindija Neandertal assigned to the phyla Actinobacteria and Proteobacteria, as well as to *Homo sapiens*. **B.** Relative enrichment (number of reads) in the U-enriched relative to the U-depleted fraction from Sidrón Neandertal assigned to the phyla Actinobacteria and Proteobacteria, as well as to *Homo sapiens*. **C.** Taxonomic tree of reads from Sidrón Neandertal assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part of the taxonomy. Assignments to the phyla Actinobacteria and Proteobacteria, as well as the species *Streptosporangium roseum* and *Homo sapiens* are named in the taxonomic tree. **D.** Cytosine to Thymine substitutions at the 5'-end of reads aligned to *S. roseum* from the Sidrón Neandertal U-selected library (U-enriched and U-depleted fractions).

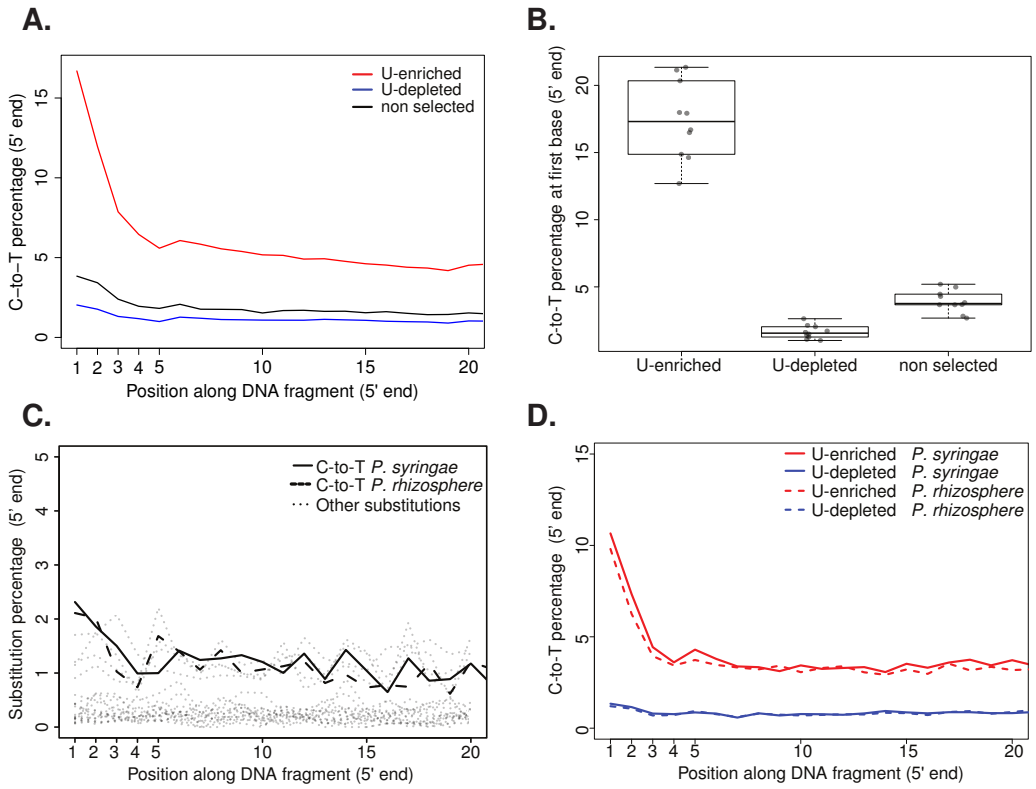

Figure 3.7: Patterns of C-to-T substitutions in plant- and *Pseudomonas*-derived reads. A. Cytosine to thymine (C-to-T) substitutions at the 5'-end of *Solanum tuberosum* sample KM177500 for a non-selected and U-selected library (U-enriched and U-depleted fractions). **B.** Distributions of C-to-T substitution percentage at first base (5'-end) for non-selected and U-selected libraries (U-enriched and U-depleted fractions). Median values are denoted as black lines and points show the original value for each individual sample. **C.** Substitution patterns at the 5'-end of *Pseudomonas syringae* and *Pseudomonas rhizosphaerae* mapped reads from a non-selected library from a *Solanum tuberosum* sample KM177500. **D.** C-to-T substitutions at the 5'-end of *P. syringae* and *P. rhizosphaerae* mapped reads from a U-selected library (U-enriched and U-depleted fractions) from a *Solanum tuberosum* sample KM177500.

since the host genome was included in the nucleotide database, positively correlated with the percentage of host endogenous DNA (Figure S2E). The inability to taxonomically assign the vast majority of reads from samples with low endogenous DNA reflects the incompleteness of the reference database compared to the diversity of the microbiomes in those samples. Additionally, single stranded DNA library preparation methods as employed during uracil enrichment generate shorter reads[52,55], which are more difficult to map to a reference genome or to assign to a reference database. This is reflected in the higher percentage of reads mapped and assigned from the dsDNA library compared with shorter reads derived from both the U-depleted and U-enriched fraction (Figure S1A-B). Originally, it was reported that the U-enriched fraction shows a mild increase in GC-content[208], however in the plant libraries analyzed here we did not find a significant difference in GC-content between the U-depleted and U-enriched fractions (Figure S1C). In theory, since Us originate from Cs, the U-enriched fraction would be enriched for GC-rich species and GC-rich genomic regions within a given genome. However, as the enrichment would depend on the diversity of taxa present and their relative age difference, and hence difference in deamination, GC-biases, if any, are expected to be highly sample-dependent.

Given the low taxonomic diversity of microorganisms in the samples included in our proof-of-principle experiment, instead of centering our analyses on the compositional assessment of microbial communities, we investigated in detail samples in which a specific microbe or group of microbes were more prevalent based on read abundance. We identified a large number of reads that were assigned to the bacterium *Pantoea vagans* in a potato (*Solanum tuberosum*) and a maize (*Zea mays*) sample (Figure 3.8A). In both samples we found patterns of C-to-T substitutions that suggest the historical nature of the sequenced reads (Figure 3.8B). Since *P. vagans* is a plant epiphyte[215], it is not entirely surprising to find it in two different plant species. We compared the potato and maize *P. vagans* with publicly available genomes using single nucleotide polymorphisms (SNPs) ascertained in these modern samples. Our analysis linked the two historical strains to a distinct cluster of modern strains based on genetic similarity (Figure 3.8C). Based on a set of 432,891 SNPs, the two historical isolates showed 95% SNP identity between them, and an average of 92% SNP identity between historical and modern strains of the same cluster. Conversely, comparisons between historical strains and any modern strain of a different cluster showed only an average of 59% identity at variable positions.

In a potato sample in which the pathogenic oomycete *Phytophthora infestans* was previously identified[72], we found a large portion of reads assigned to the bacterial genus *Pseudomonas*. Reads were assigned in particular to the species *Pseudomonas syringae* and *Pseudomonas rhizosphaerae* in different proportions (Figure 3.9). We performed de novo assembly using reads assigned to the genus *Pseudomonas* and aligned the contigs to the reference genomes of *P. syringae* and *P. rhizosphaerae* covering about 80% of both reference genomes (Figure 3.10). We subsequently filtered for contigs that aligned uniquely to either *P. syringae* and *P. rhizosphaerae* genomes and found different k-mer coverage distributions in contigs aligning uniquely to each genome (Figure 3.11), an ob-

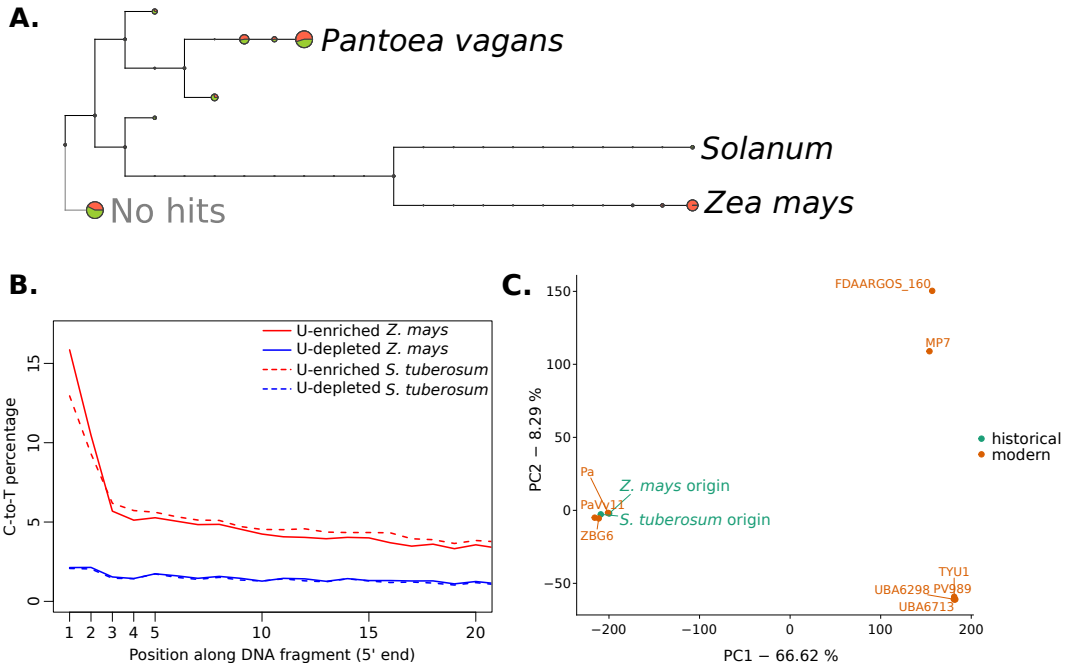

Figure 3.8: Analysis of *Pantoea vagans* sequences from different plant samples. A. Taxonomic tree of reads from *Solanum tuberosum* and *Zea mays* assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part or the taxonomy. *S. tuberosum*- and *Z. mays*-derived reads are shown in green and orange, respectively. **B.** Cytosine to thymine substitutions at the 5'-end of *P. vagans* for U-selected libraries (U-enriched and U-depleted fractions) from *Z. mays* and *S. tuberosum*. **C.** Principal component analysis of *P. vagans* from *Z. mays* and *S. tuberosum* samples, as well as nine publicly available genomes, based on single nucleotide polymorphisms. Numbers in axis labels indicate the percentage of the variance explained by each principal component (PC).

servation that reinforced our confidence in the presence of the two *Pseudomonas* species in this sample. Due to the high level of sequence divergence between the *Pseudomonas* in our sample and the reference genomes present in the database, it is difficult to assess typical deamination patterns in the dsDNA library (Figure 3.7C). However, we were able to examine damage patterns in both *Pseudomonas* species using the U-enriched fraction (Figure 3.7D), since the C-to-T signal is amplified and is much higher than the basal level of substitutions.

In summary, we showed that the U-selection method selectively enriches for authentic microbial aDNA molecules in samples from plant and animal tissues with a wide-distribution of ages and deamination levels. For instance, in *P. vagans*, U-selection increases the fraction of molecules carrying a terminal C-to-T substitution at the 5'-end 2-3 fold over the library without enrichment, relative to the total number of molecules sequenced. Sequencing both the uracil enriched fraction and the uracil depleted fraction allows to identify the presence of authentic ancient microbial taxa by assessing their relative abundance in these two fractions. Additionally, the magnifying effect of uracil enrichment on the detection of deamination patterns can help authenticating very divergent microbial taxa with moderate levels of degradation. All in all, we think that selective uracil enrichment is a valuable addition to ancient DNA metagenomics, both for de-novo taxon discovery as well as for cases where authenticity might be contentious.

Since it is extremely difficult to differentiate between ante-mortem and early post-mortem colonizers based only on deamination patterns, it is fundamental to also evaluate the biological relevance of detected taxa by comparing them with reference modern microbiomes.

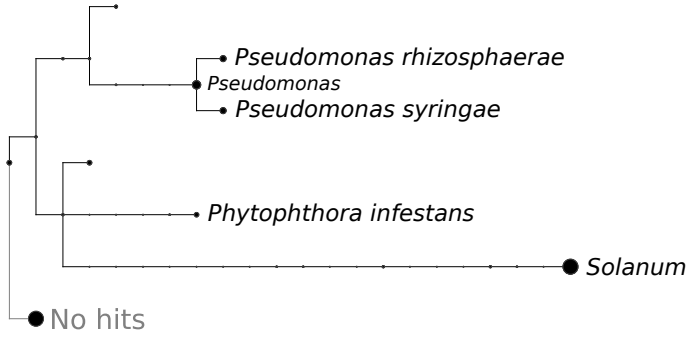

Figure 3.9: Analysis of *Pseudomonas* sequences from a *Solanum tuberosum* library. Shown is a taxonomic tree with reads from a *Solanum tuberosum* library assigned to different taxonomic levels. The size of the circle represents the amount of reads assigned to a particular part of the taxonomy. Select nodes of the species *Phytophthora infestans*, *Pseudomonas syringae* and *Pseudomonas rhizosphaerae*, as well as the genera *Pseudomonas* and *Solanum* are named in the taxonomic tree.

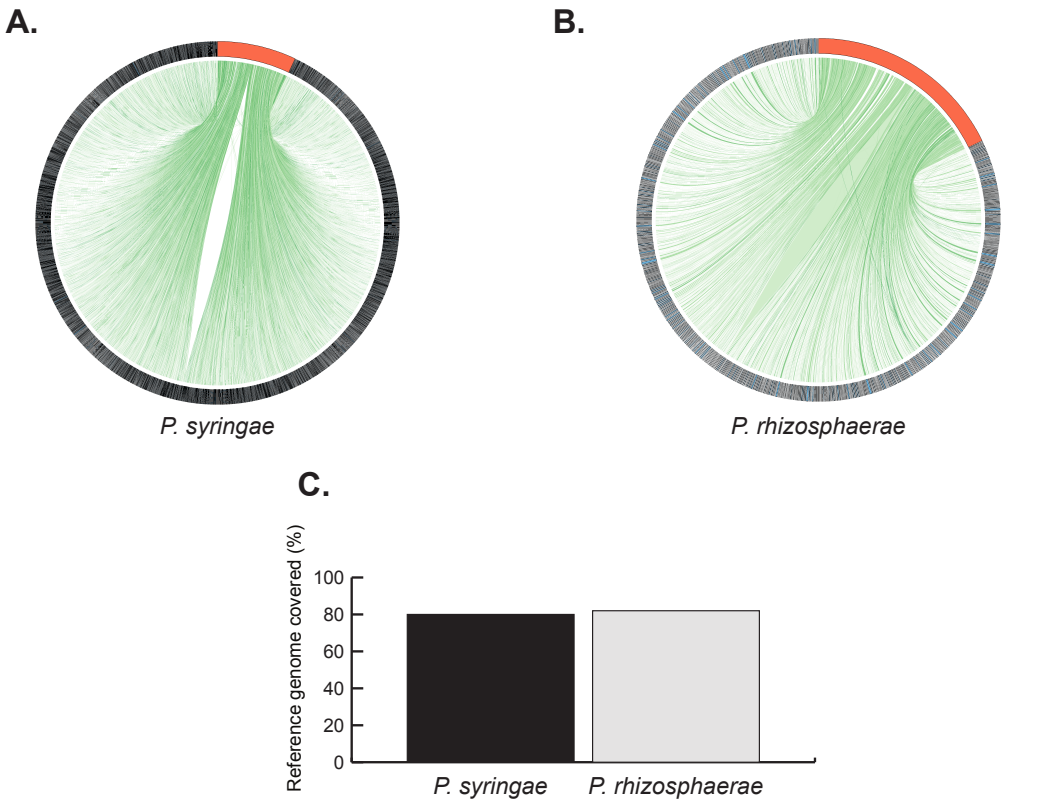

Figure 3.10: De novo assembly and genomic coverage of *Pseudomonas syringae* and *Pseudomonas rhizosphaerae* from a *Solanum tuberosum* sample. **A.** Alignments (represented as green lines) between all de novo assembled contigs (black/blue) and a *P. syringae* reference genome (orange). **B.** Alignments (represented as green lines) between all de novo assembled contigs (black/blue) and a *P. rhizosphaerae* reference genome (orange). **C.** Percentage of the *P. syringae* and *P. rhizosphaerae* reference genomes covered by de novo assembled contigs from A. and B., respectively.

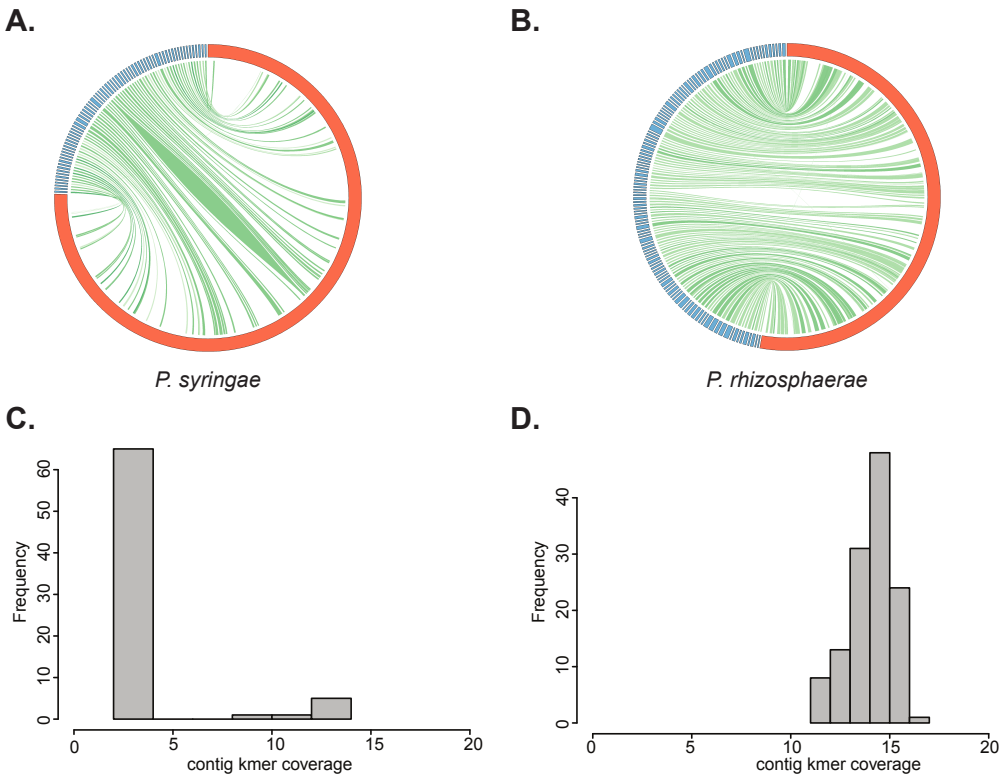

Figure 3.11: De novo assembly (uniquely mapped contigs) and contig k-mer coverage of *Pseudomonas syringae* and *Pseudomonas rhizosphaerae* from a *Solanum tuberosum* sample. **A.** Alignments (represented as green lines) between de novo assembled contigs (blue) uniquely mapped to a *P. syringae* reference genome (orange). **B.** Alignments (represented as green lines) between de novo assembled contigs (blue) uniquely mapped to a *P. rhizosphaerae* reference genome (orange). **C.** Histogram of contig k-mer coverage from de novo assembled contigs uniquely mapping to *P. syringae*. **D.** Histogram of contig k-mer coverage from de novo assembled contigs uniquely mapping to *P. rhizosphaerae*.

Table 3.2: Samples used in Section 3.4.

ID	Country of origin	Age	Species	Source
KM177500	UK	171 [5]	*Solanum tuberosum*	[6]
KM177497	UK	170 [5]	*Solanum tuberosum*	[6]
BM000815937	UK	279 [5]	*Solanum lycopersicum*	[7]
BH0000061459	USA	119 [5]	*Arabidopsis thaliana*	[8]
OSU13900	USA	82 [5]	*Arabidopsis thaliana*	[9]
NY1365364	USA	127 [5]	*Arabidopsis thaliana*	[10]
NY1365375	USA	119 [5]	*Arabidopsis thaliana*	[10]
CS5	USA	1852 [11]	*Zea mays*	[12]
CS6	USA	Undated	*Zea mays*	[12]
CS20	USA	1881 [11]	*Zea mays*	[12]
El Sidrón 1253	Spain	39,000 [11]	Neanderthal	[13]
Vindija 33.17	Croatia	Undated	Neanderthal	[14]
Vindija 33.19	Croatia	Undated	Neanderthal	[14]

[5]Calculated from collection dates (in years).
[6]Kew Royal Botanical Gardens.
[7]Natural History Museum, London.
[8]Cornell Bailey Hortorium.
[9]Ohio State University Herbarium.
[10]New York Botanical Garden.
[11]B.P. (Before present years).
[12]Turkey Pen Shelter, UTAH, USA.
[13]El Sidrón Cave, Spain.
[14]Vindija Cave, Croatia.

3.5 Discussion

The utility of DNA from ancient and historical specimens is being recognized in a growing number of fields, ranging from human[115], animal and plant genetics[68,108], to microbiology and epidemiology of infectious diseases[144]. Providing positive evidence for the authenticity of such ancient DNA from diverse sources is instrumental for all studies that make use of this resource. This can be challenging in complex mixtures of ancient DNA molecules, where sequences of interest may be at low abundance, or taxa are identified de novo.

Still, it is important to present positive evidence of the historical origin of DNA fragments, especially in cases where authenticity is contentious, and extraordinary claims are made based on a historical sample. This is particularly relevant in the context of the history of ancient DNA research, which is troubled by grandiose claims[16], many of which proved irreproducible[226] or provably false[17].

However, most of these claims date back to before library-based sequencing, when ancient DNA was amplified by PCR. The dangers associated with this practice[227] prompted the implementation of stringent safety precautions[14,228], many of which are practiced to this day.

The great advantage of library-based sequencing is the ability to reconstruct the full sequences of a large numbers of DNA molecules, without prior knowledge of what to look for[26,27]. This provides many advantages for authentication over PCR-based sequencing, especially when assessing age-associated degradation patterns[38,39]. As previously discussed, these include the distribution of molecule length, and characteristic substitution patterns caused by the deamination of cytosines. Still, library-based sequencing does not absolve from the burden of proof of authenticity, as contamination is just as critical an issue as it was during PCR-based sequencing[183].

In this chapter, we have presented two case studies which aim to address different problems in ancient DNA metagenomics and authentication. We showed, how the ubiquitous pattern of C-to-T substitutions can be captured from very few bona fide ancient DNA sequences, and how to statistically assess them. This method is especially useful when library complexity is low, and specialized library preparation protocols are not an option. In addition to presenting the statistical framework of this method and its application to a case study where ancient DNA authenticity was contentious, we also implemented the method in a stand-alone tool (`https://github.com/clwgg/ugat`). This implementation can be readily integrated into standard high-throughput sequencing workflows used in ancient DNA research, and assist in the authentication of aDNA sequences.

Additionally, we presented how novel sequencing library preparation approaches can aid the discovery of authentic ancient taxa. In contrast to targeted hybridization capture, the method we used enriches the DNA mixture in molecules carrying uracils as a result of the deamination of cytosines. This allows the enrichment of DNA of ancient or historical origin, and naturally permits the de novo discovery of bona fide ancient taxa from complex mixtures. Both of these approaches present valuable additions to the methods available for ancient DNA research, especially in cases where proof of authenticity is difficult to obtain.

4 Inference of ploidy

Contributions

Parts of the content of this chapter have also been published in the article "nQuire: a statistical framework for ploidy estimation using next generation sequencing"[229]. The following people have contributed to the work presented in this chapter: Hernán Burbano (HB), Sophien Kamoun and myself conceived and designed the project. HB and myself wrote the article with help from all authors. I developed the software application, prepared sequencing libraries, analyzed the data and generated all figures. Marina Pais and Liliana Cano performed DNA extractions.

4.1 Intraspecific ploidy variation

The number of complete sets of chromosomes an organism carries in its cells is referred to as its ploidy. Consequently, polyploidy is the presence of at least one additional set of chromosomes. Since many eukaryotic organisms are diploid, i.e. two sets of chromosomes which are paired through the fusion of two haploid gametes, polyploidy is sometimes referred to the presence of at least three complete sets of chromosomes.

Polyploidiziation events are thought to have happened in the past of many eukaryotic lineages, especially in plants[230]. However, many of these paleopolyploids revert back to functional diploidy over time, as chromosome copies diverge from each other[231]. Additionally, in sexually reproducing organisms, meiosis is required to separate sets of chromosomes into functional gametes, a process which appears to be most easily maintained if haploid gametes are created from diploid cells[232].

More recent polyploidization may also drive speciation events, for example if one of two separating lineages acquires additional sets of chromosomes, thereby strengthening the barriers to reproduction between the lineages (interspecific ploidy variation)[233].

Two types of polyploids need to be distinguished by the process of how additional sets of chromosomes are acquired[231,234]. Organisms where the sets of chromosomes originate from the same or a closely related individual are referred to as autopolyploids. In contrast, the organism is called allopolyploid if the chromosomes are acquired from a different species. In the latter case, the different sources of chromosomes may be easily distinguished by their genome sequence, and the organism is a hybrid of the parental species.

In many species, polyploid individuals which arise for example through errors during meiosis are either not viable or don't contribute to the next generation[234,235]. This is especially true for most animals and other sexually reproducing species[236]. However,

especially in species with the capacity to reproduce asexually, intraspecific ploidy variation can arise[235]. This event is arguably one of the most dramatic mutations possible, as effectively every gene in the organism's genome is duplicated. This is accompanied by increases in cellular and nuclear volume, disruption of gene expression, and often times by genomic rearrangement and aneuploidy[234,237]. Such dramatic effects on genome structure, together with mismatching numbers of chromosomes, substantiates the negative effect this can have on sexual reproduction. In fact, in organisms such as the oomycete *Phytophthora infestans*, it has been shown that ploidy level covaries with the frequency of sexual vs. asexual reproduction, which varies geographically[238]. In regions where sexual reproduction is prevalent, individuals tend to be diploid to facilitate functional sexual reproduction. Outside of the range of sexual reproduction however, individuals are predominantly polyploid.

During asexual reproduction of these types of organisms, their ploidy level and indeed their genome structure can be quite dynamic[239]. Since there is little pressure to maintain chromosomes which functionally segregate during meiosis, other types of mutations, structural rearrangements, genome duplications and aneuploidies may convey a selective advantage[240]. This can be especially true under stress[239], as shown for example in yeast strains domesticated for beer brewing[241]. Still, other types of stress may cause a polyploid to revert to diploidy, as, for example, when treating triploid *Phytophthora infestans* with sub-lethal doses of fungicide[242].

These observations show that the evolutionary consequences of polyploidization are highly complex. In the short term, polyploids may have an advantage, effectively by providing more "degrees of freedom" which evolutionary forces can act upon[234]. For example, when comparing haploid, diploid and tetraploid strains of yeast in the time it takes until one strain gains a selective advantage over another, previously genetically identical strain, the tetraploid is by far the fastest[243]. In the long term however higher polyploids may have to fight with genomic instabilities and aneuploidies, the less efficient purging of deleterious mutations, and infertility[244].

To study the evolutionary history and consequences of intraspecific ploidy variation, it is first necessary to be able to measure ploidy level. Traditionally, this is done by flow cytometry of intact nuclei, effectively assessing the amount of DNA present[245]. Apart from being cumbersome for large populations of samples, this is also problematic for ancient DNA applications, where nuclei are no longer intact.

In genome sequencing studies of organisms with intraspecific ploidy variation, it is therefore desirable to infer ploidy as a "byproduct" from the data generated to study sequence variation.

4.2 Inferring ploidy from short read data

To infer the level of ploidy from sequencing data, one generally is dependent on allelic variation at heterozygous sites. One way to query this allelic variation is by inspecting k-mer distributions[246]. In these approaches, all sub-sequences of a given length k which are present in the sequencing data are counted. Allelic variation at a site will produce

k-mer counts in accordance to the copy number of the alleles. In a histogram of all k-mer counts, these allele dosages will lead to different peaks in the histogram[246,247]. This allows to collect evidence for different ploidies even in the absence of a reference genome. However, substantial sequencing coverage is required to distinguish the peaks that different ploidy levels give rise to (Figure 4.1).

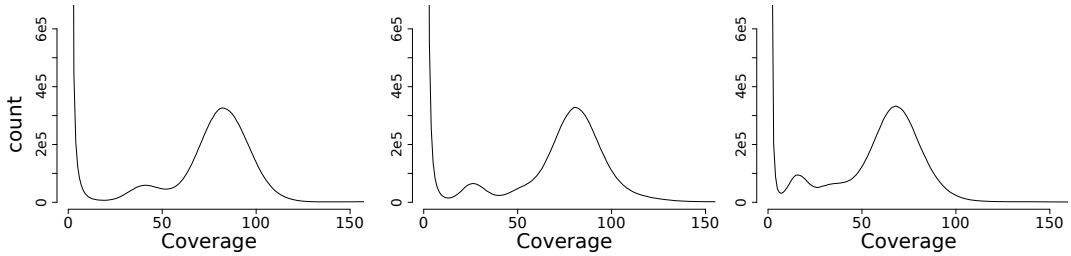

Figure 4.1: k-mer coverage count distributions for a diploid (left), triploid (center) and tetraploid (right) sample of *Saccharomyces cerevisiae*. k-mers were counted using KMC2 with $k = 21$. For sample information, see also Table 4.1.

If a reference genome is available, allelic variants can be identified by using it to align sequences in a common coordinate system[72]. Biallelic variants identified in that coordinate system are expected to segregate at different ratios for different ploidy levels, that is, 0.5/0.5 in diploids, 0.33/0.67 in triploids, and a mixture of 0.25/0.75 and 0.5/0.5 in tetraploids (Figure 4.2).

These ratios hold true only if there is sufficient heterozygosity in the source genomes which constitute the polyploid. For example, two completely homozygous diploid genomes coming together to form a tetraploid will be indistinguishable from a diploid genome by its ratios of allelic variants alone. Also, recent allopolyploids from highly diverged source genomes may present additional difficulties in the alignment to a common reference genome.

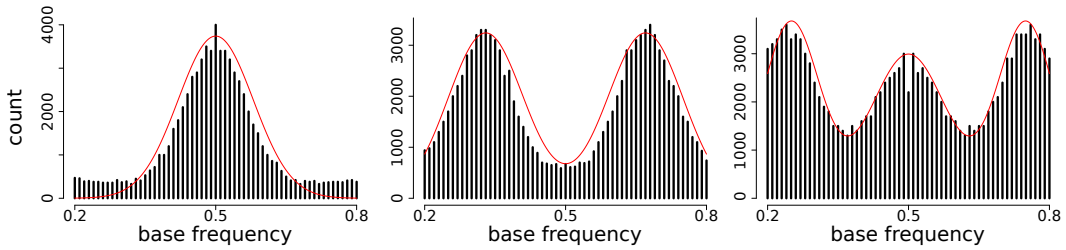

Figure 4.2: Distribution of base frequencies at variable sites where two bases are segregating for a diploid (left), triploid (center) and tetraploid (right) sample of *Saccharomyces cerevisiae*. Red lines show Gaussian distributions fit to the frequency histograms. For sample information, see also Table 4.1.

This methodology based on ratios of allelic variants has been used to detect intraspecific variation of ploidy in the oomycete *Phytophthora infestans*[72,242] and in the Baker's yeast *Saccharomyces cerevisiae*[248]. It also was successfully used for ploidy estimation in historical herbarium samples of *P. infestans* which are not suitable for flow cytometry[72].

To determine the ploidy level, the distribution of biallelic SNPs can be inspected visually[249], and qualitatively compared with simulated data[72]. However, neither of these approaches provide summary statistics that permit quantifying how well the data fit the expected distributions, which is especially critical when dealing with noisy distributions typical for highly-repetitive genomes. An additional disadvantage of this approach is that it is preceded by the identification of variable sites ("SNP calling"), which is carried out using methodologies that benefit from a previously known ploidy level[97].

4.3 nQuire

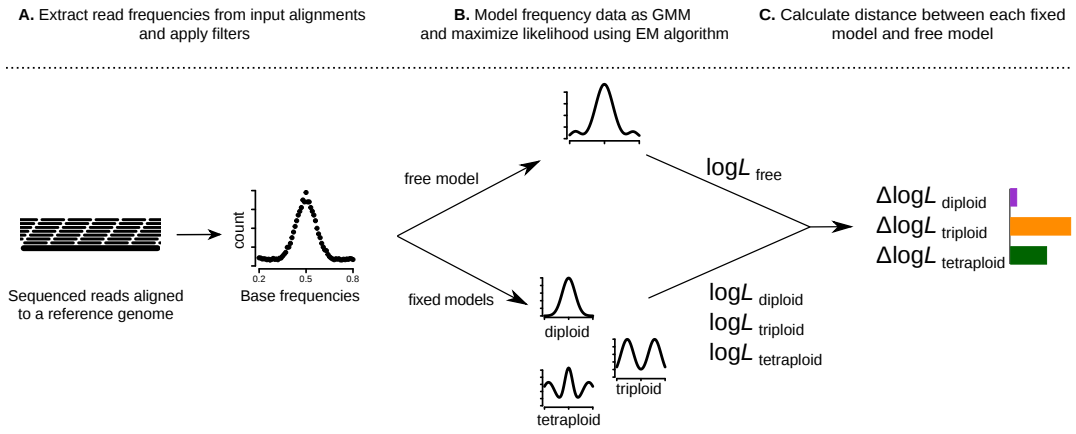

Figure 4.3: Overview of the Gaussian Mixture Model (GMM) based method. The workflow is illustrated using a diploid individual as an example. **A.** After sequenced reads are mapped to a reference genome, base frequencies are calculated at variable sites where only two bases are segregating. **B.** The base frequencies are modeled using a GMM and the likelihood is maximized using an Expectation-Maximization (EM) algorithm for both the free and the three fixed models (diploid, triploid and tetraploid). The maximized log-likelihoods ($\log L$) are extracted for subsequent model comparison. The curves show a possible final state of the GMM under the assumptions of each of the four models. **C.** The $\Delta\log L$ is calculated between the free model and each of the three fixed models (here represented as barplots). The fixed model with the smallest $\Delta\log L$ is chosen as the true ploidy level (diploid in this example).

In this chapter, we describe the implementation of `nQuire`, a command line tool which provides a statistical framework for ploidy estimation. The design goal for this tool was to distinguish diploids, triploids and tetraploids directly from sequence alignments to

a reference genome, without the requirement of high quality variant calls. Part of the rationale for this was that many variant calling procedures benefit from knowing the ploidy level a priori[97], which creates a circular problem. Additionally, the aim was to make the analysis of ploidy from short read data more streamlined and explorative, without having to first wait for and work on high quality, curated variant calls.

`nQuire` was designed to use standard file formats without the need for pre-processing or data extraction, so that it could easily be integrated into High Throughput Sequencing workflows. It is implemented in the `C` programming language, and available at `https://github.com/clwgg/nQuire`. `nQuire` uses the `htslib` library to interface with read alignment files, as well as for its multithreading implementation. The use of `htslib` in a `C` framework lends itself well to the goal of high usability, as well as computational efficiency.

4.3.1 The Gaussian Mixture Model

`nQuire` uses base frequencies at variable sites with two bases segregating to distinguish between diploids, triploids and tetraploids (Figure 4.3A). It models base frequency profiles as a mixture of three Gaussian distributions (Figure 4.3B), which are scaled relative to each other. A log-likelihood can be calculated following:

$$\log L = \sum_{i=1}^{n} \log \sum_{j=1}^{3} \alpha_j N(x_i; \mu_j, \sigma_j)$$

Here, n describes the numbers of data points and x_i describes the value of each data point (i.e. the base frequencies). At sites with two bases segregating, the frequencies of both bases are used to achieve symmetry in the frequency distribution for triploids and tetraploids. μ_j and σ_j are the parameters of the j^{th} of three Gaussian distributions N_j that are scaled relative to each other through the parameter α_j, with $\sum_{j=1}^{3} \alpha_j = 1$.

This model allows estimating the parameters of the Gaussian mixture components, as well as their mixture proportions, by maximizing the log-likelihood, either with or without constraints on the possible parameter space.

The choice to use Gaussian mixtures to approximate a binomial process is motivated by the fact that `nQuire` avoids the requirement of high quality SNP calls. The Gaussian approximation proved to be less affected by higher noise levels due to the use of base frequencies directly from sequence alignments. Especially critical are high coverage outliers, which arise for example from misalignments of paralogous sequences (Figure 4.4).

The likelihood maximization of the Gaussian Mixture Model (GMM) is implemented through an Expectation-Maximization (EM) algorithm (Section 4.3.2), which is specific to the GMM but can be extended to similar models (Figure 4.3B). The algorithm estimates all parameters at once and computes a likelihood ("free model"). Alternatively, a likelihood can be calculated with constant parameters ("fixed models") under the expectation of diploidy (one Gaussian with mean 0.5), triploidy (two Gaussians with means

0.33 and 0.67) and tetraploidy (three Gaussians with means 0.25, 0.5 and 0.75). Since all fixed models are nested within the free model, it is possible to directly compute the log-likelihood ratios, following:

$$\Delta\log L_{\text{diploid}} = \log L_{\text{free}} - \log L_{\text{diploid}}$$
$$\Delta\log L_{\text{triploid}} = \log L_{\text{free}} - \log L_{\text{triploid}}$$
$$\Delta\log L_{\text{tetraploid}} = \log L_{\text{free}} - \log L_{\text{tetraploid}}$$

The $\Delta\log L$s describe the distance between each fixed model and the best fit under the assumptions of the GMM. A substantially lower $\Delta\log L$ of one fixed model over the others supports the ploidy level described by this fixed model (Figure 4.3C). Therefore, the $\Delta\log L$ are used as summary statistics where the minimum value supports a given ploidy level.

Additionally, the GMM can be extended to a Gaussian Mixture Model with Uniform noise component (GMMU), by adding a uniform mixture component:

$$\log L = \sum_{i=1}^{n} \log \left(\alpha_4 U(x_i) + \sum_{j=1}^{3} \alpha_j N(x_i; \mu_j, \sigma_j) \right)$$

The constraint on the mixture proportions then becomes
$\sum_{j=1}^{4} \alpha_j = 1$.

The uniform noise component is used to allow base-line noise removal (Figure 4.5). This can be useful when the Gaussian peaks are observable but embedded in a basal noise, which could be caused by highly repetitive genomes or low coverage.

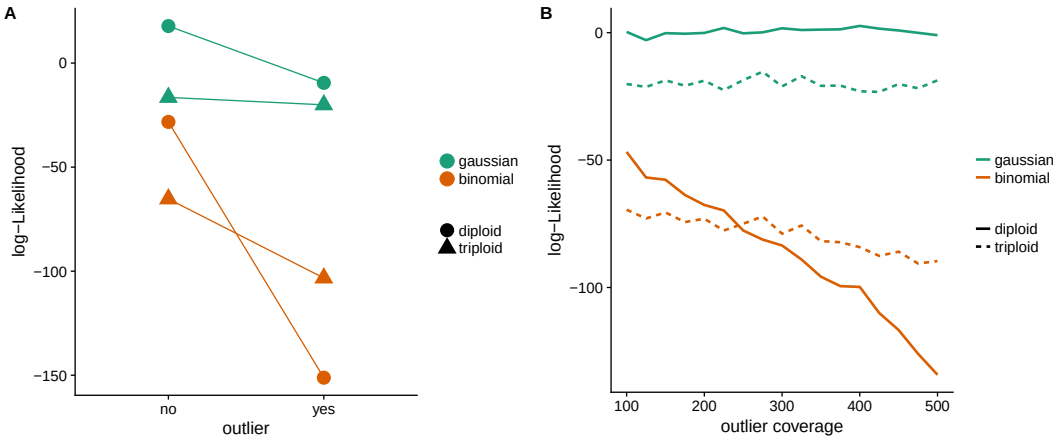

Figure 4.4: Effect of high coverage outliers on the likelihood of Gaussian and binomial mixtures. The figure shows the results from a simulation of a diploid individual with 10 biallelic sites, sequenced to 100x coverage. **A.** Simulations were performed without and with outliers. For the former, we simulated 100x coverage at all 10 positions, and drew a base count from a binomial of size 100 and probability 0.5. For the latter, one out of the ten positions was simulated at 400x coverage and the base count was drawn from the same binomial as above. This simulates a scenario, where reads from three additional (potentially) paralogous, homozygous alleles are falsely aligned to a heterozygous diploid site. In the simulations without an outlier (left column), the diploid fixed model shows a higher log-likelihood than the triploid fixed model both for Gaussian and binomial mixtures. In contrast, in the presence of the high coverage outlier (right column), the drop in log-likelihood for the binomial mixture is more drastic than for the Gaussian mixture, as the high coverage site is given more weight in the binomial. **B.** The figure shows how the coverage of the outlier influences the log-likelihood of the binomial mixture. Simulations were performed as in (**A.**) with presence of an outlier, with the difference that the outlier always has a minor allele frequency of 0.2, at different coverages ranging from 100x to 500x. The log-likelihood of the binomial mixture for the diploid fixed model drops below the log-likelihood of the triploid fixed model between 200x and 300x coverage of the outlier.

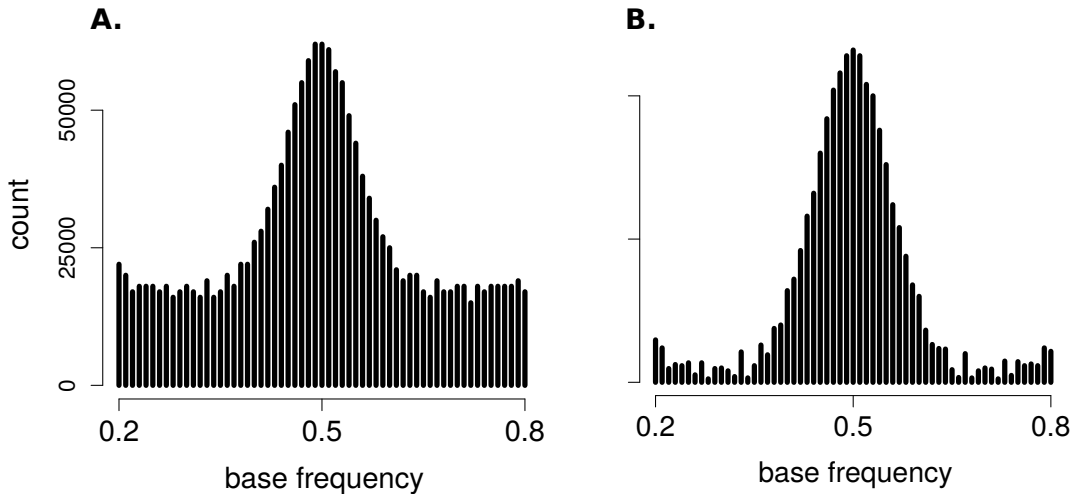

Figure 4.5: Effect of denoising using the uniform mixture component. A. The base frequencies of the diploid *P. infestans* sample 99189 before denoising. **B.** The base frequencies of the diploid *P. infestans* sample 99189 after denoising using the mixture proportion of the uniform mixture component after likelihood maximization. For sample information, see also Table 4.1.

4.3.2 Implementation

`nQuire` uses a subcommand system for the different functions supplied to the user. These include subcommands for data extraction, visualization, and analysis. The first subcommand a user will generally employ is the `create` command. This command reads a BAM file containing sequences aligned to a reference genome, and extracts base counts at sites where two bases are segregating, after a set of filters are applied. These filters include minimum base frequency, minimum mapping quality, minimum coverage, and maximum coverage. Sensible defaults are configured, which have shown to be applicable over a wide range of applications, but users may tailor them to their specific needs. Additionally, the `create` subcommand allows to filter the BAM file by specific regions of interest, which are defined in BED format. Extracted base counts are are written to a `.bin` file, which contains base counts and other meta information in binary format. All information written and read from these files is stored as integers of specified length, to avoid storing floating point information. To make the files transferable between UNIX systems of different endianness, `nQuire create` uses the standard library `<arpa/inet.h>` to store data in "network byte order" using the `hton()` functions ("host-to-network"). When data is read from the `.bin` file, the `ntoh()` functions are used ("network-to-host") to convert the data back to "host byte order".

The primary purpose of creating the intermediate `.bin` file is to allow quicker interaction with the data and easy iteration of different analyses, once the full alignment dataset is filtered for informative sites, since reading the `.bin` file is much quicker than reading the full BAM file.

By default, the `.bin` file does not carry information about the reference sequence and position from which the base count data originates. An extension to the `.bin` file which can be created using an optional flag does store this information. To avoid storing the reference sequence name as a string, it uses its ID, as it is specified in the BAM header format specified by `htslib`. To query the sequence name, the extended `.bin` file can be annotated using the original BAM file. To correctly identify the BAM file from which a `.bin` file is created, and trigger a warning if the wrong BAM file is supplied, the extended `.bin` file also stores the djb2-hash of the BAM header.

Once the `.bin` file is created, it can be used as the input to all other subcommands. First, the content of a `.bin` file can be visualized using the subcommands `view` and `histo`. The `view` subcommand simply prints the base counts in plain text, optionally annotated with the additional data stored in the extended `.bin` file. The `histo` subcommand prints an ASCII histogram to the command line, which shows the histogram of base frequencies similar to the representation in Figure 4.2.

The main subcommand for ploidy inference from a `.bin` file is called `lrdmodel`. It employs the Gaussian Mixture Model (GMM) and uses an Expectation-Maximization (EM) algorithm for maximum likelihood inference.

The EM algorithm makes use of the latent variables Z_i. They represent the assignment of a data point to one of the mixture components. In the E-step, the current estimates for μ_j, σ_j and α_j are used to calculate:

$$P(Z_i = j|x_i) = \frac{\alpha_j N(x_i; \mu_j, \sigma_j)}{\sum_{j=1}^{3} \alpha_j N(x_i; \mu_j, \sigma_j)} = \gamma_{Z_i}(j)$$

This probability is calculated for all x_i and $j \in \{1, 2, 3\}$ to form a $n \times 3$ matrix, the columns of which represent the probability of each data point belonging to either of the three mixture components. From this matrix, column sums are calculated as $S_j = \sum_{i=1}^{n} \gamma_{Z_i}(j)$, which represent the size of each mixture component. In the following M-step, the estimates of μ_j, σ_j and α_j are updated:

$$\hat{\mu_j} = \frac{1}{S_j} \sum_{i=1}^{n} \gamma_{Z_i}(j) x_i$$

$$\hat{\sigma_j^2} = \frac{1}{S_j} \sum_{i=1}^{n} \gamma_{Z_i}(j)(x_i - \mu_j)^2$$

$$\hat{\alpha_j} = \frac{S_j}{n}$$

The log-likelihood is calculated after the M-step, and the next E-step is initiated unless the log-likelihood has changed by less then $\epsilon = 0.01$ from the previous M-step.

As shown above, the algorithm allows the estimation of μ_j, σ_j and α_j simultaneously. This is done under the previously introduced "free model". The log-likelihood of the free model upon convergence represents the optimal fit under the assumptions of the model.

Select parameters can also be set to fixed values, in which case they are not updated in the M-step. This is used for the three "fixed models", where the log-likelihood is maximized under the expectation of diploidy (one Gaussian with mean 0.5), triploidy (two Gaussians with means 0.33 and 0.67) and tetraploidy (three Gaussians with means 0.25, 0.5 and 0.75):

$$\log L_{\text{diploid}} = \sum_{i=1}^{n} \log N(x_i; 0.5, \sigma)$$

$$\log L_{\text{triploid}} = \sum_{i=1}^{n} \log \sum_{j=1}^{2} 0.5 \cdot N(x_i; \mu_j, \sigma_j), \quad \mu_j \in \{0.33, 0.67\}$$

$$\log L_{\text{tetraploid}} = \sum_{i=1}^{n} \log \sum_{j=1}^{3} 0.33 \cdot N(x_i; \mu_j, \sigma_j), \quad \mu_j \in \{0.25, 0.5, 0.75\}$$

In the three fixed models, only σ_j is estimated, while μ_j and α_j are fixed as shown above. The `lrdmodel` subcommand then reports the maximized $\log L$ under the free model and the three fixed models, as well as the three $\Delta\log L$ (see also Section 4.3.1). The `lrdmodel` subcommand allows for multiple `.bin` files to be specified at once, if many samples are to be analyzed. It uses the `pthread` wrapper `thread_pool.h` supplied by `htslib` to run the maximum likelihood estimation for different samples in parallel.

Another subcommand using the EM algorithm implemented in nQuire is `denoise`. Here, the mixture of three Gaussians is extended with a fourth, uniform component (see also Section 4.3.1). Since this component has uniform probability density, it effectively just adds a mixture component α_4. The `denoise` subcommand uses the EM to estimate all parameters, similarly to the free model, with the inclusion of the uniform α_4. The estimate of α_4 is then used to establish a threshold at which to downsample each bin of the histogram equally (see also Figure 4.5).

4.4 Multivariate normal clustering

The `lrdmodel` subcommand of `nQuire` returns both the maximized $\log L$ under the free model and the three fixed models, as well as the three $\Delta\log L$. As described in Section 4.3.1, The $\Delta\log L$ can be interpreted as a goodness-of-fit of each ploidy level relative to the best fit under the free model. However, many modern studies of intraspecific ploidy variation will use population samples for which it may be cumbersome to assess the ploidy level of each sample individually.

To analyze such a population sample together, one can make use of the maximized $\log L$ of the three fixed models. The rationale here is, that even though factors like genome assembly quality and repetitive content may affect the comparability of these likelihoods between organisms, within a species the relative magnitude of the likelihoods given a certain ploidy level can be expected to be comparable. Relative magnitudes of the maximized $\log L$ can be obtained by normalizing the $\log L$ of the fixed models by the

maximized log$L$ of the free model. This is especially useful since the magnitude of the likelihoods are affected by single-sample effects such as sequencing coverage, and, since the fixed models are nested within the free model, this normalization scheme will scale the values that are being compared between 0 and 1.

Once the maximized log$L$ of each sample is normalized in this way, the population of samples needs to be classified into the three ploidy levels the fixed models correspond to. One way to achieve this, is to cluster the samples into three groups within the three dimensions spanned by the normalized, maximized log$L$.

The Gaussian Mixture Model (GMM) at the core of nQuire uses univariate Gaussians to model base frequencies. Each of the mixture components in the GMM are characterized by their mixture proportion α, and the mean μ and variance σ^2 of the univariate Gaussian itself. A natural extension to this model are multivariate normal (MVN) mixtures. Each mixture component in the MVN is now a multivariate normal distribution over d dimensions. They are characterized again by a mixture proportion α, but the mean μ is now a d - dimensional vector with one mean along each dimension d. Instead of a simple variance term, each multivariate normal is described by a $d \times d$ variance-covariance matrix Σ.

To use such a structure for the classification of samples into diploids, triploids and tetraploids using normalized, maximized log$L$, a mixture of three, three-dimensional normal distributions can be used. `mclust5` is a package for the `R` programming language which facilitates clustering using multivariate normal mixtures[250]. `mclust5` additionally allows to specify models in which the volume, shape and orientation of each mixture component is free to vary, or is restricted to be the same across all components. This is based on an Eigendecomposition of the covariance matrix Σ, the factors of which correspond to these three parameters[251]. For clustering the normalized, maximized log$L$, a configuration allowing equal volume, but varying shape and orientation showed the best overlap between cluster assignments of samples and their manually curated ploidy levels (Figure 4.8).

In addition to using `mclust5`, we have implemented a utility to use multivariate normal mixtures for clustering in higher dimensions. It is implemented in the `C` programming language, and available at `https://github.com/clwgg/MVNclust`. Primarily, this implementation is meant as a prove of concept for the extension of the EM algorithm from the univariate- to the multivariate case. It uses the GNU Scientific Library (GSL) for its Basic Linear Algebra Subprograms (BLAS), and MVN density estimation. Additionally, it uses the Eigendecomposition of the covariance matrix as implemented in `mclust5`[250,251], but only implements the case MVNs of equal volume and varying shape and orientation.

4.5 Additional methods

4.5.1 *Phytophthora infestans* libraries

The two benchmarking libraries from *P. infestans* were generated according to the protocol by Meyer and Kircher [53] from DNA extracted from lab cultures[252]. These libraries were sequenced to high coverage on an Illumina HiSeq 3000 machine in paired end 150bp mode. This sequencing data is available at the European Nucleotide Archive (ENA) under study number PRJEB20998.

4.5.2 Read mappings

Paired end reads generated for this study, as well as those from Zhu, Sherlock, and Petrov [248], were mapped to their respective reference genomes (*P. infestans*[176], *S. cerevisiae* panel from Zhu, Sherlock, and Petrov [248]) using BWA-MEM (version 0.7.10) with default settings[177].

4.5.3 k-mer histograms

21-mers were counted directly from raw sequencing data using KMC2 (version 2.1.1[253]). The default minimum and maximum counts of 2 and 255 were changed to 1 and 1×10^7, respectively, From the count table produced by KMC2, a histogram was produced using the `ReadNextKmer()` interface of the KMC2 `cpp` API.

4.6 Analysis and results

Table 4.1: Samples used to evaluate and benchmark `nQuire`. Benchmark samples from *Saccharomyces cerevisiae* and *Phytophthora infestans* were used. The smallest $\Delta\log L$ for each sample is highlighted in bold.

Sample	Ploidy	Species	$Cov.^+$	$\Delta\log L_{2n}$	$\Delta\log L_{3n}$	$\Delta\log L_{4n}$
CBS7837	2n	*S. cerev.*	116	**6319**	35721	23033
CBS2919	3n	*S. cerev.*	111	33614	**920**	29347
CBS9564	4n	*S. cerev.*	101	37003	22971	**6429**
99189	2n	*P. infes.*	210	**119218**	369682	293194
88069	3n	*P. infes.*	368	599933	**42717**	390002

nQuire directly processes BAM files and is designed to be efficient in memory usage and runtime. To process a 1GB *S. cerevisiae* BAM file (100x coverage), nQuire needs 70 seconds to build appropriate data structures, 6 seconds to run the models and calculate the maximum likelihood estimates, and uses a maximum of 8 Mb of RAM, whereas for processing a 10GB *P. infestans* BAM file (100x coverage) it needs 760 seconds, 100 seconds and 60 Mb of RAM, respectively. These benchmarks were performed on a single

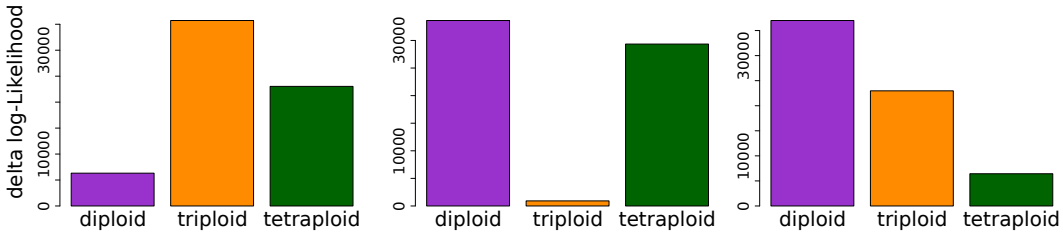

Figure 4.6: The $\Delta\log L$ of all fixed models for a diploid (left), triploid (center) and tetraploid (right) sample of *Saccharomyces cerevisiae* (also presented in Table 4.1).

core of a Intel® Core™ i5-4670 CPU on a system with 16Gb of DDR3-1600 RAM and an SSD.

We evaluate nQuire's performance using three *S. cerevisiae* samples at ~100x coverage, which represent each of the three ploidy levels evaluated by the model, as well as two *P. infestans* samples, one diploid and one triploid, at 210x and 368x coverage, respectively. The $\Delta\log L$ of each of the fixed models to the free model at full coverage is shown in Table 4.1. At those coverages, the $\Delta\log L$ of the best model is more than two times closer to the free model than the second best. Additionally, it coincides in all samples with the ploidy level inferred by visually inspecting the empirical distributions of base frequencies at full coverage (Figure 4.6 and Figure 4.10, center row).

To investigate the impact of coverage on the performance of the GMM, we downsampled mapped reads from the *S. cerevisiae* (Figure 4.7) and *P. infestans* (Figure 4.10, bottom row) strains shown in Table 4.1 to different coverage levels. In all cases the $\Delta\log L$s of the two improper models increases with increasing coverage. In contrast, the $\Delta\log L$ of the best model stabilizes at low coverage and doesn't increase further as coverage increases. The coverage at which the $\Delta\log L$ of the best model is stable will be different for genomes of different size and complexity, as shown in the difference between the two organisms used for benchmarking.

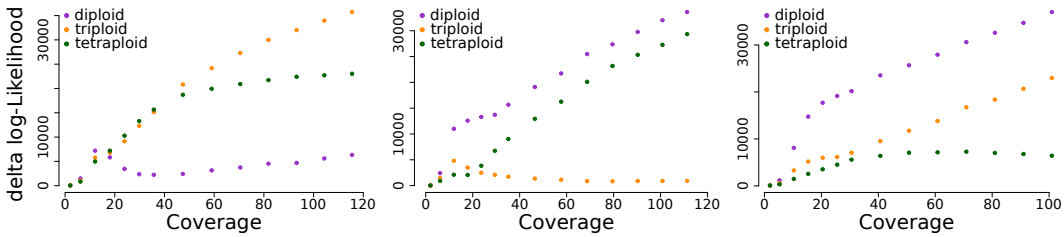

Figure 4.7: Coverage dependence of the $\Delta\log L$. Shown is the $\Delta\log L$ of all fixed models as a function of genome coverage for a diploid (left), triploid (center) and tetraploid (right) sample of *Saccharomyces cerevisiae*.

In cases where multiple samples are sequenced simultaneously, it might be impractical to assess ploidy in each sample individually. In these cases, we propose to use maxi-

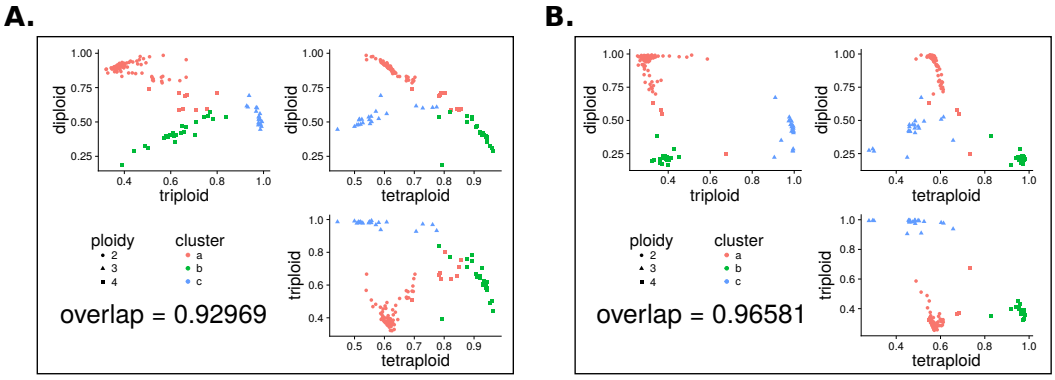

Figure 4.8: Clustering sets of samples into ploidy groups. Samples were clustered based on the normalized maximized log-likelihood of each fixed model representing each of the three ploidy levels. The clustering in three dimensions is shown with a set of three plots, one for each combination of ploidies. Shapes represent the manually assessed true ploidy levels, while colors show the assignments returned by the clustering algorithm. The agreement between the two is shown as a fraction. This analysis was conducted for samples before (**A.**) and after denoising (**B.**).

mized log-likelihoods of the three fixed models, normalized by that of the free model, to cluster samples into ploidy groups. The rationale is that within one species, the relative likelihoods of the fixed models will be similar within each ploidy level. As a proof of concept, we applied this to all di-, tri- and tetraploids from the *S. cerevisiae* test set[248], and clustered the samples into three groups in three dimensions using multivariate normal clustering (see Section 4.4). The sample set was manually scored for ploidy, and the overlap between clusters and manually assessed ploidy level was calculated (Figure 4.8). Running nQuire on raw data showed high recovery of ploidy level (93%, Figure 4.8A), which was further improved through the denoising implementation utilizing the GMMU (96%, Figure 4.8B).

Recent polyploidization is often associated with aneuploidies. To be able to detect those, nQuire allows to split the analysis of a sample by regions defined in BED format. We used this to reanalyze the sample YJM466 from the *S. cerevisiae* test set[248]. This sample had been shown to be triploid on whole genome level, but tetraploid for chromosome 6 and diploid for chromosome 9. The $\Delta\log L$s for the three fixed models individually calculated for each of the 16 chromosomes of *S. cerevisiae* confirmed this observation (Figure 4.9).

A natural extension of analyzing the genome by chromosomes is to use sliding windows to detect possible transitions between ploidies in aneuploid individuals. We used the three *S. cerevisiae* datasets shown in Table 4.1 at their full coverage to benchmark the number of randomly sampled positions needed to accurately assign ploidy (Figure 4.11). For these test samples, 100-200 random sites at 100x coverage are enough to correctly assign ploidy based on the $\Delta\log L$. However, this will vary for regions of the genome

with different complexity and repetitiveness.

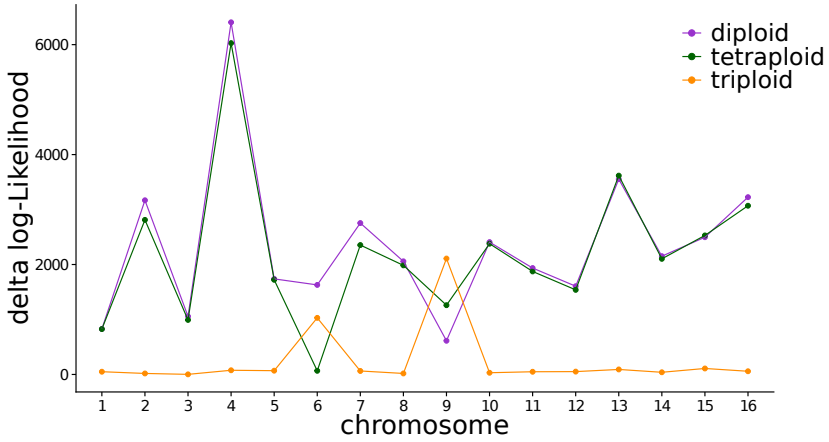

Figure 4.9: Detection of chromosome-wide aneuploidies. The ploidy estimation can be run for each chromosome independently, which enables detection of aneuploidies. The sample displayed here shows genome wide triploidy, but splitting the analysis by chromosome shows tetraploidy for chromosome 6 and diploidy for chromosome 9, as detected by $\Delta \log L$.

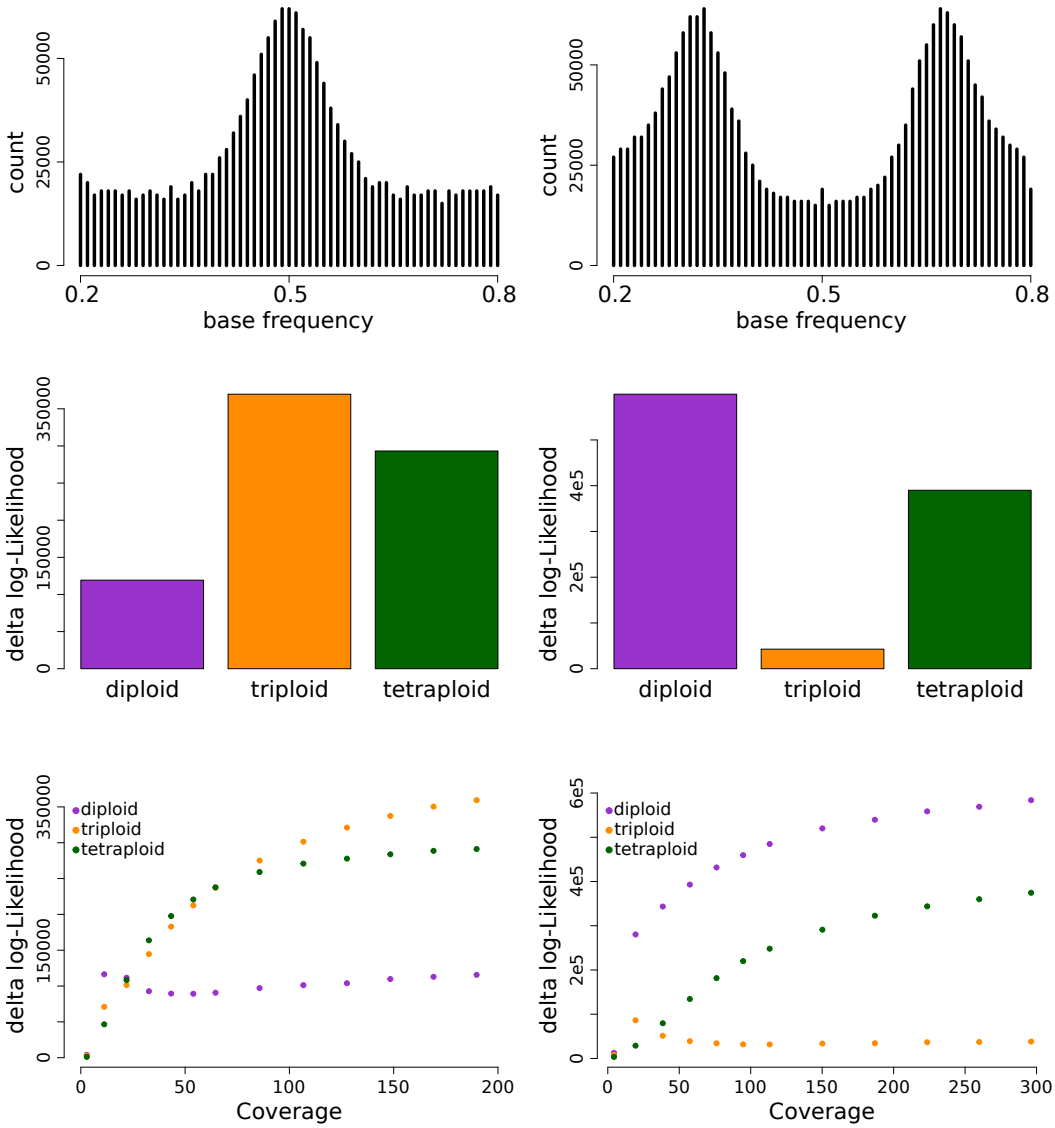

Figure 4.10: Evaluation and benchmarking of `nQuire` using a diploid (left) and triploid (right) sample of *Phytophthora infestans*. Distribution of base frequencies at variable sites where two bases are segregating are shown on top. The barplots depict the $\Delta \log L$ of all fixed models (also presented in Table 4.1). The bottom row depicts change of $\Delta \log L$ of all fixed models as a function of genome coverage.

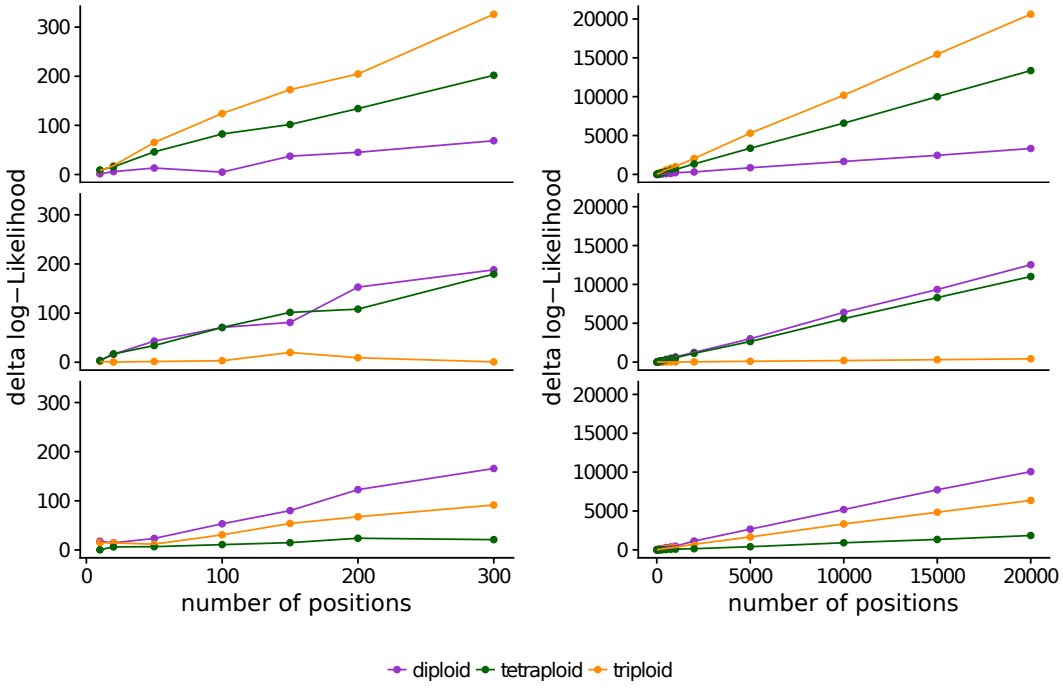

Figure 4.11: Number of positions required for reliable ploidy assignment. The diploid (top), triploid (middle) and tetraploid (bottom) *S.cerevisiae* datasets from Table 4.1 were used to randomly sample 10 to 300 positions (left). The subsampling was extended to 20,000 positions (right).

4.7 Discussion

In addition to nucleotide and structural variation, some organisms also vary intraspecifically in their ploidy level, which constitutes another source of variation that evolutionary forces can act upon[234]. `nQuire` permits assessment of ploidy variation from sequencing data usually generated for variant detection. In contrast to previous methods that visually analyze the distributions of SNPs at biallelic heterozygous sites[72,242,249], `nQuire` quantitatively distinguishes between different ploidy levels based on the distribution of base frequencies at variable sites, using relative differences in likelihoods. In comparison to the approach proposed by Gompert and Mock[254], `nQuire` avoids the requirement of high quality SNP calls. The higher level of noise in the data resulting from this is accounted for by using Gaussian distributions. They approximate a binomial process, but are impacted less by the effects of high coverage outliers (Figure 4.4). Additionally, nQuire is a Linux command line tool that uses standard file formats as input and handles large genomes at high coverage efficiently.

In all test cases, triploids were the easiest to distinguish, most likely caused by the lack of probability density around 0.5 compared to the other two models. While diploids and tetraploids are more difficult to tease apart, our results on coverage dependence show that at sufficient coverage, the data fits the true model much better than either of the two alternatives (Figure 4.7).

For cases where the Gaussian peaks were largely overlapped by uniform noise, the free model can be extended to include a uniform component, whose mixture proportion can be used - after likelihood maximization - for base-line removal.

We show that this procedure improves the recovery of the true ploidy level when samples are clustered based on maximized likelihoods under the assumptions of the fixed models (Figure 4.8B).

We also show that few high quality positions are enough to estimate the correct ploidy level (Figure 4.11). Such high quality positions can be selected by using stringent filtering criteria. Several filters are implemented in nQuire directly, thus no pre-processing of the BAM file (after duplicate removal) is necessary. They include minimum and maximum coverage, as well as mapping quality and base frequency filters. The default values of these filters have been configured to fit most applications. The exact coverage and number of positions needed for a reliable estimation of ploidy will however depend on the complexity and repetitiveness of the genome. Additionally, it is possible to obtain high quality positions by using BED files to define regions of low repetitiveness, where base frequencies can be more confidently assessed.

In sequencing projects which incorporate both ancient and modern DNA, `nQuire` facilitates the ploidy estimation of the entire sample set. This is not possible with flow cytometry, since it requires intact nuclei. Still, if modern samples are available for flow cytometry, they can be easily incorporated into the clustering procedure and used to validate the clusters identified by `nQuire`.

5 Inference of genotypes

Contributions

The content of this chapter is unpublished at the time of submission, and the following people have contributed:

Hernán Burbano and myself conceived and designed the project. I developed the software applications and generated the figure.

5.1 Sampling genotypes from low coverage data

A crucial step in almost all sequencing and re-sequencing studies is the assessment of sequence variation within and between populations of samples[81,255]. Most commonly, sites which show variation are identified in a reference set of samples for which high quality data is available. Samples of lower quality can then be genotyped based on this reference set[256]. The confidence in these genotypes mostly depends on the sequencing coverage, i.e. how many sequenced fragments cover a given variant site. The coverage also dictates whether heterozygous sites can be assessed[86]. The biggest issue here is that DNA sequencing is not an error-free process. Assuming sequencing error to be evenly distributed along the genome, enough non-erroneous bases have to cover a variant site to confidently distinguish true genotypes from sequencing error.

There are two main sources of bias associated with this method. First, the discordance between the reference set of samples in which sites of variation are ascertained, and the set of samples that are being genotyped, leads to ascertainment bias[257]. Strong ascertainment bias occurs if the reference set does not represent the genetic diversity that is being genotyped. An example of this would be if variable sites are detected in one population of a species, and these sites are used to genotype a different, genetically distinct population. In this case, both populations may have many private variants which drive their differentiation, which will be missed by this genotyping scheme. Still, another more subtle type of ascertainment bias occurs even if large scale patterns of diversity are covered in the reference set. This is because rare alleles will, by definition, only occur in a few individuals in a population[77]. While some truly rare alleles may be represented in the reference set, many will only be discovered once more individuals are added. In a genotyping scheme however, all rare alleles which are not present in the reference set will be missed. This is especially problematic if the reference set is much smaller than the set of samples being genotyped, since a larger set of samples will contain more rare alleles.

The second source of bias is not due to the source of the reference set of samples used to assess variant sites, but with the reference genome itself[258,259,260]. Almost all re-sequencing studies utilize a single reference genome, which is used to map short read data against. The reference genome thereby acts as a coordinate system for sequences of other samples, so that sequences which cover specific sites can be compared across samples. However, often the reference genome acts not only as a coordinate system, but also as a reference sequence. A common differentiation in re-sequencing studies is whether a sample carries the "reference allele" or the "alternative allele". This refers to a site of known variation, where both the allele present in the reference genome, as well as another, "alternative" allele has been observed. Reference bias occurs, if the "reference allele" is assumed if there is insufficient evidence for the "alternative allele" in a sample being genotyped. A common source of insufficient evidence can be for example low coverage, where it is difficult to distinguish genotype from sequencing error[261]. If the "reference allele" is assumed in these cases, the usually arbitrarily chosen genotype of the reference genome will be over-represented in samples of low coverage and bias their genotypes.

Reference bias is however not the only problem which comes with low quality, low coverage genotype data. Most modern pipelines for variant discovery and genotyping, like GATK[92], freebayes[95] or DeepVariant[262], are optimized for high quality, high coverage data. Using different models and heuristics, they aim do identify variant sites in (sets of) samples and call homozygous and heterozygous variants[217]. However, low coverage sequencing does not provide the type of data that enable these tools to confidently assess variation[97]. In the worst case this may lead to overcalling of reference alleles. A less problematic, but potentially overly conservative way of dealing with low confidence is to report non-informative sites in the form of missing data. Still, the models and heuristics that enable these tools to perform well with high quality data, may introduce unknown biases in the types of variants identified from low coverage data. Another way of dealing with low coverage, low quality sequencing data is to forgo "classical" variant calling all together, and randomly sample pseudo-haploid genotypes from the little sequencing data available[46,263]. This has the advantage of being computationally much more efficient than model-based variant calling. Additionally, if error were completely random and equally distributed across samples, this approach still allows an unbiased way to estimate genetic relatedness from extremely low coverage data[46], albeit with high variance.

In this procedure, a reference genome again acts as a coordinate system which allows the comparison of bases overlapping specific sites across samples. If genetic variation data from a reference set of samples is available, the genotyping of additional samples is simple. At each position where variation is observed in the reference set, a base is sampled randomly from all bases covering the site. If no sequence from the sample of interest covers the reference position, it is set to missing data. This genotyping scheme may still suffer from ascertainment bias, but reference bias is reduced since the "reference allele" is never considered as a fallback in case of insufficient evidence for the "alternative allele". If sequencing error is randomly distributed, the extend of genotyping error at one-fold coverage will correspond approximately to the sequencing error rate. At higher coverage, true bases will preferentially be sampled as they start to overwhelm erroneous

base calls. Likewise, the different alleles at heterozygous sites will have equal chance of being sampled. The result of this genotyping procedure is an unphased, pseudo-haploid genotype, in which each base has a chance of being erroneous, but all sampled bases that constitute the genotype together allow estimation of relatedness between samples.

Ancient DNA from historical samples is a common source of low-coverage, low quality data, due to degradation and dilution of DNA sequences of interest[38,46,261]. These specimens often times do not allow the collection of more data to achieve higher coverage, as sequence complexity may have been exhausted. Naturally, the sampling of pseudo-haploid genotypes has found frequent use in ancient DNA applications, such as sequencing and genotyping of ancient hominin fossils[46,50]. Still, an easy to use implementation of the procedure which uses standard file formats and is computationally efficient has been lacking. In this chapter we discuss the implementation of genotyping by random sampling of bases. In addition to a procedure based on reference variants, we present an approach to detect sites which show variation de novo. This is applicable in cases where no reference sample set exists, or where ascertainment bias might be a major source of error.

5.2 Analysis of shared derived alleles

Once pseudo-haploid genotypes are sampled, a natural next step is to ask how these genotypes ("query" in the following) relate to individuals of known provenance ("test" in the following). A straight forward way to estimate relatedness is to count the number of bases, or alleles, that different genotypes share with each other. This is especially useful if a phylogenetic hypothesis exists about how test samples are related[46,50]. In the presence of outgroup sequences, this allows the polarization of alleles into ancestral and derived states. One can then count how often a query individual shares a derived state with any combination of test individuals. If the test individuals represent the observable genetic variation well, this test allows the placement of query individuals along the branches of the tree relating the test individuals.

This approach has been useful for example to distinguish sequences from different ancient hominin groups, and allowed their classification as Neanderthal or Denisovan[50,264]. Differences in the amount of shared derived alleles among query individuals has also allowed to estimate relative levels of diversity and divergence.

The analysis of derived allele sharing complements the sampling of pseudo-haploid genotypes especially well, since sampled alleles can be readily compared with the allelic state in outgroups and test individuals. This allows the rapid estimation of genotypes and characterization of relatedness from low-coverage sequencing data.

5.3 bsh-ref and bsh-denovo

To facilitate the genotyping of DNA sequences based on pseudo-haploid sampling, we implemented a set of programs called `bsh` (for "BAM-Sample-Haplotypes"). One ob-

jective for `bsh` was to use standard data formats which readily integrate into genomics and population genetics workflows. Another motivation was to interface with standard libraries to operate on some of these formats, as they provide superior performance over parsing their plain text representations. This is important when trying to scale sample sizes to modern population genetics experiments.

The first `bsh` tool, `bsh-ref`, aims to sample pseudo-haploid genotypes based on already ascertained segregating sites. This is helpful for example in cases where a set of new samples are to be placed in the context of already characterized genetic diversity. Since the sampling of genotypes is usually used for ancient DNA at low coverage, this is especially relevant when placing a set of aDNA samples in the context of modern diversity.

The second tool, `bsh-denovo`, aims to provide a similar analysis pipeline as `bsh-ref` for sample sets without reference data. This means, that positions which segregate within the sampled population need to be identified de novo. Once these segregating sites are identified, `bsh-denovo` allows to sample pseudo-haploid genotypes across the entire sample set.

5.3.1 Data formats

For both `bsh` tools, new samples from which genotypes are to be sampled enter the pipeline as BAM files, containing read mappings to a reference genome[89]. Since no annotations or other genomic features are needed, there is no particular requirement on the quality of the reference genome. It solely acts as a common coordinate system for all samples, so that sites within this coordinate system can be queried regarding the extend of sequence variation in reads covering the position, across samples.

The sites of known variation are supplied to `bsh-ref` in a MAP file, which is a format commonly used in population genetics and was popularized originally by the software package PLINK[265]. Within PLINK, MAP files are usually used in pair with PED files. The MAP file contains the coordinates of variable positions along the reference genome, while the PED file contains the sequence information across a set of samples. This PED file can be supplied to `bsh-ref` as well, and will be integrated with the new genotypes sampled from alignments in the BAM file, based on the coordinates specified in the MAP file. The output is a PED file containing both the reference data set, as well as the newly sampled pseudo-haploid genotypes. Note, that `bsh-ref` allows to pseudo-haploidize the reference data as well, by sampling one of the two alleles at heterozygous positions.

Since `bsh-denovo` aims to identify segregating sites in the sample set, it only requires a BAM file with read alignments of all samples. For computational efficiency, a single BAM file is supplied, which needs to be created by merging single-sample BAM files while specifying the sample identity in each alignments RG-tag ("Read Group"), as specified by the SAM/BAM format specifications. The output is a MAP and PED file pair, containing the coordinates of the newly identified segregating sites in the sample set, as well as the pseudo-haploid genotypes sampled at each of these positions.

5.3.2 Implementation

The backbone of both `bsh` tools is the `htslib` library. It is the standard library for interfacing with many file formats common in genomics, including BAM files, and it is implemented in the `C` programming language. It provides two interfaces for reading BAM files: The `bam_read()` interface allows reading the file one alignment at a time, while the `bam_plp()` interface provides a pileup of all reads covering a reference position. This pileup interface allows to traverse the reference sequence, querying all bases covering an individual reference position. A particular advantage of the BAM format is the possibility to create an index file, which accompanies the alignments and sequence data stored in BAM format. This index, together with the pileup interface, allows random access to all bases covering any reference position, without having to iterate through the entire file.

The use of indexed BAM files is especially important for `bsh-ref`, where the positions which are to be queried from the alignments are already predefined in the MAP file. Therefore, the first step within `bsh-ref` is to read all coordinates from the MAP file and store them. Then, each BAM file is queried at these specified coordinates using the BAM index and the pileup interface, and all bases covering each position of interest are stored. In addition to the base information, `bsh-ref` tracks the length of the sequences each base originated from, its position in the sequence, as well as information on whether or not the alignment showed terminal C-to-T substitutions. This tracking allows for post-hoc diagnostics on the possible effects of aDNA damage on sampled bases. Furthermore, the tracking of terminal C-to-T substitutions allows to optionally restrict sampled bases to sequences carrying such substitutions. This is helpful when the magnitude of damage is high, and the aDNA mixture might be contaminated with modern DNA alignable to the same reference (e.g. ancient DNA from Neanderthal specimens aligned to the human reference genome)[50,167]. Once all positions of interest are queried from a BAM file, a new PED line is written to the output file by sampling bases from the stored sequence information to create a pseudo-haploid genotype for this sample.

Since `bsh-denovo` is aimed at the use-case where no variable positions are known a-priori, all bases have to be queried across all samples to discover segregating sites. First, the header of the merged, multi-sample BAM is used to identify the set of samples present. A hash table is created with the RG-strings as keys, and a custom data struct as values. The RG-strings identify the samples, and are also present in each alignment of a sequence from a given sample. For the hash table, the `khash` implementation included in `htslib` is used. In the merged BAM file which contains information from all samples, the pileup interface is used to traverse the reference sequence from start to end. At each position, all bases associated with each sample are stored in the hash table using the RG-string key. Once all reads covering the current position have been parsed, a set of filters is applied using the sequence information now stored in the hash. These filters allow control over the type of sites which will be identified as variable across the population.

The first filter restricts the extend of missing data that is allowed across samples. The second filter defines the frequency a base must reach within a sample to be considered as

contributing to the bases segregating at this position. This filter allows for example to include or exclude bases which are heterozygous within samples. The third filter sets the minimum frequency the minor allele must reach across samples, for the position to be identified as segregating and to be reported by `bsh-denovo`. If all these filters pass, the position will be reported. For this, the coordinate information is written to the MAP file. Since the PED format reports positions in columns, the PED file is not written immediately as the BAM file is traversed. Instead, a base is sampled randomly for each sample present in the BAM file, and stored in a linked list. Once the entire BAM file is traversed, the PED file is written by traversing the linked list sample-by-sample. After the coordinate of the current position is written to the MAP file, and the sampled bases are stored in the linked list, the information in the hash table is reset to be filled with sequence information at the next position.

In contrast to `bsh-ref`, `bsh-denovo` performs both the discovery of segregating sites as well as the sampling at these newly identified positions. As the sampling of pseudo-haploid genotypes is primarily aimed at low-coverage data, and commonly done in ancient DNA applications, the de novo discovery of sites increases the risk of false positive genotypes driven by sequencing error or ancient DNA damage. During position discovery, this can be mitigated to some extend using the filters described above. Still, to also reduce false positives at the base-sampling stage, `bsh-denovo` implements a sampling scheme popularized under the name "Consensify"[263]. Instead of sampling one base randomly, the consensify method aims to sample three bases, without replacement. if two samples agree on the same base, this base is reported. If three different bases are sampled, missing data is reported. If there is only two bases covering a position, they have to agree to be reported. The minimum coverage using this method is two.

5.3.3 Shared derived alleles

An additional feature implemented in `bsh-ref` is the analysis of derived alleles that a set of query samples share with a set of test samples. To use this feature, three sets of samples have to be supplied to `bsh-ref` through the reference PED file, and the BAM files from which pseudo-haploid genotypes are being sampled (Figure 5.1). At least one outgroup sample has to be defined from the reference set, which will be used to polarize alleles into ancestral and derived states. Ideally though, multiple outgroup samples will be defined, in which case a base will only be called ancestral if it is shared across all these samples. In addition, at least one test sample needs to be supplied from the reference set, which will be used to identify derived states which differ from the allele identified as ancestral in the outgroups. If multiple test samples are defined, all combinations of derived states across all test samples will be identified. The ancestral and derived states identified using these first two sets of samples are then used to identify shared derived alleles in all query samples.

A base at a site is termed "informative", if it has been observed in at least one test sample, and is different from the identified ancestral state. This identifies all derived states in all sets of test samples. A base is termed "shared" in addition to being "informative", if it is identical to the base present in the query sample currently under

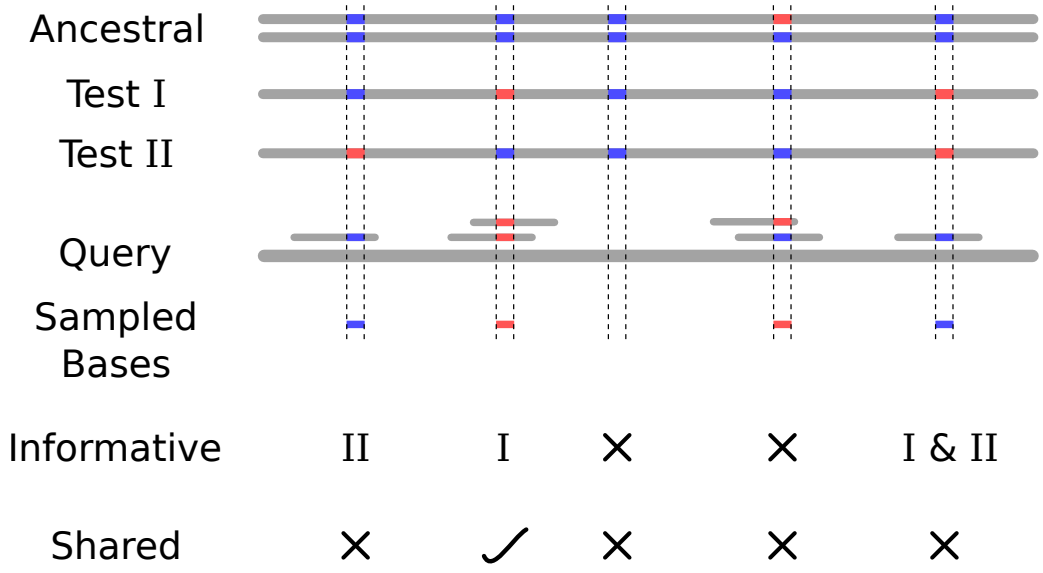

Figure 5.1: Schematic overview of the analysis of shared derived alleles. First, the ancestral allele is established in a set of outgroup individuals. A position is termed "informative" for a set of test individuals (here: Test I and II), if it carries an allele different from the established ancestral state. For each informative site, it is then assessed whether the base sampled from the query individual shares this derived state.

investigation (Figure 5.1). All samples provided through BAM files are automatically used as query samples, but it is also possible to define additional query samples from the reference dataset supplied through the PED file.

In the following, we describe how `bsh-ref` keeps track of derived alleles across all possible combinations of test samples. In a set of n test samples, each base (`A`, `C`, `G`, `T`) can be present in any combination of 2^n test samples, including none of them or all of them. This also means, that there are 2^n combinations of test samples, with which a query sample may share a derived allele. To keep track of counts of either "informative" or "shared" sites within a sample, `bsh-ref` first allocates arrays of length 2^n (Listing 5.1).

At each position of interest, the pattern of alleles shared across sets of test samples is established. This is done using four bit-arrays, one for each base, of type `unsigned long long`. The bases in each test sample are traversed in order, and each test sample corresponds to a bit position. If a sample's bit is set for one of the four bases, this means the base was present in this test sample at the current position. The combinations of set bits corresponds to test samples sharing this base (Listing 5.2). This implementation restricts the number of test samples which can be queried for their derived alleles to the number of bits in an `unsigned long long`, which on a standard 64-bit architecture is 64. This number greatly exceeds the amount of distinct test samples one would normally query against (Listing 5.1, lines 4-7). Of course, it also exceeds the amount of memory

```
int n_test = n_test_indiv;
int n_comb = pow(2, n_test);

if (n_test > CHAR_BIT * sizeof(unsigned long long)) {
  /* Too many test samples to keep in ULL bit array */
  return -1;
}

int *infor = NULL;
int *share = NULL;
infor = calloc(n_comb, sizeof *infor);
share = calloc(n_comb, sizeof *share);
```

Listing 5.1: Bit-array for shared derived allele test - part 0. Allocating count arrays to keep track of informative and shared site counts across sites.

that can practically be allocated for the count arrays of length 2^n, and no more than 10 test individuals would generally be utilized for this test.

```
char *b = array_of_test_sample_bases;
int i, k;

char ref[4] = {'A','C','G','T'};
unsigned long long pbase[4] = {0,0,0,0};
for (i = 0; i < n_test; ++i) {
  for (k = 0; k < 4; ++k) {
    if (b[i] == ref[k]) {
      /* Set pbase bit for current test sample */
      pbase[k] |= 1 << i;
    }
  }
}
```

Listing 5.2: Bit-array for shared derived allele test - part 1. Tracking of shared bases across test samples, using the bit-array `pbase[4]`.

The ancestral base, once established, is tracked in a second bit array together with the base from the query sample under investigation (Listing 5.3). This makes the final step of the analysis much easier.

The next step after establishing the combinations of bases observed in test samples, as well as the ancestral state, is to assess whether there are “informative” and “shared” bases. A base is “informative” for a given set of test samples where it is present, if it is different from the ancestral state (Listing 5.4 lines 9-11). It is “shared” with that set of test samples, if it is “informative”, and identical to the sampled base (Listing 5.4 lines

```
1  char s = sampled_base;
2  char anc = ancetral base;
3
4  int sbase[4] = {0,0,0,0};
5  for (k = 0; k < 4; ++k) {
6    /* Keep track of sampled base and ancetral base */
7    if (s == ref[k]) sbase[k] |= 1;
8    if (anc == ref[k]) sbase[k] |= 2;
9  }
```

Listing 5.3: Bit-array for shared derived allele test - part 2. Tracking of sampled- and ancestral base, using the bit-array `sbase[4]`.

9, 12-14). The interesting detail here is, that the decimal representation of the bit array can be used as an index to the integer-arrays counting "informative" and "shared" sites. These indexes identify each combination of test samples uniquely (Listing 5.4 lines 2-5, 11, 14).

```
1  /*
2    Example: derived base A is shared in test samples 1 and 2
3             => bit array pbase[0] = .... 00000011
4             this is 3 in decimal
5             infor[3] and share[3] correspond to combination 1-2
6  */
7
8  for (k = 0; k < 4; ++k) {
9    if (pbase[k] && (sbase[k] >> 1) ^ 1) {
10     /* Base is set, and different from ancestral (XOR with 1) */
11     ++infor[pbase[k]];
12     if (sbase[k] & 1) {
13       /* Sampled base is the same as derived base in pbase[k] combination. */
14       ++share[pbase[k]];
15     }
16   }
17 }
```

Listing 5.4: Bit-array for shared derived allele test - part 3. Assessment of informative and shared sites from the previously established bit-arrays.

Once all sites of a query sample are analyzed, the "informative" and "shared" counts are reported to an output file.

5.4 Discussion

The `bsh` tools presented here allow the estimation of pseudo-haploid genotypes through sampling. This is useful in cases, where stringent, high quality variant detection is either not possible due to the data types collected, or not necessary due to the scientific question a user is trying to answer. Such a question may be the broad scale genetic classification of a set of samples, for example by integrating the genotypes estimated with `bsh` with a set of reference genotypes of known provenance. This is especially common in ancient DNA applications, where a historical sample should be placed in the genetic context of either other historical samples[50,264], or modern diversity[266]. Also common for ancient DNA is genetic data of low quality, which does not facilitate the inference of high-quality genotypes[46,261]. For these cases, the random sampling of bases at variable positions tries to facilitate the inference of relatedness in the presence of error with as little bias as possible. In ancient DNA, an additional source of error is caused by post-mortem degradation, especially due to the conversion of cytosine to uracil[38]. Several additions to the `bsh` tools are motivated by the presence of this type of age-associated DNA damage. These include the diagnostics produced by `bsh-ref` which report the positional bias within reads of bases sampled, the option to sample bases from reads conditional on the presence of substitutions consistent with the presence of age-associated damage[50], as well as the inclusion of the "consensify"[263] method for base sampling in `bsh-denovo`.

The `bsh` tools are widely applicable to cases where genotypes are to be estimated by sampling, especially due to the use of standard file formats. Both input and output files use formats which are commonly produced and consumed by pipelines used in evolutionary genomics. No user manipulation is needed, and all up- and downstream operations can be done with widely used tools such as `samtools`[89] and `plink`[265]. The only input which needs to be created manually is the file defining samples for testing the amount of shared derived alleles offered by `bsh-ref`. In addition to the file formats used, the `bsh` tools are implemented in the `C` programming language. This is partly for easier access to the standard routines for interfacing the BAM files, as supplied by `htslib`, but also to optimize run time and memory requirements. As the sample sizes for experiments in population- and evolutionary genetics increase, it is important for tools such as `bsh`, which are aimed to facilitate a first look into the genetic diversity captured by a sample set, to run quickly to allow the rapid creation and testing of hypotheses.

While the sampling of pseudo-haploid genotypes at variable positions has been used before[46,50], previous implementations were often not shared with the community, did not use standard file formats, or did not aim for computational efficiency and usability. The `bsh` tools are publicly available, with detailed usage instructions and the ability to report bugs and request features (`https://github.com/clwgg/bsh-ref` and `https://github.com/clwgg/bsh-denovo`).

`bsh-ref` was originally conceived for the project presented in the following Chapter 6. The use case for this project was exactly as described for `bsh-ref`: A reference panel of genetic diversity at a defined set of variant positions was known, and bases were to be sampled at these sites from low coverage short read alignments. The use case of

`bsh-denovo` is complementary to that, as it allows the discovery of variable sites if no reference data is available. Usually, variable sites are discovered by procedures analogous to the identification of high-quality variants. `bsh-denovo` offers the identification of variable sites for cases where data quality may not suffice for using standard variant calling infrastructure, just as `bsh-ref` offers the estimation of pseudo-haploid genotypes from low-quality data in the presence of reference genotypes.

Apart from sampling pseudo-haploid genotypes, `bsh-ref` also implements counting the derived alleles that query samples share with a set of test samples. This facilitates quick analysis of the relatedness of samples of interest with defined reference samples. The application of it in Chapter 6 will show that this simple analysis shows high congruence with other common methods of assessing the relatedness of genetic samples and demonstrate its use for explorative analysis.

The current implementation of counting shared derived alleles assesses all combinations of test samples. While this is interesting for some applications, a common use case for an analysis of this type is to query shared derived alleles along some pre-defined topology of how samples are related. Alternatively, the counts of shared derived alleles may inform a user on a possible tree topology if none is known a-priori. For these applications, a future development of `bsh-ref` could be to add the ability to interact with tree topologies, either to restrict the counting of derived alleles to combinations concordant with the topology, to infer possible topologies by assessing derived alleles shared among test samples, or to evaluate the presence of discordant alleles explicitly.

Another possible extension is the capacity to define populations to test, rather than individual samples. This would allow the extension of these analyses to allele frequencies instead of allele counts, similarly to the widely popular family of f- and D-statistics[46,125,126].

6 Sedimentary ancient DNA

Contributions

The project presented in this chapter developed out of a collaboration published in the article "Neandertal and Denisovan DNA from Pleistocene sediments"[198]. The content of this chapter is unpublished at the time of submission, and the following people have contributed:

Hernán Burbano (HB), Matthias Meyer (MM) and myself conceived and designed the project. HB and myself selected the sequences to target in the capture experiment. MM and Viviane Slon designed the laboratory experiments, which were performed by Brigit Nickel to generate the sequencing data. Marie Soressi led the excavation at Les Cottés and provided archaeological expertise. I analyzed the data and generated all figures.

6.1 Sequencing of sedimentary ancient DNA

Even before the advent of second generation sequencing, researchers were interested in investigating the suitability of environmental DNA as a source of genetic sequence information from organisms long after their death[59]. The term environmental DNA summarizes DNA extracted not from specimens like mammalian bones or plant leaves, but directly from environments such as sediments from archaeological excavations[198], from permafrost[267], or from lake beds[145].

DNA from these sources presents a great opportunity to recover sequence information which is otherwise difficult to acquire, for example, due to the scarcity of mammalian fossil remains. This is especially true also for remains from different hominin lineages. For instance, the Denisovan lineage has only been characterized based on DNA sequences from a handful of remains in one cave in the Siberian Altai mountains[264]. However, many archaeological sites can be associated to hominin and modern human industries, even if no skeletal remains can be found. For these sites, there is great potential in the ability to characterize DNA associated with these sediments[198].

This potential is by no means limited to hominin or mammalian species. In contrast to bones preserved from these organisms, there has been limited success in extracting DNA from fossilized plant remains[64]. This is in part due to the intrinsically low DNA concentration in wood[66], but also due to the limited preservation of other plant remains. For example, seeds associated with human inhabitation are often charred, which negatively affects DNA preservation[65]. In addition, seeds which are not charred or processed in some way may quickly be eaten by animals, rather than preserved over thousands of years. For these cases, environmental DNA which might be deposited at a site through

decaying leaves or roots can present an opportunity to interrogate ancient plant DNA[147]. Such DNA evidence can then be incorporated with other archeobotanical evidence to reconstruct paleo-environments[145].

However, ancient DNA extracted from environmental sources comes with many difficulties, as discussed in previous chapters. For example, the complex mixture of DNA will be composed of molecules of ancient as well as modern origin, which can be difficult to tease apart. This can be caused by the high abundance of modern microbial DNA, which dilute ancient molecules from taxa of interest. The low abundance of sequences of interest makes it difficult to assess their genetic information, but also their authenticity. As contamination is a constant problem in all of ancient DNA research, it is of utmost importance to present evidence of authenticity for sequences from all taxa whose presence informs the conclusions drawn from these analyses.

When extracting ancient DNA from plant or animal remains, it is often possible to date the specimen from which DNA is extracted, by methods such as radiocarbon dating[268]. Sediments can also be dated directly[269], but it is more difficult to associate the age of the sample with the age of the DNA contained in it. In sediments, DNA can be bound directly to soil particles[270,271], or be contained in microscopic remains embedded in the sediment[146,198]. As such, it is more difficult to associate the age of the sediment directly with the age of the DNA, as movement of DNA, though infrequent, has been observed[272,273]. This makes it especially important to interpret DNA results in the archeological context of the site, and the temporal context of the DNA found[146,147]. For example, dense sampling across the stratigraphy of sites which contain DNA from extinct organisms has shown, that such DNA was associated only with layers dated to periods compatible with the extinction history of the associated organisms[198].

These difficulties often make it more practical to employ targeted approaches to recover DNA from a defined set of taxa. This can be achieved through targeted hybridization capture, which enriches DNA mixtures in molecules which hybridize with a library of bait sequences[44]. An example of such a study was presented by Slon et al. [198] in 2017, where sediments from a set of seven archeological sites from the Pleistocene were queried for a range of mammalian mitochondrial sequences (Figure 6.1A). These sequences allowed to investigate the distribution of major animal taxa, including hominins. Since the mitochondria of modern humans, Neanderthals and Denisovans carry diagnostic sites, it also allowed to assess DNA sequences of hominin mitochondria, without the need to rely on skeletal remains. Since these mitochondrial sequences were enriched over the predominantly microbial DNA molecules in the sediments by targeted capture, it also permitted the authentication of mammalian and hominin sequences through the presence of age-associated degradation patterns.

In addition to the mitochondrial captures, Slon et al. [198] prepared shotgun sequencing libraries for a subset of DNA extracts from all caves. These allowed the characterization of the overall taxonomic composition of sequences, but also facilitated more explorative investigations of taxa of potential further interest. When investigating the abundances of major plant genera for example, we detected the *Vitis* genus as an outlier in the Les Cottés site in western France (Figure 6.1B).

6.1.1 The Les Cottés site

The Les Cottés site is located in western-central France, in the southwestern part of the Parisian Basin (Figure 6.1A). It is located in front of the entrance to a cave, which was first excavated in 1881[274]. The following excavations however, including those which produced the sedimentary DNA analyzed by Slon et al. [198], focused on the area in front of the cave, and the cave entrance.

One of the unique features of this site is the association of several hominin industries with different strata along its well established stratigraphy[274]. Lower layers are associated with the Mousterian and Chatelperronian, which were produced by Neanderthals not long before their disappearance from the fossil record. Above these are layers associated with the Protoaurignacian and Early Aurignacian, which are attributed to early modern humans. Layers associated with hominin occupation are interspersed with archaeologically sterile layers. The entire stratigraphy has been dated by radiocarbon dating and luminescence-based methods, with age estimates ranging from 45-55kya for the lower layers, and 30-40kya for the upper layers[274,275] (Figure 6.3A). Climatically, the lower layers are associated with an open steppe environment, while upper layers resemble a more arctic environment approaching the last glacial maximum.

Despite the richness in archaeological artifacts from a multitude of hominin industries, skeletal remains have been found only from an anatomically modern human. Faunal remains have also been characterized, several of which were recovered using mitochondrial DNA capture by Slon et al. [198]. In the layers associated with Neanderthal occupation however, no hominin mitochondrial DNA was recovered.

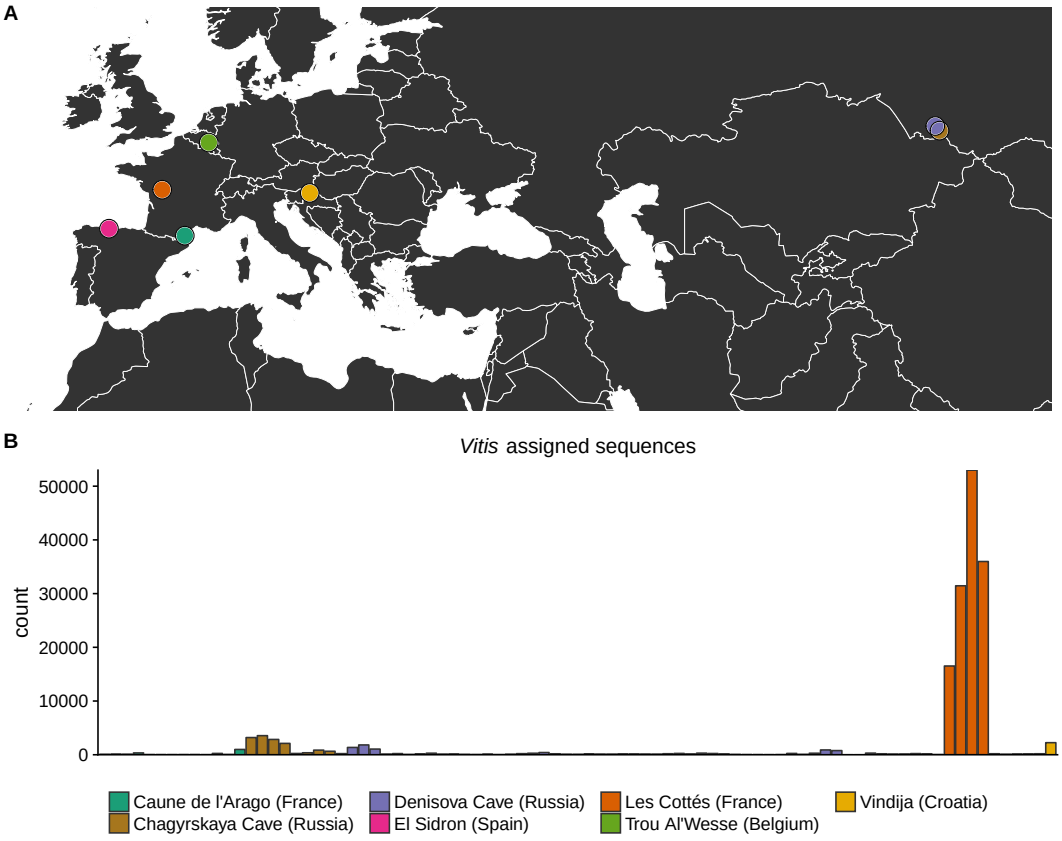

Figure 6.1: Initial screening of caves investigated by Slon et al. [198] A. Distribution of archaeological sites from which sedimentary DNA was analyzed. **B.** Counts of sequences assigned to the *Vitis* genus or lower taxonomic levels.

6.2 *Vitis vinifera*

We screened the shotgun sequencing data from sedimentary DNA produced by Slon et al. [198] for the abundance of several major plant genera. The only outlier we found, which rose above a basal level of noisy assignments, was the genus *Vitis* (Figure 6.1). This genus is represented in the databases primarily by the reference genome of the common grape vine *Vitis vinifera*, which was sequenced from an inbred line derived from cultivated Pinot Noir[276].

Globally however, the *Vitis* genus encompasses about 60 largely inter-fertile wild species, with the two main centers of diversity in the southwest of North America as well as east Asia[277]. The only species native to Europe is *Vitis vinifera*, with a distribution spreading from western Asia across the Caucasus and throughout the Mediterranean. Among wild *Vitis* species, nuclear sequence markers suggest that the European *vinifera* represents the most derived representative of the genus, with North American species being most ancestral[277]. In addition, wild *Vitis vinifera* is also the closest relative of

cultivated grape, which is distinguished as a subspecies: wild *Vitis vinifera* is referred to as the *subsp. sylvestris*, and its domesticated relative is called *subsp. vinifera*[278].

Domesticated grapes of *subsp. vinifera* come in several forms, most importantly to be processed into wine, but also for fresh or dried consumption of fruits[279]. Grape cultivation and domestication is thought to have started around 8,000 years ago in the Caucasus region[280]. Historical evidence for the use of grapes in wine production is supported by archeological and chemical evidence from this time period, primarily from studying storage jars[281,282,283]. From this initial center of domestication, early cultivars spread westwards throughout the Mediterranean over the following millenia[278,284], where the production and consumption of wine became an integral component of cultures such as the ancient Greeks and the Roman Empire[285,286].

While modern cultivars still primarily have ancestry most closely related to eastern *Vitis vinifera subsp. sylvestris*, there has been evidence that especially cultivars from western Europe carry ancestry of western wild *subsp. sylvestris* as well[279], suggesting a more complex ancestry of cultivars in the west. In addition, more recent breeding efforts to increase disease resistance in cultivars have included other wild species of the *Vitis* genus[278].

To be able to characterize these diverse ancestries in modern *Vitis vinifera* cultivars, Myles et al. [287] designed a genotyping array based on whole genome sequencing data generated using reduced representation sequencing. A set of 8,898 SNPs was selected based on this whole genome data, with the aim to allow the differentiation of wild *Vitis* species, as well as *Vitis vinifera* cultivars[287]. This Vitis9KSNP array was subsequently used to genotype a large set of samples from the USDA grape germplasm collection, including over 1,000 cultivars as well as wild *Vitis vinifera subsp. sylvestris* and a large variety of other wild *Vitis* species[279,287,288].

The availability of such a comprehensive dataset of *Vitis* genetic diversity made it both feasible and interesting to further investigate the putative *Vitis* sequences found in the sediments of Les Cottés.

6.3 Methods

6.3.1 DNA extraction, sequencing and processing

DNA extraction and sequencing was performed as described by Slon et al. [198]. In brief, we used an extraction protocol that efficiently recovers short DNA sequences expected in ancient DNA[51]. From these aDNA extracts, single stranded, double-indexed libraries were produced in an automated fashion as described by Gansauge et al. [55].

From these libraries, 76-cycle, paired-end sequencing was carried out with parameters suited to double-indexed libraries[169] on MiSeq or HiSeq 2500 v3 platforms (Illumina). Base calling was carried out using Bustard (Illumina) or freeIbis[289]. Adapters were trimmed, and overlapping forward and reverse reads were merged using leeHom[210]. Only sequences carrying the exact expected indexing combinations were retained for downstream processing. For further analysis we established a minimal length cutoff of

30 base pairs to decrease the number of spurious alignments.

6.3.2 Taxonomic characterization of shotgun data

Sequences were aligned to the full non-redundant nucleotide (nt) collection of the National Center for Biotechnology Information (NCBI) (downloaded January 2015) using the MEGAN alignment tool (MALT, version 0.0.12[195]) in BlastN mode. MALT uses a BlastN-like algorithm to generate local alignments of all obtained sequences to matching DNA sequences in a given database. After generating alignments, a lowest common ancestor (LCA) algorithm[137] is used to bin sequences along the NCBI taxonomy. The LCA assigns a sequence to the node in the taxonomy that is the lowest common ancestor of all nodes, which have alignments with bit-scores within 10% of the best alignment. This means that sequences with unspecific alignments over multiple taxa end up with assignments at higher taxonomic levels.

MALT saves the alignments, as well as the binning information in an RMA file, which can then be processed and inspected using the Metagenome Analyzer MEGAN (version 5.11.3[137]). MEGAN allows visual inspection of how sequences are binned along the taxonomy, and provides a command line interface (CLI) for basic operations involving RMA files.

Here, MEGAN was used to extract the number of sequences that were assigned to different plant genera, as an initial screening step to investigate the composition of plant taxa in DNA from cave sediments.

6.3.3 Targeted hybridization capture

6.3.3.1 Probe design

A set of 8,898 Single Nucleotide Polymorphisms (SNPs) was established by Myles et al. [287], and included on the Vitis9KSNP array. This array was used to characterize the diversity of the *Vitis* genus[279]. After filtering, a dataset of 6,114 SNPs was available from this genotyping effort (supplied by the authors, contact: Sean Myles). This SNP set was used to design probe sequences, with the aim to capture molecules overlapping these variable positions. For each SNP, 12 probes of 60bp length were designed according to the surrounding sequence, with 5bp tiling. Half of these 12 probes carried the reference allele, and the other half carried the alternative allele of the targeted site.

In addition, the entire *Vitis vinifera* chloroplast sequence was used to design an additional set of 60bp probes, with 10bp tiling. To fill up the array, a random subset of 500 additional nuclear SNPs was selected, for with probes were designed in accordance to the original set of 6,114 SNPs.

All probe sequences were subjected to repetitiveness filtering based on genome-wide frequencies of 15-mers, as established by Hodges et al. [290] and Burbano et al. [44]. In addition, only unique probes were included in the final design.

6.3.3.2 Capture procedure

The final set of 92,838 probes were ordered to be synthesized from CustomArray, Inc. The synthesized 102bp oligos included the designed bait sequences (60bp), as well as adapters for PCR primers APL2 and APL6 as described in Fu et al. [291] (2x 21bp). These adapters allow the amplification of the probe set, as well as the incorporation of biotin to be used to capture DNA using Streptavidin beads. These biotinylated probes were used to capture ancient DNA libraries following Maricic, Whitten, and Pääbo [45], with the modifications described in Slon et al. [198].

Sequencing and initial sequence processing of these enriched libraries was performed in accordance to the protocol used for the initial shotgun libraries.

6.3.4 Mapping of short read data

Sequences which passed the 30bp length cutoff after successfully merging the forward and reverse reads were mapped to the *Vitis vinifera* reference genome version 8x[276] using bowtie2 (version 2.2.4[216]) with default parameters. The 8x version of the reference genome was used, since the positions of variants included in the Vitis9KSNP array were established based on this version.

Potential PCR duplicates were removed based on identical start- and end coordinates of mapped sequences, using the DeDup utility of the EAGER pipeline (version 0.12.0[292]).

6.3.5 Analysis of chloroplast sequences

Merged sequences above 30bp length from all captured libraries were individually mapped to the *Vitis vinifera* chloroplast sequence (NC_007957[293]) using BWA-MEM (version 0.7.10[177]) with default settings. Successfully aligned sequences were extracted, and classified taxonomically using MALT as described for shotgun data in Section 6.3.2. The CLI of MEGAN also described in this section was used to extract all sequences assigned to the *Vitis* genus or lower taxonomic levels.

This procedure should result in a set of sequences which is highly curated to include only sequences with high sequence similarity to the *Vitis vinifera* chloroplast. To reduce alignments to nuclear insertions of chloroplast sequences into the *Vitis vinifera* genome, we mapped this set of sequences to the reference genome, as well as the chloroplast genome, using bowtie2, and removed putative PCR duplicates as described in Section 6.3.4.

This dataset was used to call variants in the chloroplast genome in all samples with an average chloroplast coverage of above five, using the bcftools (version 1.8,[217]) utilities mpileup ("bcftools mpileup -q 1 -I -Ou -a 'AD' -r 'chloroplast' -f $REF $IN") and call ("bcftools call -v --ploidy 1 -m -O v").

6.3.6 Assessment of deamination patterns

The characteristic deamination pattern in ancient DNA, which is observable from sequencing data by the presence of C-to-T conversions at the 5' end of sequences, were assessed using mapDamage (version 2.0.2-12[180]). The frequency of C-to-T conversions at the 5'-end was extracted from the mapDamage output file "5pCtoT_freq.txt", while the frequency of C-to-T conversions at the 3'-end was calculated from the output file "misincorporations.txt" directly (since single-stranded libraries show C-to-T damage patterns at both ends).

6.3.7 Assessment of capture target coverage

To assess the success of retrieving nuclear SNP data using hybridization capture, it was important to assess the sequencing depth at positions targeted by the capture. For this, the coordinates of targeted SNPs were converted to BED format, and the bedtools utility `coverageBed` (version 2.24.0[294]) was used to retrieve sequencing depth at the sites defined in the BED file (supplied with the `-a` option), using the `-hist` output option.

6.3.8 Sampling of variable positions

The dataset of *Vitis* genomic diversity as assessed by Myles et al. [279] using the Vitis9KSNP array was used as the reference dataset, on the basis of which we assessed the diversity in captured libraries. It was used in the PLINK text format, which can act directly as input to the base sampling method `bsh-ref` described in Chapter 5. In cases were pseudo-haploid genotypes were to be analyzed, the reference genotypes were haploidized by choosing one of the alleles at heterozygous sites at random. From the captured libraries, genotypes were sampled directly from read alignments using `bsh-ref`. Additionally, these datasets were filtered using PLINK (version v1.90b4.1)[265,295], when applicable. The sampling and filtering was used to generate different sets of variants, which will be described in the following.

6.3.8.1 Variant sets and filters

- Initial capture PCA

Read mappings from the initial sample set of two libraries from each of the layers US06 and US08 were merged either all together, or merged by layer. From these mappings, one base was sampled per merged dataset, at variable sites based on the *Vitis* diversity reference set. These sampled pseudo-haploid genotypes were combined with haploidized genotypes from all 1,031 *Vitis vinifera subsp. vinifera* samples and all 69 *Vitis vinifera subsp. sylvestris* samples included in the reference set. This sample set was used for the PCA displayed in Figure 6.2D.

- All libraries PCA

Pseudo-haploid genotypes were sampled from read mappings of each of the 35 captured libraries. The genotypes of 69 *subsp. sylvestris* samples were haploidized as well. After sampling, PLINK was used to filter out individuals with more than 40% missing data, and variants with more than 5% uncalled individuals, since some libraries in the full set had only few sequences alignable to the *Vitis vinifera* reference. This filtering removed 15 of the 104 individuals, and 2,785 of the 6,114 variants. The filtered, pseudo-haploid dataset was used to produce the PCA in Figure 6.8.

- Shared derived alleles

For the analyses of shared derived alleles (Figures 6.4,6.7, 6.11), the full reference dataset of 1,817 samples was haploidized, and combined with 10 replicates of pseudo-haploid samples of the entire set of 35 captured libraries. For the analysis presented in Figure 6.11B, the base sampling from captured libraries was restricted to sequences carrying at least one C-to-T conversion at the 5'- or 3'-end (since single-stranded libraries show C-to-T damage patterns at both ends).

- D-statistics

For the population-based D-statistics (Figure 6.5), the full reference dataset was first filtered for variants which had Hardy-Weinberg equilibrium exact test mid-p values[296] of above 0.01, and that had less than 10% of uncalled individuals. This was done primarily to remove variants with highly skewed allele frequencies due to cultivation and clonal propagation of *Vitis vinifera subsp. vinifera* individuals, which dominate the reference dataset. The remaining 1,041 variants were haploidized, and combined with sampled, pseudo-haploid genotypes from all captured libraries. These were again filtered to remove individuals with more than 40% missing data, and variants with more than 5% uncalled individuals, which removed 19 individuals, and no additional variants.

- Genetic distance

For the analysis of within-library genetic distance (Figure 6.9), the full reference dataset and 10 replicate samples of all captured libraries were haploidized, and filtered to remove individuals with more than 40% missing data, and variants with more than 5% uncalled individuals.

- Nucleotide diversity

To analyze nucleotide diversity (Figure 6.10), the full reference dataset was used in diploid state, and two alleles were sampled for each captured library. These genotypes were filtered to remove individuals with more than 40% missing data, and variants with more than 5% uncalled individuals. Diploid genotypes were used for this analysis, to more accurately represent the level of diversity in the reference dataset populations.

6.3.9 Principle component analysis

Principle component analyses were conducted with smartpca from the EIGENSOFT package (version 6.1.4[106]), using only genotypes from *Vitis vinifera subsp. sylvestris*. Genotypes from additional samples, such as those captured from Les Cottés, were then projected into this space (also implemented in smartpca).

6.3.10 Analysis of shared derived alleles

The fraction of derived alleles that a query sample shares with a set of test samples representing the major clusters identified in the PCA in Figure 6.2D was assessed using `bsh-ref` following the methodology described in Chapter 5. First, test samples were identified for each cluster (*subsp. sylvestris* east, west, Spain, as well as *subsp. vinifera*), by their minimal within-cluster euclidean distance in the first two dimensions of the PCA (Figure 6.2D). Next, suitable outgroup individuals were chosen by selecting the individuals with the lowest missing data from the wild *Vitis* species *Vitis rotundifolia* and *Vitis cinerea*, which are both outgroups to the entire *Vitis vinifera* diversity[277].

Using this set of test- and outgroup individuals, the fraction of shared derived alleles was assessed in 10 replicate samples of all captured libraries, as well as all *subsp. sylvestris* samples from the reference set using `bsh-ref`. These query samples were filtered by the number of informative sites at the root of the *Vitis vinifera* clade (see also Chapter 5), which was required to be above 100 (above 30 for the C-to-T restricted filtering).

6.3.11 D-statistics

For the calculation of D-statistics, all libraries from Les Cottés were treated as one population. The set of 63 *Vitis rotundifolia* samples was used as an outgroup. The D-statistics, as well as the blocked-jackknife based standard error, were calculated using POPSTATS[297].

6.3.12 Analyses of genetic distance and diversity

Genetic distances between all samples were calculated using PLINK with the `--distance` option. Nucleotide diversity was calculated using the `--window-pi` utility of vcftools (version 0.1.12b)[96], with a window size of 1Mb.

6.4 *Vitis* in Les Cottés sediments

6.4.1 Shotgun sequencing of sedimentary DNA

Shotgun sequences generated from sedimentary DNA extracts from all seven caves were scanned for the abundance of different plant genera, using alignment-based taxonomic binning based on the NCBI taxonomy and the NCBI nucleotide BLAST database. The only genus which was over-represented in one of the caves was the *Vitis* genus, which was at high abundance in sequences from the Les Cottés cave (Figures 6.1, 6.2A). This genus is primarily represented in databases by the grape vine *Vitis vinifera*, which includes cultivated grapes. Direct alignments of sedimentary DNA sequences from Les Cottés to the *Vitis vinifera* reference genome showed sequences to be randomly distributed along the genome, with low edit distances between aligned sequences and the reference. In addition, the sequences showed deamination patterns characteristic for ancient DNA, which are observable in read alignments of sequences generated by single-stranded library preparation as C-to-T conversions at both the 5'- and the 3'-ends of sequences (Figure 6.2B).

These observations suggested that sedimentary DNA from Les Cottés contains nuclear ancient DNA closely related to *Vitis vinifera* at appreciable abundance. The fraction of *Vitis* sequences generated using a shotgun approach was however less then 0.5%, making a brute-force approach of sequencing nuclear data from these extracts impractical. To estimate the potential depth of *Vitis* sequences retrievable from these extracts, we used qPCR to quantify the number of molecules present after adapter ligation, which yielded between $2 - 4 \times 10^9$ molecules per Les Cottés library. The number of molecules in the extract, together with the fraction of *Vitis* sequences of $0.1\% - 0.5\%$, and an average fragment size of ~50bp allowed us to estimated that each position of the *Vitis vinifera* genome could be covered by $1 - 2$ sequences per Les Cottés DNA extract.

This estimate, as well as the existence of a reference panel of *Vitis* diversity[279], led us to characterize these nuclear sequences in more detail using targeted hybridization capture.

6.4.2 Targeted capture of genomic DNA

The *Vitis* reference dataset by Myles et al. [279] was used to design a targeted capture experiment to retrieve nuclear *Vitis* SNP data from the Les Cottés sediments. Probe sequences were selected to be complementary to the regions flanking 6,614 SNPs characterized on the Vitis9KSNP genotyping array, and used for in-solution capture of sedimentary DNA extracts. This was effective in recovering nuclear sequence data for between 50% and 80% of targeted SNPs covered by 1 to 3 sequences (Figure 6.2C). In addition, it was specific to extracts which had shown an abundance of *Vitis* sequences in the shotgun data, as no SNP data was recovered from extracts from Trou Al'Wesse, which was used as a negative control.

From this low-coverage data, we generated pseudo-haploid genotypes by sampling bases aligned to polymorphic sites targeted by the capture. These genotypes were used

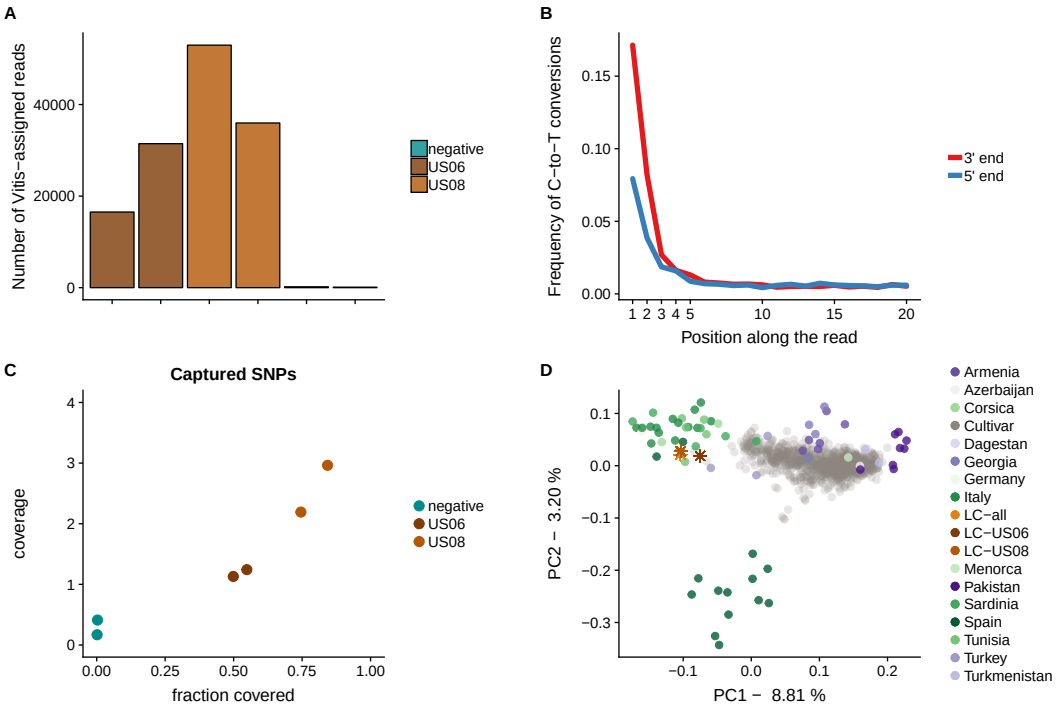

Figure 6.2: Non-targeted taxon discovery and initial capture. A. Number of reads assigned to the Genus *Vitis* or lower taxonomic levels, using MALT on the NCBI Blast nt database. Shown are the four libraries from Les Cottés, as well as two libraries from a different cave site (Trou Al'Wesse) as a negative control. **B.** C-to-T substitution frequencies in a pool of all four Les Cottés libraries, at the 5'- (blue) and 3'-end (red) of sequenced fragments. **C.** The covered fraction of SNPs targeted using in-solution capture, plotted against the average coverage of covered SNPs for the Les Cottés libraries, and Trou Al'Wesse as a negative control. **D.** The first two axes of a Principle Component Analysis conducted on genotypes of *Vitis vinifera subsp. sylvestris* are plotted against each other, with *Vitis vinifera subsp. vinifera* and Les Cottés genotypes projected into this space. Western *sylvestris* samples are shown in shades of green, while Eastern *sylvestris* samples are shown in shades of purple. The *vinifera* cultivar samples are shown in grey.

to put the Les Cottés sequences into the context of modern *Vitis vinifera* diversity using a Principle Component Analysis (PCA). Principle components were calculated using genotypes from individuals of the wild relative of modern cultivated grape, *Vitis vinifera subsp. sylvestris*. Into this space, both the sampled genotypes from Les Cottés, as well as modern cultivar genotypes of the subspecies *Vitis vinifera subsp. vinifera* were projected (Figure 6.2D). This showed, that sequences from Les Cottés were most closely associated with wild *Vitis vinifera subsp. sylvestris* from western Europe, compatible with the provenance of these samples.

In addition to the sites targeted in the *Vitis vinifera* nuclear genome, we also included the entire chloroplast sequence in our probe set. To investigate the suitability of these sequences to identify phylogenetically informative variants, we first aligned all sequences to the chloroplast reference genome, and classified all aligning sequences taxonomically. The assignments of sequences along the NCBI taxonomy showed, that the vast majority of sequences were assigned to higher taxonomic levels along bacterial and plant lineages, suggesting that these sequences had high quality alignments to a wide variety of sequences in the database (Figure 6.6C). We then identified variants in the subset of sequences assigned to the *Vitis* genus or lower taxonomic levels. In chloroplast sequences from a single individual, one expects the frequency of variants in the chloroplast to approach one at higher coverage, since the chloroplast genome is haploid. However, in chloroplast sequences captured from sediments, we identified more intermediate allele frequencies in libraries sequenced at higher coverage (Figures 6.6A,B). These results, as well as the taxonomic assignments, suggest that captured sequences alignable to the *Vitis vinifera* chloroplast represent a complex mixture of sequences from a variety of taxa, and even those which are assigned to *Vitis* are genetically diverse.

6.4.3 Genomic variation along the stratigraphy

The initial success in retrieving nuclear sequence data from sediments led us to investigate the distribution and diversity of *Vitis* sequences along the Les Cottés stratigraphy, which constitutes a gradient of age, climate, and association with hominin occupation of the cave (Figure 6.3A). The fraction of targeted SNPs recovered from the full set of 35 DNA extracts ranged from almost 0% to almost 100% (Figure 6.3B), and nuclear sequence data was recovered from all layers carrying comparable amounts of C-to-T conversions at first base (Figure 6.3C).

Again, we generated pseudo-haploid genotypes for each of these libraries by randomly sampling bases overlapping polymorphic sites. In the space spanned by the first two principle components calculated using *Vitis vinifera subsp. sylvestris* genotypes from the reference data set, these Les Cottés genotypes were again most closely associated with western *subsp. sylvestris*, without any discernible structure between layers (Figure 6.8).

To analyze the genetic relationship of the libraries from Les Cottés with the *Vitis vinifera* samples in the reference panel in more detail, we assessed the extend to which they share derived alleles with individuals representing the major clades of *subsp. sylvestris*, as well as the cultivar *subsp. vinifera*. On average, the Les Cottés sequences again showed

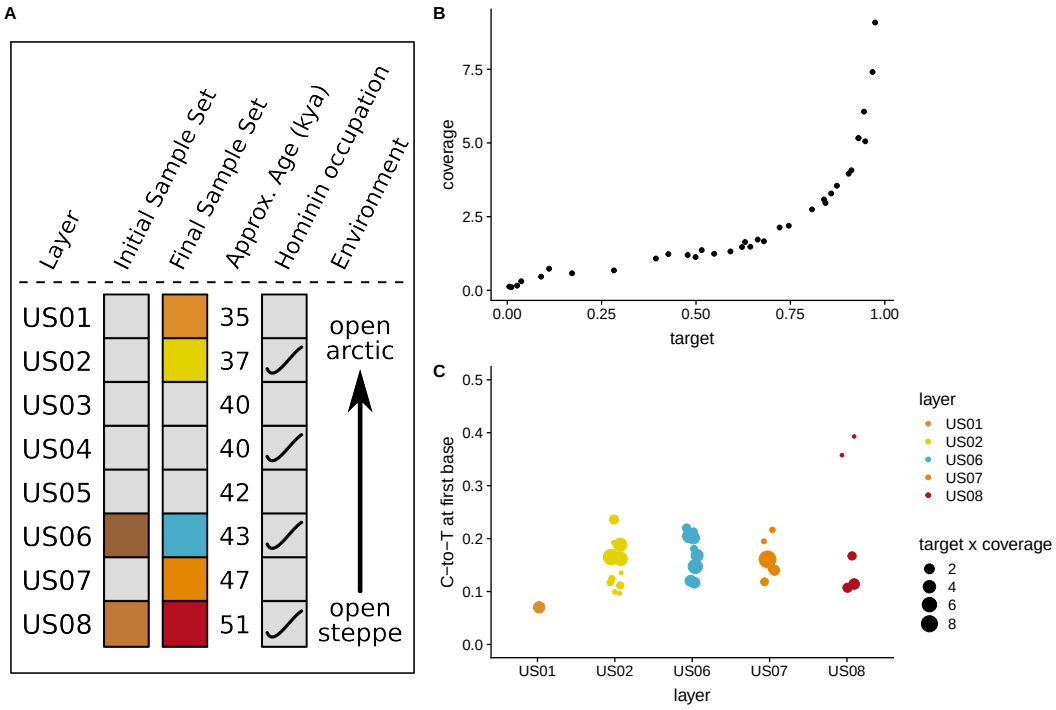

Figure 6.3: Additional capture experiments from Les Cottés sediments. A. Metainformation on the stratigraphy of the Les Cottés site, as well as the datasets used in this study. **B.** The fraction of targeted SNPs covered by at least one base, plotted against the average coverage of those SNPs for all captured libraries from Les Cottés sediments. **C.** The frequency of C-to-T substitutions at first base in each library, separated by sedimentary layer. The blob sizes represent the product of the fraction of targeted SNPs covered and their average coverage (see also B).

the highest fraction of shared derived alleles with the western *subsp. sylvestris* clade (Figure 6.4A). This signal was stable also when restricting the sampling of genotypes only to sequences carrying C-to-T conversions on either end (Figure 6.11). Looking at 10 replicate samples from each sequenced library however, there was considerable variation in the fraction of shared derived allele with the western *subsp. sylvestris* representative, which was comparable with the variation observed in the entire western *subsp. sylvestris* sample set (Figure 6.4B). Still, even those libraries with the lowest fraction of derived alleles shared with the sample representing western *subsp. sylvestris* had more derived alleles shared along the internal Spanish-western branch of the tree, rather than the eastern-cultivar branch (Figure 6.7).

In contrast to DNA extracted from a single sample, the *Vitis* DNA extracted from sediments contained in each sequencing library most likely originates from a group of genetically distinct individuals. To test this, we analyzed the relationship between the coverage of targeted SNPs in each library, and the mean genetic distance of ten subsamples of the same library. This showed considerable genetic variation within libraries which increased with coverage, suggesting that each library from Les Cottés already represented a mixture of sequence information from multiple individuals (Figure 6.9).

Following this observation, we treated the entirety of libraries from Les Cottés as a population of genotypes. To characterize this population we assessed its nucleotide diversity as a whole, or only including individuals with at least 5 fold coverage on average, and compared it with modern *Vitis vinifera* populations (Figure 6.10). This analysis suggests, similarly to the analysis assessing the variation in shared derived alleles (Figure 6.4B), that the Les Cottés *Vitis* population contains considerable genetic diversity.

It has previously been suggested, that modern *Vitis vinifera subsp. vinifera* cultivars from western Europe are to some extend the result of cultivars from the east introgressing with wild *Vitis vinifera subsp. sylvestris* from the west[279]. As the *Vitis* population from Les Cottés represents genetic diversity that is most similar to *subsp. sylvestris* from the west, we wanted to assess the introgression dynamics of modern cultivars with western *subsp. sylvestris* as well as Les Cottés *Vitis* using D-statistics. Using either of these two populations to represent wild western European ancestry as a potential source for introgression into modern cultivars of eastern, central or western origin result in significantly positive D-statistics (Figure 6.5A) for western cultivars. However, the magnitute of D was higher when using Les Cottés Vitis, so we used an additional D-statistic to explicitly test whether any subpopulation breaks the clade formed by western *Vitis vinifera subsp. sylvestris* and Les Cottés *Vitis*. Interestingly, the only subpopulations breaking this clade due to their increased affinity to Les Cottés were western cultivars, indicating that these modern cultivars are more closely related to the *Vitis* population from Les Cottés than they are to modern western *subsp. sylvestris* (Figure 6.5B).

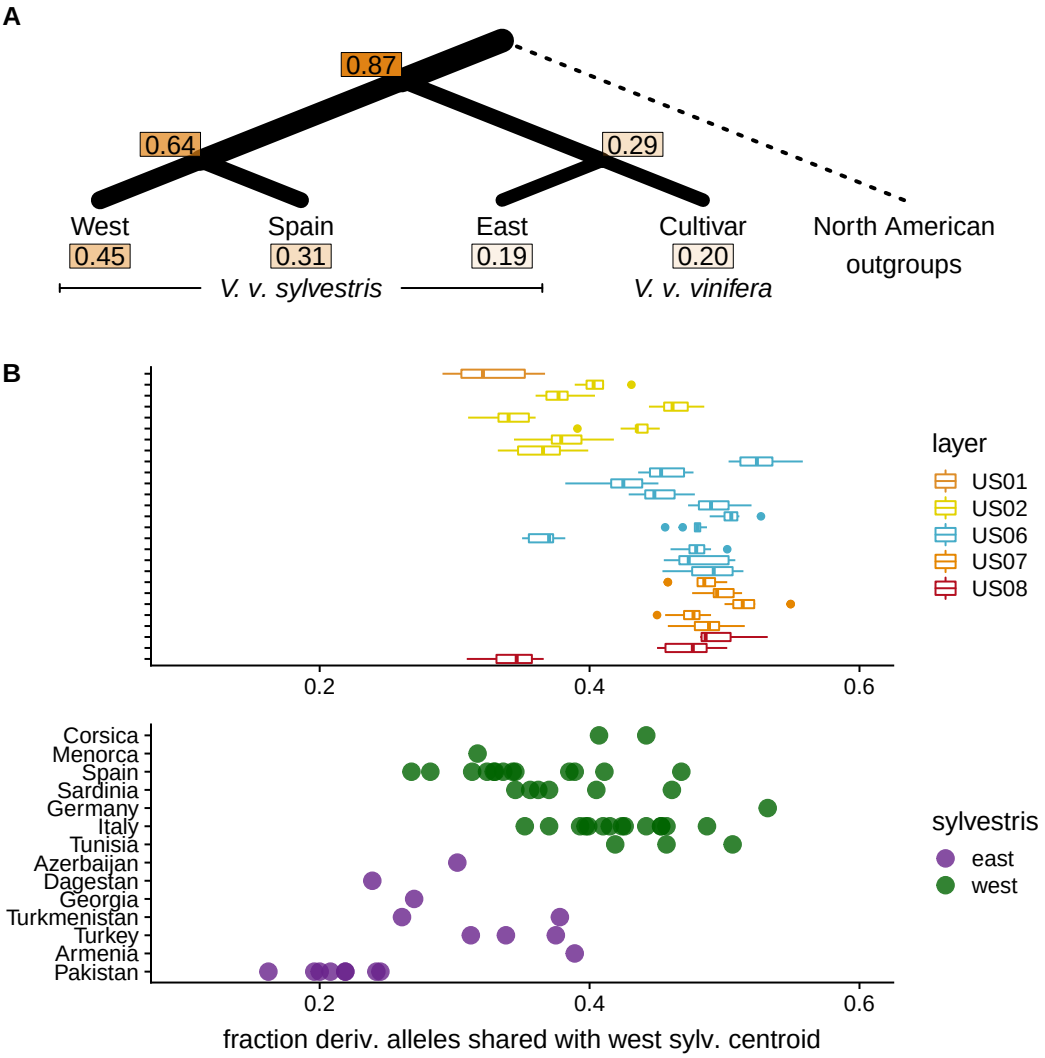

Figure 6.4: Analyses of shared derived alleles along the *Vitis vinifera* tree. For each Les Cottés library, the genotypes were sampled ten times. The fraction of shared derived alleles in these sampled genotypes with individuals representing different *Vitis vinifera* clades was assessed after the ancestral allele was ascertained in North American *Vitis* outgroups. **A.** The average fraction of shared derived alleles at different nodes of the *Vitis vinifera* tree, across all Les Cottés libraries. **B.** The fraction of shared derived alleles of each Les Cottés library with the individual representing western *Vitis vinifera subsp. sylvestris* (leftmost leaf in the tree in A). Boxplots represent ten genotype samples for each Les Cottés library. This fraction is also shown for all genotyped *Vitis vinifera subsp. sylvestris* individuals, colored by their classification as a member of the Eastern or Western clade.

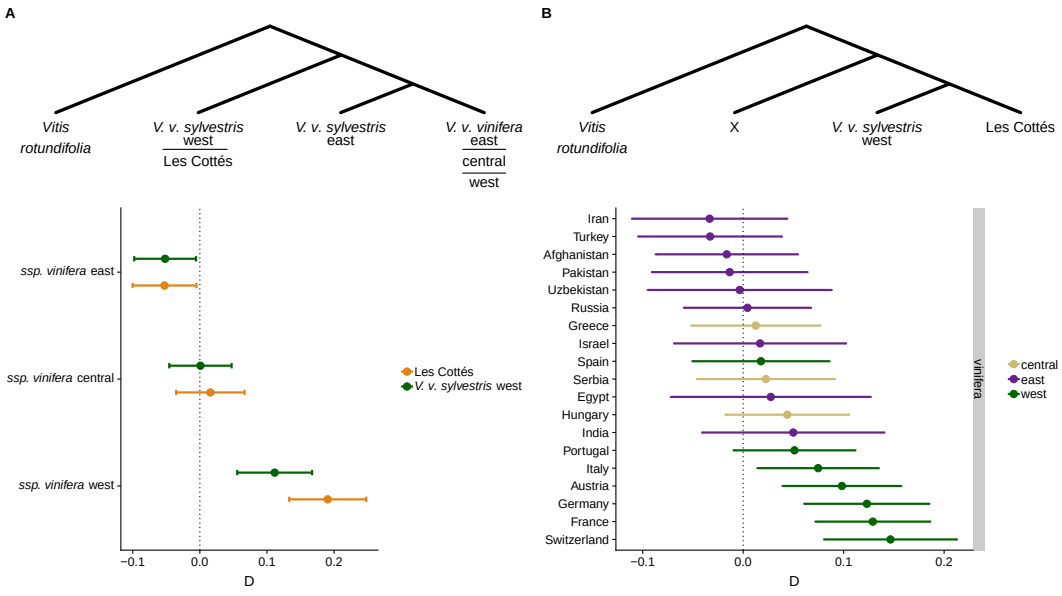

Figure 6.5: Different affinity of western *subsp. sylvestris* and Les Cottés to modern western cultivars. For all D-statistics, the point represents the estimate of D, with error bars representing two standard errors as estimated by a blocked jackknife. **A.** D-statistics of type D(outgroup, west *sylvestris*; east *sylvestris*, cultivar), using either modern western *Vitis vinifera subsp. sylvestris* or Les Cottés as the wild western population, and dividing the *Vitis vinifera subsp. vinifera* cultivars up into three geographic clusters. **B.** D-statistics of type D(outgroup, X; west *sylvestris*, Les Cottés), to test whether any *Vitis vinifera subsp. vinifera* sub-populations break the clade formed by western *Vitis vinifera subsp. sylvestris* and Les Cottés' *Vitis*. Tests are shown for sub-populations including at least four individuals.

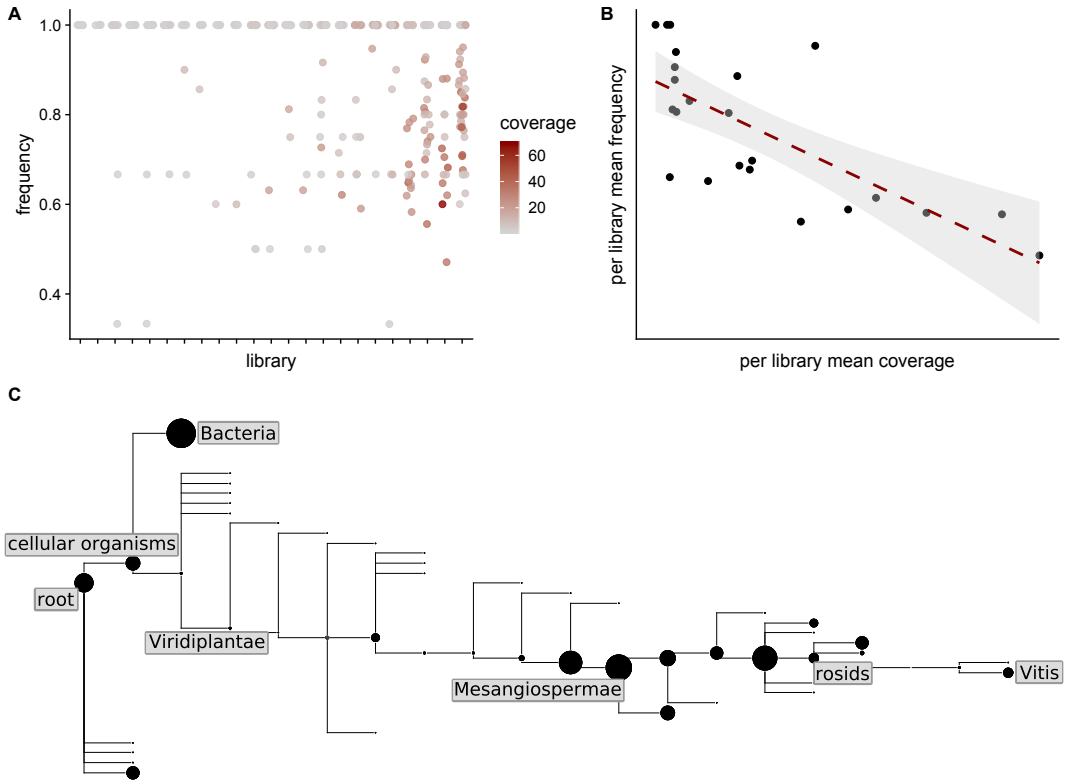

Figure 6.6: Analysis of captured chloroplast sequences. A. The major allele frequency of each variant called in the chloroplast in libraries from Les Cottés. The libraries are sorted along the x-axis by ascending mean coverage. Points are colored with a gradient indicating the coverage of a called variant. **B.** The relationship of the average coverage of called variants in a Les Cottés library and their average major allele frequency. The dashed line indicates a linear regression. **C.** Assignments of captured sequences which were aligned to the *Vitis vinifera* chloroplast and then classified along the NCBI taxonomy using MALT. The size of nodes is proportional to the number of reads assigned to them.

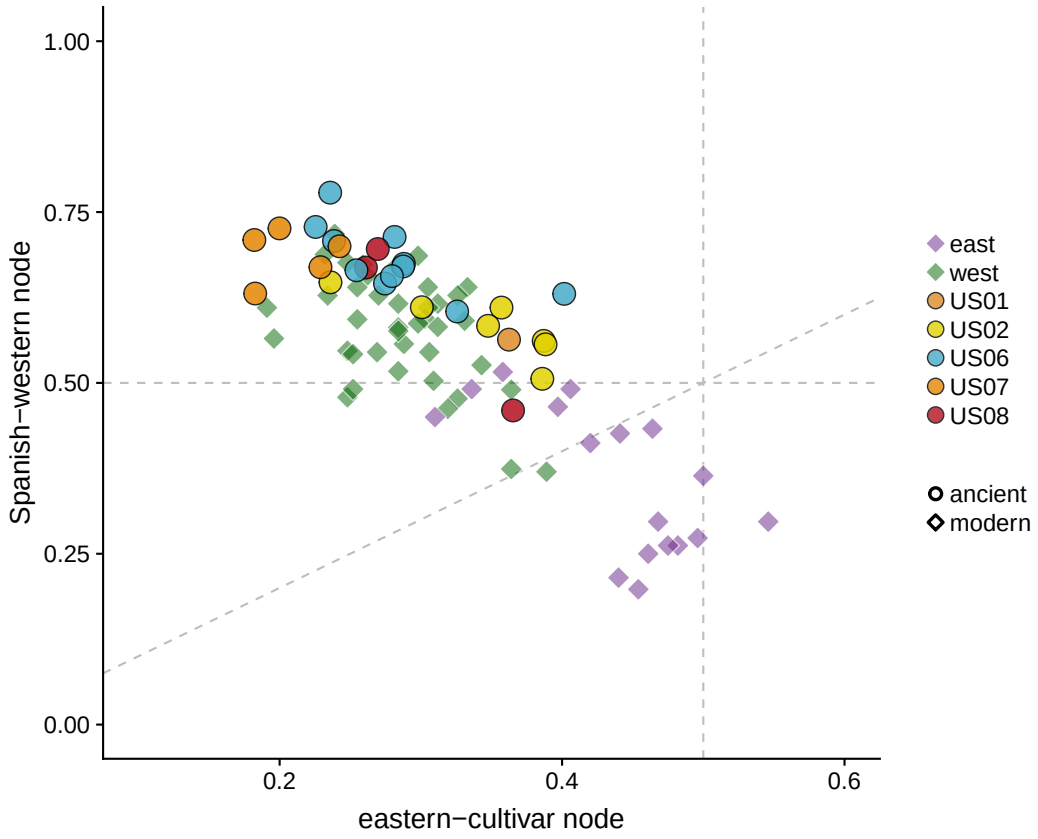

Figure 6.7: Balance of shared derived alleles. The fraction of shared derived alleles at the internal nodes of the tree shown in Figure 6.4A, for all genotypes presented in Figure 6.4B. Dotted lines indicate the diagonal, as well as the 50-50 split.

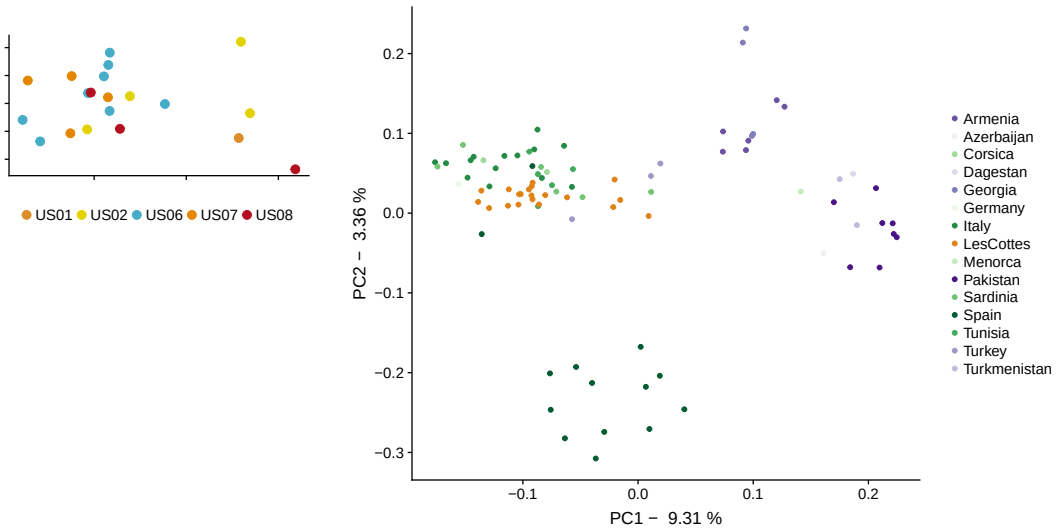

Figure 6.8: Placement of Les Cottés genotypes in the context of *Vitis vinifera subsp. sylvestris* diversity. The first two axes of a Principle Component Analysis (PCA) conducted on *Vitis vinifera subsp. sylvestris* genotypes, with sampled genotypes from Les Cottés projected into this space. The *subsp. sylvestris* individuals are colored in shades of purple or green by their classification as eastern or western, respectively. The zoom-in on the left shows the Les Cottés libraries in the same PC space, colored by their layer of origin.

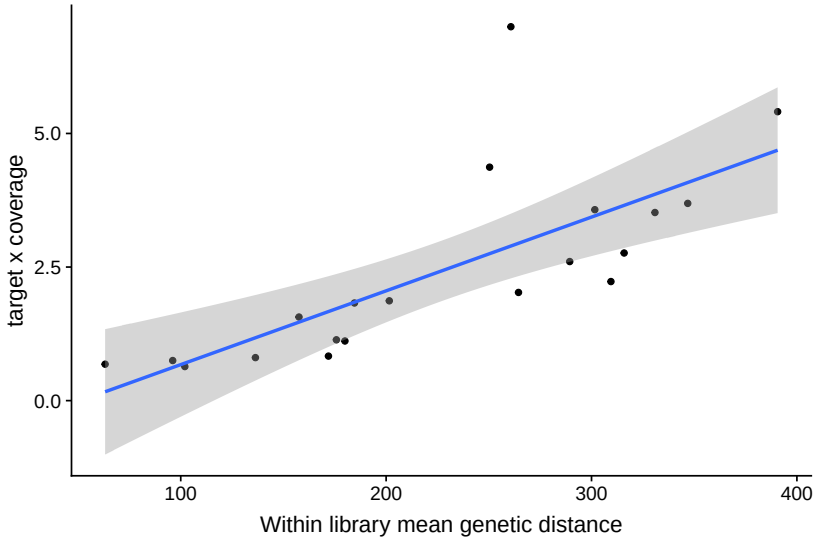

Figure 6.9: Genetic diversity within Les Cottés libraries. The mean genetic distance of ten subsamples of the same Les Cottés library is plotted against the product of the fraction of targeted SNPs covered and their average coverage.

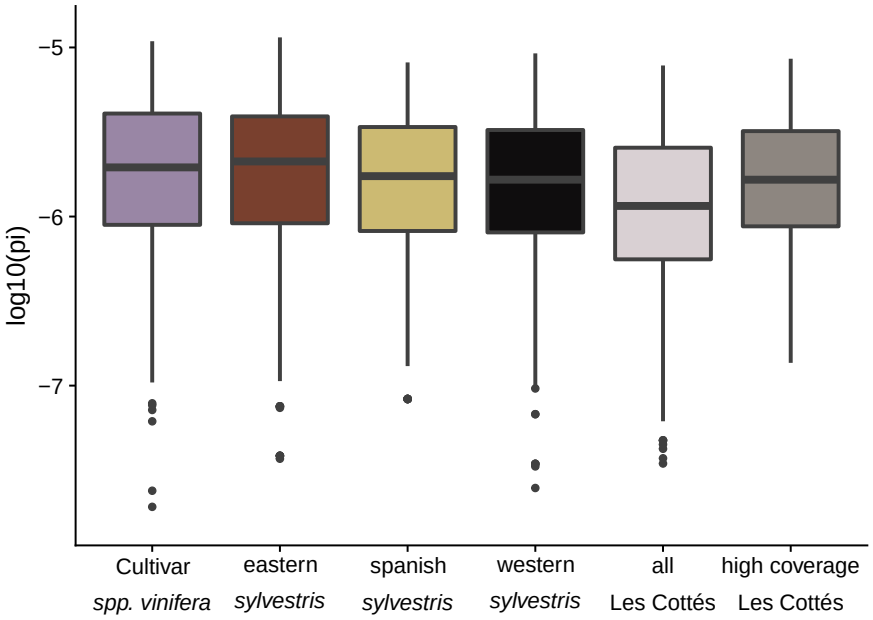

Figure 6.10: Nucleotide diversity of different *Vitis vinifera* populations. The populations analyzed are the cultivar *subsp. vinifera*, the eastern, Spanish and western sub-populations of the wild *subsp. sylvestris*, and the Les Cottés population including either all individuals or only those with at least five bases covering a SNP on average.

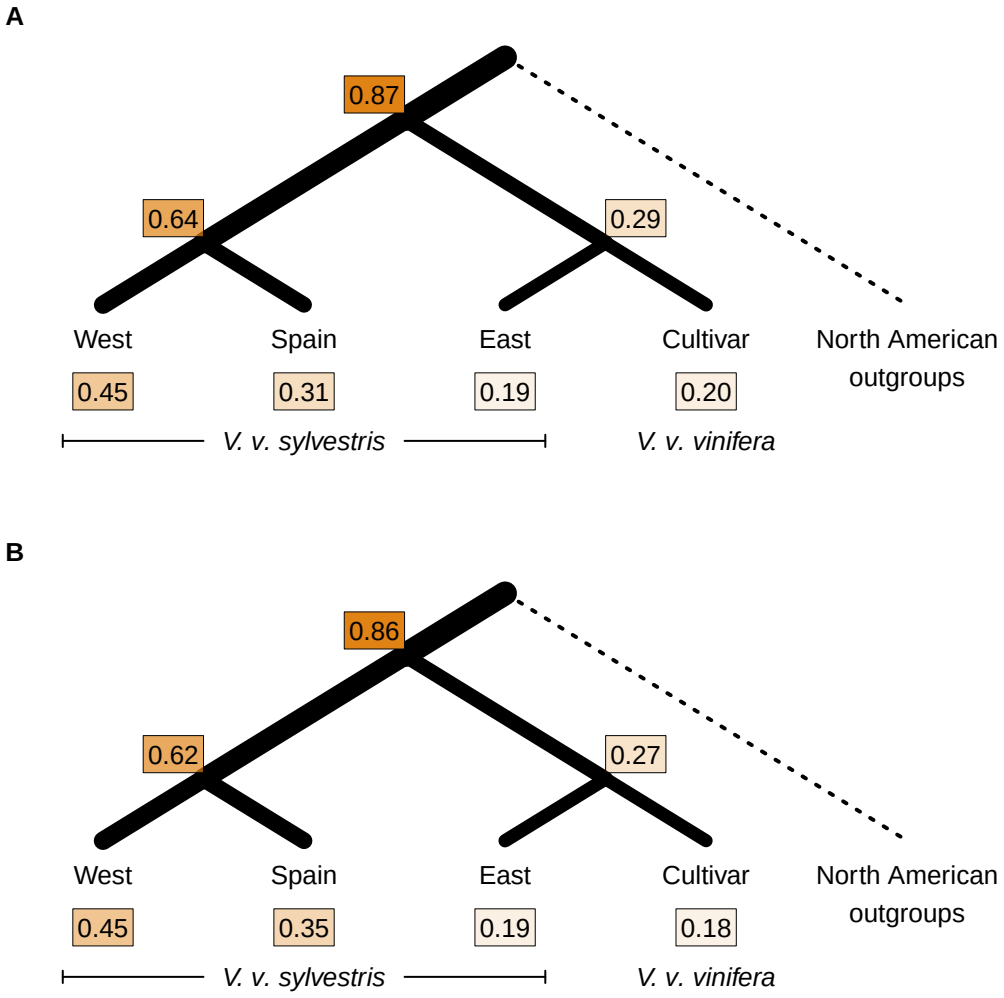

Figure 6.11: Effect of C-to-T restriction on the analysis of shared derived alleles. The fraction of shared derived alleles along the *Vitis vinifera* tree was analyzed as in Figure 6.4A, and displayed here either without (**A**) or with (**B**) applying a filter which restricts the sampling of genotypes to sequences which carry terminal C-to-T substitutions.

6.5 Discussion

In this chapter, we have presented the successful retrieval of genomic sequence information from archaeological sediments by targeted hybridization capture. While shotgun data from such sources can be used for broad-scale compositional analysis, the genetic information content for the species which constitute this composition is often low[145]. This is primarily due to the low relative abundance of sequences derived from most organisms, but may also be due to the complexity of the mixture, where the presence of confounders can make reliable sequence alignments difficult.

Because of this, a compositional analysis using shotgun sequencing can be complemented by targeted enrichment of sequences of interest[198]. Most commonly, these targeted approaches are informed by data from modern reference panels. While this somewhat restricts the species which can be investigated in this way, it also means, that data is available to which the captured samples can be compared. This context is important to allow sensible interpretation of the retrieved sequences, as their information content is limited in isolation[146,147].

We targeted sites in the nuclear genome of *Vitis vinifera*, which had been shown to be polymorphic within the genus and species, and were characterized on the Vitis9KSNP genotyping array[279,287]. The recovery of sequences overlapping these positions from sedimentary DNA extracts allowed us to characterize these sequences genetically, but also to analyze them in the context of modern diversity. These analyses showed that captured sequences were most closely related to wild *Vitis vinifera subsp. sylvestris* from western Europe. Along the stratigraphy of the Les Cottés site, we found *Vitis* DNA in all layers, with comparable amounts of age-associated damage patterns (Figure 6.3). Thus, neither the presence of *Vitis* DNA, nor its state of degradation is stratified by the age of the sediment from which the DNA was extracted.

On average, the sequences generated from Les Cottés share most derived alleles with wild *subsp. sylvestris* from western Europe, which was consistent also when restricting the sampling of genotypes to sequences carrying C-to-T conversions (Figure 6.11). This suggests, that the signal is driven by authentic ancient DNA sequences rather than a mixture of ancient *Vitis* and potential modern contaminants. When investigating the fraction of derived alleles that each library shares with western European *subsp. sylvestris*, we observed high levels of variation within and across layers (Figure 6.4B), but genotypes from all layers fall within the western European side of the gradient between eastern and western wild species (Figures 6.4B, 6.7). Surprisingly though, the extend of variation among genotypes from Les Cottes alone was comparable with the variation across all other western European *subsp. sylvestris* samples. Still, there is no clear separation of distinct groups of ancestry among libraries from Les Cottés, and we observe high diversity already within libraries, which suggests that they are complex mixtures of DNA from different individuals (Figure 6.9). Because of this, we treated the collection of libraries from Les Cottés as a population of genotypes which is most closely related to western European *Vitis vinifera subsp. sylvestris*, and shows comparable levels of diversity (Figure 6.10).

Having found and characterized this alternative population representing western Euro-

pean wild ancestry, we were interested to disentangle how it relates to the domestication history of cultivated grape. The domestication process is thought to have started in the Caucasus approximately 8,000 years ago from progenitors of eastern wild *Vitis vinifera subsp. sylvestris*[280]. From there, early cultivars moved westwards through the Mediterranean[278,284]. During this movement, it has been proposed that wild material which was encountered locally introgressed and contributed to these early cultivars[279]. The analysis of drift patterns using either modern western wild material or the population identified at Les Cottés as potential sources of gene flow into western cultivars supports this hypothesis (Figure 6.5A).

This raised the question whether modern wild material from western Europe and the population from Les Cottés were distinct in their relatedness to western cultivars. As previous analyses suggested that these two populations form a wild western European clade, we used D-statistics to test whether any other population breaks this clade. Indeed, western cultivar populations are the only group which break this clade and show a higher affinity to the Les Cottes population than to modern wild material from western Europe (Figure 6.5B).

These analyses of gene flow with western cultivars imply that both modern western *subsp. sylvestris* as well as the Les Cottés population can represent some aspect of wild western ancestry in these cultivars. Surprisingly though, the sequences recovered from the sediments at Les Cottés resemble this source of ancestry more closely than the set of characterized modern wild individuals.

To accommodate the limited stratification of the presence of *Vitis* DNA across layers and the comparable extend of age-associated degradation, as well as the introgression signal we observe, it must be considered that the *Vitis* DNA found at Les Cottés could be younger than the sediments from which the DNA was extracted. This could be caused for example by root material extending through the archaeologically relevant layers from plants on top of the stratigraphy, and explain why *Vitis* DNA is present in all layers without association e.g. to the climatic conditions inferred for different strata. It is difficult to date this population based on DNA evidence alone, and we could not find DNA evidence of climate indicator species for which pollen are present in the sediment. However, the DNA is likely of historical origin, based on the substantial signal of age-associated degradation found in all libraries. Assuming these sequences are of historical origin, but more recent than the uppermost layer of the Les Cottés stratigraphy, could place the Les Cottés population in closer temporal proximity to early *Vitis vinifera* cultivars reaching the western Mediterranean. This could yield the introgression signal we observe, which suggests that Les Cottés *Vitis* is more closely related to the source of gene flow into western cultivars than modern wild material is.

In summary, this work has shown both the difficulty and dangers, as well as the power and opportunity of using sediments as a source for ancient DNA to further our understanding of evolutionary processes. On the one hand, we show that it was necessary to pair dense sampling of the stratigraphy with targeted hybridization capture to be able to characterize the genetic diversity of *Vitis* at Les Cottés, but also to allow its inter-

pretation in the context of the archaeology of this site. On the other hand, once such a dataset in generated, it allows meaningful interpretation of the recovered genetic data in the context of modern *Vitis vinifera* diversity and its relation to domestication processes. For plant species such as grapes this is especially helpful, as efforts to characterize the domestication history genetically using archeobotanical remains have had limited success[298]. For such cases, sedimentary DNA can provide a great resource. However, it is important to interpret these data in the archeological context, as well as in the context of modern diversity, due to the difficulties of dating and authenticating such sequences. A meaningful interpretation is only possible in the presence of well established modern reference data of the diversity of an organism and its relatives. For *Vitis vinifera* such resources are continuously improving, as high-throughput sequencing prices drop and whole genome sequencing efforts become more feasible[299,300].

7 Conclusions and Outlook

The technological advances that have allowed us to sequence entire genomes with reasonable effort have revolutionized many aspects of biology[301]. As more genome sequences from a larger variety of organisms are produced, our ability to infer evolutionary processes from molecular evolution to population genetics increases[82]. In addition to better understanding such processes, this has implications for human health[302], conservation biology[303] and breeding of domesticated plants and animals[74,304,305]. A lot can be learned already from sequencing genomes of extant organisms, for example by inter-species comparative genomics[306] or intra-species population- and quantitative genetics[255,307].

However, the ability to sequence the genomes of ancient and historical organisms long after their death can greatly empower such analyses[308]. This is especially important when studying genetic diversity that has been lost through time, and might be difficult or impossible to infer from extant sequencing data alone. Such ancient DNA can be used to study migration events, which can be difficult to disentangle from extant data either because of complex patterns of gene flow, or due to large-scale replacement events through time[109]. In addition, it allows to study the history and consequences of archaic gene flow, such as for example introgression from Neanderthals into modern human genomes[117,121]. More broadly, ancient DNA allows the direct sequencing of extinct organisms, the genomes of which are by definition lost to time[8]. In plants and animals, ancient DNA can be used to study domestication processes and ecological dynamics[67,136], which greatly benefit from additional temporal resolution. Such temporally resolved datasets also allow tracking allele frequencies through time, which can aid our understanding of many evolutionary processes such as selective events[131].

Nonetheless, the time that has passed from an organisms death to the extraction of DNA molecules does leave its mark[149]. The molecular machines which constantly repair DNA in cells of a living organisms stop working once the tissues and cells die, and DNA molecules are left exposed to enzymatic attacks and oxidative processes. This leads to fragmentation of the DNA backbone, as well as other types of chemical damage such as cross-links to other macromolecules or damage to the bases themselves[3]. It also means that DNA from such specimens are usually at low concentrations, and easily overwhelmed by the DNA of secondary colonizers of the tissue or other sources of contamination[7]. These obstacles need to be overcome by ancient DNA researchers to utilize the valuable genomic information contained in DNA molecules extracted from ancient and historical specimens. To develop and apply specialized procedures to aid working with ancient DNA, it is therefore also necessary to first investigate and describe the characteristics of ancient DNA extracted from a variety of sources.

The pioneering work of ancient DNA from mammalian bones had shown which types of degradation processes can be studied directly from high-throughput sequencing data[38].

While the presence of these patterns has been reported almost uniformly across ancient DNA sequencing projects, their relationship with the age of samples remained inconclusive[152]. This is most likely due to environmental differences, as the chemical properties of the environment in which a specimen was preserved and stored can strongly affect DNA degradation[160]. While no correlation of age and fragmentation had been found in sequencing data from bones[152], qPCR of a time series dataset had allowed the assessment of fragmentation kinetics and the calculation of a DNA decay rate in this type of tissue[162].

We have used high-throughput sequencing of ancient DNA extracted from herbarium samples to study the kinetics and temporal dynamics of age-associated degradation patterns. Herbaria are vast collections of dried plant specimens, which hold great promise as a source of ancient DNA to study effects of global change, modern plant cultivation, and the rediscovery of the New World[69,309]. As such, it is important to understand the extend and dynamics of damage in DNA extracted from these samples. In addition, the environmental influence on DNA preservation may be more limited compared to mammalian fossils, which would help in the discovery of temporal signals of ancient DNA decay. We show that DNA fragmentation, as well as the deamination of cytosines to uracils are strongly associated with the age of the sample. In addition, we were able to use the temporal signal of DNA fragmentation to calculate a decay rate of DNA molecules in herbarium samples, which supported the observation that DNA from such tissue is highly fragmented. These insights can be used to inform the development of methods which help to utilize herbaria as an ancient DNA resource, for instance by optimizing the retrieval of ultra-short molecules[165].

In addition to these patterns of age-associated degradation of DNA molecules, their fragmentation also renders these samples highly prone to contamination. The low concentration of endogenous DNA, as well as their short size, means that exogenous DNA can quickly overwhelm the DNA of the organism of interest[7]. This has been a major challenge to overcome throughout the history of ancient DNA research, and several early discoveries have since been attributed to contamination[17,18]. However, this problem persists even in the age of high-throughput sequencing of ancient DNA[183]. A great advantage of library-based sequencing is the direct observation of damage patterns, which can be used to distinguish authentic ancient DNA from modern contamination[38,39]. The most characteristic and ubiquitous of such patterns are C-to-T conversions caused by deamination of cytosines to uracils. These patterns have been found even in samples from permafrost[48] and lake sediments[145]. In addition, they showed the strongest association with sample age in our investigation of herbarium samples, as well as in mammalian bones[152].

However, the taxonomic complexity of ancient DNA extracts can sometimes hinder the straight-forward detection of these patterns[199]. Still, it is imperative that positive evidence of authenticity is provided. Therefore, we have investigated and developed methods which aid the authentication process in these difficult cases. First, we have investigated a library preparation protocol which separates molecules which carry uracils from those that do not[208]. This allows sequencing a DNA extract in two fractions: one enriched, and one depleted in molecules carrying these degradation signatures caused

by the deamination of cytosines. Sequencing both of these fractions can be coupled with taxonomic assignment of sequences, and the relative abundance of taxa in the two fractions of a library can be used to identify taxa of ancient or historical origin. In addition, the uracil-enriched fraction can be used to magnify degradation patterns in cases where they are difficult to observe, for example due to high divergence between a sample and the closest available reference genome. We expect that this approach will be helpful in the analysis of ancient metagenomes, especially those of high taxonomic complexity.

We also have developed methods to aid the authentication and analysis of ancient DNA sequencing data generated using standard ancient DNA library preparation procedures. The first such method is primarily aimed at studies of ancient DNA from taxonomically complex mixtures. Here, sequences of interest are often at low abundance, which can make the detection of characteristic deamination patterns difficult. Our method captures the presence of this pattern and summarizes it in a single value, which describes the goodness of fit of an exponential decay curve. Combining this approach with random sampling of sequences of known provenance allows the generation of empirical distributions, which a query distribution can be tested against. In addition, it allows to assess the sequencing depth required to reliably identify deamination patterns needed for authentication. Our results show, that only a few hundred sequences are required for authentication of even weakly deaminated libraries.

Another way of overcoming low proportions of endogenous DNA in complex ancient DNA mixtures is by using targeted hybridization capture[44,45]. For this, probe sequences are designed to be complementary to sequences of interest. Immobilizing these probes either on a glass slide (array capture) or on beads using the interaction of biotin and streptavidin (in-solution capture), allows to enrich library molecules which hybridize with these probes. This procedure is especially useful if sequences or positions of interest are already assessed for example in modern populations. Targeted capture then allows to retrieve sequence data at such positions of interest, to relate the genomic information from these specimens with modern diversity. We have developed a tool to aid this process, primarily for low coverage sequencing data as generated from ancient DNA extracts. It uses coordinates of sites defined by a reference genome, together with sequence alignments to this reference, to sample bases from aligned sequences at targeted sites defined by the coordinates. This allows to quickly place new genotypes in the context of already characterized diversity, without the need for model-based variant calling. We also developed a similar tool for use cases where no pre-defined sites of variation exist. In this case, variable sites are discovered de novo from a set of samples aligned to a common reference. Together, these tools are useful for quick assessments of genetic diversity and relatedness between samples from low coverage sequencing data, and can be used both in the presence or absence of well characterized reference samples. Since such reference data is available for few species only, the latter use case makes this method accessible also to non-model organisms or organisms of limited commercial value.

To broadly characterize genetic diversity and relatedness, it is often enough to generate low-coverage sequencing data and assess genotypes through procedures such as the

sampling scheme we have presented. Some genomic features however do require higher coverage data, and the ability to reliably assess allele frequencies within samples. If an organism shows intra-specific variation in ploidy for example, it is of interest to characterize this additional axis of genomic variation in each sample. Ideally this would be done from sequencing data directly, which then allows to study both genomic diversity and ploidy from one sequencing experiment. This is possible for example when trying to distinguish di-, tri- and tetraploids, as the allele frequencies at heterozygous sites are markedly different. We developed a method to infer these ploidies directly from sequence alignments, without the need for variant calling. This again allows the rapid assessment of this type of diversity, and allows the incorporation of both modern and ancient sequencing data. In addition, it can be used to detect aneuploidies by estimating ploidy level locally across the genome.

We have applied what we have learned about ancient DNA from diverse sources, and some of the methods we have developed, to characterize *Vitis vinifera* sequences we found in DNA extracts from archaeological sediments. We have made use of targeted capture and the sampling of bases at variable sites, to place these sequences in the context of modern *Vitis* diversity. This has allowed us to relate these sedimentary sequences with the domestication history of modern, cultivated grape vine. To the best of our knowledge, this constitutes the first successful assessment of relatedness from genomic ancient DNA retrieved from sedimentary DNA extracts.

In summary, the work presented here contributes to many aspects of investigating and using ancient DNA as a source of genetic sequence information to be used for evolutionary inference. This starts at characterizing DNA molecules and the degradation processes which have shaped them over time. In addition, because of the metagenomic nature of ancient DNA and the risk of contamination, a crucial step of ancient DNA sequence analysis is the taxonomic characterization of sequences as well as the authentication of their ancient origin. To this we have contributed by developing frameworks and describing methods which can aid this process in cases where it is not straight-forward. We have also developed tools to analyze genetic variation data at low coverage and low quality, as it may be retrieved from complex ancient DNA extracts. These insights and tools were then applied to analyze nuclear sequence data from a wild relative of cultivated grape vine directly from the sediments of an archaeological cave site.

Our work opens up the ancient DNA framework to be applied to a larger variety of samples from a larger variety of organisms. This especially includes ancient plant DNA from herbaria, as well as from archaeological sediments. However, we hope to have contributed also to the continuous development and implementation of good practices when working with DNA from these types of sources, especially concerning the authentication of ancient DNA. In addition, most of the tools we provide, while developed for ancient DNA applications, can be used for a variety of sequencing projects from modern samples, or those incorporating both modern and ancient sequencing data.

References

[1] R. Higuchi et al. “DNA sequences from the quagga, an extinct member of the horse family”. *Nature* 312.5991 (1984), pp. 282–284.

[2] S. Pääbo. “Molecular cloning of Ancient Egyptian mummy DNA”. *Nature* 314.6012 (1985), pp. 644–645.

[3] S. Pääbo. “Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification”. *Proc. Natl. Acad. Sci. U. S. A.* 86.6 (Mar. 1989), pp. 1939–1943.

[4] R. K. Saiki et al. “Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia”. *Science* 230.4732 (Dec. 1985), pp. 1350–1354.

[5] R. K. Saiki et al. “Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase”. *Science* 239.4839 (Jan. 1988), pp. 487–491.

[6] S. Pääbo, R. G. Higuchi, and A. C. Wilson. “Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology”. *J. Biol. Chem.* 264.17 (June 1989), pp. 9709–9712.

[7] S. Pääbo et al. “Genetic analyses from ancient DNA”. *Annu. Rev. Genet.* 38 (2004), pp. 645–679.

[8] M. Hofreiter et al. “Ancient DNA”. *Nat. Rev. Genet.* 2.5 (May 2001), pp. 353–359.

[9] A. Cooper and H. N. Poinar. “Ancient DNA: Do It Right or Not at All”. *Science* 289.5482 (Aug. 2000), pp. 1139–1139.

[10] D. Serre et al. “No evidence of Neandertal mtDNA contribution to early modern humans”. *PLoS Biol.* 2.3 (Mar. 2004), E57.

[11] H. N. Poinar et al. “Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis”. *Science* 281.5375 (July 1998), pp. 402–406.

[12] M. Krings et al. “A view of Neandertal genetic diversity”. *Nat. Genet.* 26.2 (Oct. 2000), pp. 144–146.

[13] M. Krings et al. “Neandertal DNA sequences and the origin of modern humans”. *Cell* 90.1 (July 1997), pp. 19–30.

[14] O. Handt et al. “The retrieval of ancient human DNA sequences”. *Am. J. Hum. Genet.* 59.2 (Aug. 1996), pp. 368–376.

[15] R. J. Cano et al. “Amplification and sequencing of DNA from a 120-135-million-year-old weevil”. *Nature* 363.6429 (June 1993), pp. 536–538.

[16] S. R. Woodward, N. J. Weyand, and M. Bunnell. “DNA sequence from Cretaceous period bone fragments”. *Science* 266.5188 (Nov. 1994), pp. 1229–1232.

[17] H. Zischler et al. “Detecting dinosaur DNA”. *Science* 268.5214 (May 1995), 1192–3, author reply 1194.

[18] T. Lindahl. “Recovery of antediluvian DNA”. *Nature* 365.6448 (Oct. 1993), p. 700.

[19] J.-J. Hublin and S. Pääbo. “Neandertals”. *Curr. Biol.* 16.4 (Feb. 2006), R113–4.

[20] E. S. Lander et al. “Initial sequencing and analysis of the human genome”. *Nature* 409.6822 (Feb. 2001), pp. 860–921.

[21] R. Staden. “A strategy of DNA sequencing employing computer programs”. *Nucleic Acids Res.* 6.7 (June 1979), pp. 2601–2610.

[22] S. Anderson. “Shotgun DNA sequencing using cloned DNase I-generated fragments”. *Nucleic Acids Res.* 9.13 (July 1981), pp. 3015–3027.

[23] P. L. Deininger. “Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis”. *Anal. Biochem.* 129.1 (Feb. 1983), pp. 216–223.

[24] J. P. Noonan et al. “Genomic sequencing of Pleistocene cave bears”. *Science* 309.5734 (July 2005), pp. 597–599.

[25] M. Margulies et al. “Genome sequencing in microfabricated high-density picolitre reactors”. *Nature* 437.7057 (Sept. 2005), pp. 376–380.

[26] H. N. Poinar et al. “Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA”. *Science* 311.5759 (Jan. 2006), pp. 392–394.

[27] R. E. Green et al. “Analysis of one million base pairs of Neanderthal DNA”. *Nature* 444.7117 (Nov. 2006), pp. 330–336.

[28] J. Shendure and H. Ji. “Next-generation DNA sequencing”. *Nat. Biotechnol.* 26.10 (Oct. 2008), pp. 1135–1145.

[29] J. P. Noonan et al. “Sequencing and analysis of Neanderthal genomic DNA”. *Science* 314.5802 (Nov. 2006), pp. 1113–1118.

[30] R. E. Green et al. “A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing”. *Cell* 134.3 (Aug. 2008), pp. 416–426.

[31] C. Adessi et al. “Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms”. *Nucleic Acids Res.* 28.20 (Oct. 2000), E87.

[32] M. Fedurco et al. “BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies”. *Nucleic Acids Res.* 34.3 (Feb. 2006), e22.

[33] G. Turcatti et al. “A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis”. *Nucleic Acids Res.* 36.4 (Mar. 2008), e25.

[34] D. R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. *Nature* 456.7218 (Nov. 2008), pp. 53–59.

[35] M. Kircher and J. Kelso. “High-throughput DNA sequencing–concepts and limitations”. *Bioessays* 32.6 (June 2010), pp. 524–536.

[36] H. P. J. Buermans and J. T. den Dunnen. “Next generation sequencing technology: Advances and applications”. *Biochim. Biophys. Acta* 1842.10 (Oct. 2014), pp. 1932–1941.

[37] N. J. Loman et al. “Performance comparison of benchtop high-throughput sequencing platforms”. *Nat. Biotechnol.* 30.5 (May 2012), pp. 434–439.

[38] A. W. Briggs et al. “Patterns of damage in genomic DNA sequences from a Neandertal”. *Proceedings of the National Academy of Sciences* 104.37 (Sept. 2007), pp. 14616–14621.

[39] J. Krause et al. “A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia”. *Curr. Biol.* 20.3 (Feb. 2010), pp. 231–236.

[40] M. Hofreiter et al. “DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA”. *Nucleic Acids Res.* 29.23 (Dec. 2001), pp. 4793–4799.

[41] T. Lindahl et al. “DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli”. *J. Biol. Chem.* 252.10 (May 1977), pp. 3286–3294.

[42] A. W. Briggs et al. “Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA”. *Nucleic Acids Res.* 38.6 (Apr. 2010), e87.

[43] A. W. Briggs et al. “Targeted retrieval and analysis of five Neandertal mtDNA genomes”. *Science* 325.5938 (July 2009), pp. 318–321.

[44] H. A. Burbano et al. “Targeted investigation of the Neandertal genome by array-based sequence capture”. *Science* 328.5979 (May 2010), pp. 723–725.

[45] T. Maricic, M. Whitten, and S. Pääbo. “Multiplexed DNA sequence capture of mitochondrial genomes using PCR products”. *PLoS One* 5.11 (Nov. 2010), e14004.

[46] R. E. Green et al. “A draft sequence of the Neandertal genome”. *Science* 328.5979 (May 2010), pp. 710–722.

[47] D. Reich et al. “Genetic history of an archaic hominin group from Denisova Cave in Siberia”. *Nature* 468.7327 (Dec. 2010), pp. 1053–1060.

[48] L. Orlando et al. “Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse”. *Nature* 499.7456 (July 2013), pp. 74–78.

[49] M. Meyer et al. “A mitochondrial genome sequence of a hominin from Sima de los Huesos”. *Nature* 505.7483 (Jan. 2014), pp. 403–406.

[50] M. Meyer et al. “Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins”. *Nature* 531.7595 (Mar. 2016), pp. 504–507.

[51] J. Dabney et al. “Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments”. *Proc. Natl. Acad. Sci. U. S. A.* 110.39 (Sept. 2013), pp. 15758–15763.

[52] M.-T. Gansauge and M. Meyer. “Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA”. *Nat. Protoc.* 8.4 (Apr. 2013), pp. 737–748.

[53] M. Meyer and M. Kircher. “Illumina sequencing library preparation for highly multiplexed target capture and sequencing”. *Cold Spring Harb. Protoc.* 2010.6 (June 2010), db.prot5448.

[54] M. Meyer et al. “A high-coverage genome sequence from an archaic Denisovan individual”. *Science* 338.6104 (Oct. 2012), pp. 222–226.

[55] M.-T. Gansauge et al. “Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase”. *Nucleic Acids Res.* (Jan. 2017).

[56] A. Schlumbaum, J.-M. Neuhaus, and S. Jacomet. “Coexistence of Tetraploid and Hexaploid Naked Wheat in a Neolithic Lake Dwelling of Central Europe: Evidence from Morphology and Ancient DNA”. *J. Archaeol. Sci.* 25.11 (Nov. 1998), pp. 1111–1118.

[57] R. G. Allaby, M. Banerjee, and T. A. Brown. “Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat”. *Genome* 42.2 (Apr. 1999), pp. 296–307.

[58] V. Jaenicke-Després et al. “Early allelic selection in maize as revealed by ancient DNA”. *Science* 302.5648 (Nov. 2003), pp. 1206–1208.

[59] E. Willerslev et al. “Diverse plant and animal genetic records from Holocene and Pleistocene sediments”. *Science* 300.5620 (May 2003), pp. 791–795.

[60] S. A. Palmer et al. “Archaeogenomic evidence of punctuated genome evolution in Gossypium”. *Mol. Biol. Evol.* 29.8 (Aug. 2012), pp. 2031–2038.

[61] L. Kistler et al. “Transoceanic drift and the domestication of African bottle gourds in the Americas”. *Proc. Natl. Acad. Sci. U. S. A.* 111.8 (Feb. 2014), pp. 2937–2941.

[62] R. R. da Fonseca et al. “The origin and evolution of maize in the Southwestern United States”. *Nat Plants* 1 (Jan. 2015), p. 14003.

[63] M. Mascher et al. “Genomic analysis of 6,000-year-old cultivated grain illuminates the domestication history of barley”. *Nat. Genet.* 48.9 (Sept. 2016), pp. 1089–1093.

[64] A. Schlumbaum, M. Tensen, and V. Jaenicke-Després. “Ancient plant DNA in archaeobotany”. *Veg. Hist. Archaeobot.* 17.2 (Mar. 2008), pp. 233–244.

[65] J. Threadgold and T. A. Brown. "Degradation of DNA in artificially charred wheat seeds". *J. Archaeol. Sci.* 30.8 (Aug. 2003), pp. 1067–1076.

[66] F. Gugerli, L. Parducci, and R. J. Petit. "Ancient plant DNA: review and prospects: Research review". *New Phytol.* 166.2 (Feb. 2005), pp. 409–418.

[67] K. Swarts et al. "Genomic estimation of complex traits reveals ancient maize adaptation to temperate North America". *Science* 357.6350 (Aug. 2017), pp. 512–515.

[68] R. M. Gutaker and H. A. Burbano. "Reinforcing plant evolutionary genomics using ancient DNA". *Curr. Opin. Plant Biol.* 36 (Feb. 2017), pp. 38–45.

[69] P. L. M. Lang et al. "Using herbaria to study global environmental change". *New Phytol.* 221.1 (Jan. 2019), pp. 110–122.

[70] P. S. Soltis. "Digitization of herbaria enables novel research". *Am. J. Bot.* 104.9 (Sept. 2017), pp. 1281–1284.

[71] M. Staats et al. "DNA damage in plant herbarium tissue". *PLoS One* 6.12 (Dec. 2011), e28448.

[72] K. Yoshida et al. "The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine". *Elife* 2 (May 2013), e00731.

[73] International HapMap Consortium. "The International HapMap Project". *Nature* 426.6968 (Dec. 2003), pp. 789–796.

[74] B. C. Y. Collard and D. J. Mackill. "Marker-assisted selection: an approach for precision plant breeding in the twenty-first century". *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363.1491 (Feb. 2008), pp. 557–572.

[75] International HapMap Consortium. "A haplotype map of the human genome". *Nature* 437.7063 (Oct. 2005), pp. 1299–1320.

[76] T. A. Manolio et al. "Finding the missing heritability of complex diseases". *Nature* 461.7265 (Oct. 2009), pp. 747–753.

[77] R. Nielsen. "Genomics: In search of rare human variants". *Nature* 467.7319 (Oct. 2010), pp. 1050–1051.

[78] N. Siva. "1000 Genomes project". *Nat. Biotechnol.* 26.3 (Mar. 2008), p. 256.

[79] 1000 Genomes Project Consortium et al. "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491.7422 (Nov. 2012), pp. 56–65.

[80] D. Weigel and R. Mott. "The 1001 genomes project for Arabidopsis thaliana". *Genome Biol.* 10.5 (May 2009), p. 107.

[81] 1001 Genomes Consortium. "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". *Cell* 166.2 (July 2016), pp. 481–491.

[82] H. Ellegren. "Genome sequencing and population genomics in non-model organisms". *Trends Ecol. Evol.* 29.1 (Jan. 2014), pp. 51–63.

[83] M. L. Metzker. “Sequencing technologies - the next generation”. *Nat. Rev. Genet.* 11.1 (Jan. 2010), pp. 31–46.

[84] P. Flicek and E. Birney. “Sense from sequence reads: methods for alignment and assembly”. *Nat. Methods* 6.11 Suppl (Nov. 2009), S6–S12.

[85] A. Dilthey et al. “Improved genome inference in the MHC using a population reference graph”. *Nat. Genet.* 47.6 (June 2015), pp. 682–688.

[86] H. Li, J. Ruan, and R. Durbin. “Mapping short DNA sequencing reads and calling variants using mapping quality scores”. *Genome Res.* 18.11 (Nov. 2008), pp. 1851–1858.

[87] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows–Wheeler transform”. *Bioinformatics* (2009).

[88] B. Langmead et al. “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome”. *Genome Biol.* 10.3 (Mar. 2009), R25.

[89] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[90] 1000 Genomes Project Consortium et al. “A map of human genome variation from population-scale sequencing”. *Nature* 467.7319 (Oct. 2010), pp. 1061–1073.

[91] O. Harismendy et al. “Evaluation of next generation sequencing platforms for population targeted sequencing studies”. *Genome Biol.* 10.3 (Mar. 2009), R32.

[92] A. McKenna et al. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. *Genome Res.* 20.9 (Sept. 2010), pp. 1297–1303.

[93] R. Nielsen et al. “Genotype and SNP calling from next-generation sequencing data”. *Nat. Rev. Genet.* 12.6 (June 2011), pp. 443–451.

[94] M. A. DePristo et al. “A framework for variation discovery and genotyping using next-generation DNA sequencing data”. *Nat. Genet.* 43.5 (May 2011), pp. 491–498.

[95] E. Garrison and G. Marth. “Haplotype-based variant detection from short-read sequencing” (July 2012).

[96] P. Danecek et al. “The variant call format and VCFtools”. *Bioinformatics* 27.15 (Aug. 2011), pp. 2156–2158.

[97] G. A. Van der Auwera et al. “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline”. *Curr. Protoc. Bioinformatics* 43 (2013), pp. 11.10.1–33.

[98] H. E. L. Lischer and L. Excoffier. “PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs”. *Bioinformatics* 28.2 (Jan. 2012), pp. 298–299.

[99] R. Khan and D. Mittelman. “Consumer genomics will change your life, whether you get tested or not”. *Genome Biol.* 19.1 (Aug. 2018), p. 120.

[100] J. Hardy and A. Singleton. “Genomewide association studies and human disease”. *N. Engl. J. Med.* 360.17 (Apr. 2009), pp. 1759–1768.

[101] L. A. Hindorff et al. “Potential etiologic and functional implications of genome-wide association loci for human diseases and traits”. *Proc. Natl. Acad. Sci. U. S. A.* 106.23 (June 2009), pp. 9362–9367.

[102] J. D. Wall. “Estimating ancestral population sizes and divergence times”. *Genetics* 163.1 (Jan. 2003), pp. 395–404.

[103] J. D. Wall. “Detecting ancient admixture in humans using sequence polymorphism data”. *Genetics* 154.3 (Mar. 2000), pp. 1271–1279.

[104] N. A. Rosenberg et al. “Genetic structure of human populations”. *Science* 298.5602 (Dec. 2002), pp. 2381–2385.

[105] J. K. Pritchard et al. “Association mapping in structured populations”. *Am. J. Hum. Genet.* 67.1 (July 2000), pp. 170–181.

[106] N. Patterson, A. L. Price, and D. Reich. “Population structure and eigenanalysis”. *PLoS Genet.* 2.12 (Dec. 2006), e190.

[107] A. L. Price et al. “New approaches to population stratification in genome-wide association studies”. *Nat. Rev. Genet.* 11.7 (July 2010), pp. 459–463.

[108] D. E. MacHugh, G. Larson, and L. Orlando. “Taming the Past: Ancient DNA and the Study of Animal Domestication”. *Annu Rev Anim Biosci* 5 (Feb. 2017), pp. 329–351.

[109] I. Lazaridis. “The evolutionary history of human populations in Europe”. *Curr. Opin. Genet. Dev.* 53 (June 2018), pp. 21–27.

[110] O. Savolainen, M. Lascoux, and J. Merilä. “Ecological genomics of local adaptation”. *Nat. Rev. Genet.* 14.11 (Nov. 2013), pp. 807–820.

[111] S. I. Wright et al. “The effects of artificial selection on the maize genome”. *Science* 308.5726 (May 2005), pp. 1310–1314.

[112] M. A. Gore et al. “A first-generation haplotype map of maize”. *Science* 326.5956 (Nov. 2009), pp. 1115–1117.

[113] H. Li and R. Durbin. “Inference of human population history from individual whole-genome sequences”. *Nature* 475.7357 (July 2011), pp. 493–496.

[114] J. D. Wall and M. Slatkin. “Paleopopulation genetics”. *Annu. Rev. Genet.* 46 (Sept. 2012), pp. 635–649.

[115] P. Skoglund and I. Mathieson. “Ancient Human Genomics: The First Decade”. *Annu. Rev. Genomics Hum. Genet.* (Apr. 2018).

[116] K. Prüfer et al. “The complete genome sequence of a Neanderthal from the Altai Mountains”. *Nature* 505.7481 (Jan. 2014), pp. 43–49.

[117] S. Sankararaman et al. “The genomic landscape of Neanderthal ancestry in present-day humans”. *Nature* 507.7492 (Mar. 2014), pp. 354–357.

[118] B. Vernot and J. M. Akey. "Resurrecting surviving Neandertal lineages from modern human genomes". *Science* 343.6174 (Feb. 2014), pp. 1017–1021.

[119] S. Sankararaman et al. "The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans". *Curr. Biol.* 26.9 (May 2016), pp. 1241–1247.

[120] F. Racimo et al. "Evidence for archaic adaptive introgression in humans". *Nat. Rev. Genet.* 16.6 (June 2015), pp. 359–371.

[121] M. Dannemann and F. Racimo. "Something old, something borrowed: admixture and adaptation in human evolution". *Curr. Opin. Genet. Dev.* 53.e26782v1 (June 2018), pp. 1–8.

[122] M. Dannemann and J. Kelso. "The Contribution of Neanderthals to Phenotypic Variation in Modern Humans". *Am. J. Hum. Genet.* 101.4 (Oct. 2017), pp. 578–589.

[123] Q. Fu et al. "An early modern human from Romania with a recent Neanderthal ancestor". *Nature* 524.7564 (Aug. 2015), pp. 216–219.

[124] V. Slon et al. "The genome of the offspring of a Neanderthal mother and a Denisovan father". *Nature* (Aug. 2018), p. 1.

[125] N. Patterson et al. "Ancient admixture in human history". *Genetics* 192.3 (Nov. 2012), pp. 1065–1093.

[126] D. Reich et al. "Reconstructing Indian population history". *Nature* 461.7263 (Sept. 2009), pp. 489–494.

[127] L. L. Cavalli-Sforza et al. *The History and Geography of Human Genes*. Princeton University Press, 1994. ISBN: 9780691087504.

[128] J. Novembre and M. Stephens. "Interpreting principal component analyses of spatial population genetic variation". *Nat. Genet.* 40.5 (May 2008), pp. 646–649.

[129] T. A. Brown et al. "The complex origins of domesticated crops in the Fertile Crescent". *Trends Ecol. Evol.* 24.2 (Feb. 2009), pp. 103–109.

[130] J. P. Bollback, T. L. York, and R. Nielsen. "Estimation of 2Nes from temporal allele frequency data". *Genetics* 179.1 (May 2008), pp. 497–502.

[131] I. Mathieson and G. McVean. "Estimating selection coefficients in spatially structured populations from time series data of allele frequencies". *Genetics* 193.3 (Mar. 2013), pp. 973–984.

[132] A. Ludwig et al. "Coat color variation at the beginning of horse domestication". *Science* 324.5926 (Apr. 2009), p. 485.

[133] T. Bersaglieri et al. "Genetic signatures of strong recent positive selection at the lactase gene". *Am. J. Hum. Genet.* 74.6 (June 2004), pp. 1111–1120.

[134] J. Burger et al. "Absence of the lactase-persistence-associated allele in early Neolithic Europeans". *Proc. Natl. Acad. Sci. U. S. A.* 104.10 (Mar. 2007), pp. 3736–3741.

[135] A. Durvasula et al. “African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana”. *Proc. Natl. Acad. Sci. U. S. A.* 114.20 (May 2017), pp. 5213–5218.

[136] M. Exposito-Alonso et al. “The rate and potential relevance of new mutations in a colonizing plant lineage”. *PLoS Genet.* 14.2 (Feb. 2018), e1007155.

[137] D. H. Huson et al. “MEGAN analysis of metagenomic data”. *Genome Res.* 17.3 (Mar. 2007), pp. 377–386.

[138] K. I. Bos et al. “A draft genome of Yersinia pestis from victims of the Black Death”. *Nature* 478.7370 (Oct. 2011), pp. 506–510.

[139] S. Rasmussen et al. “Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago”. *Cell* 163.3 (Oct. 2015), pp. 571–582.

[140] C. Warinner et al. “Pathogens and host immunity in the ancient human oral cavity”. *Nat. Genet.* 46.4 (Apr. 2014), pp. 336–344.

[141] C. J. Adler et al. “Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions”. *Nat. Genet.* 45.4 (Apr. 2013), 450–5, 455e1.

[142] L. S. Weyrich et al. “Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus”. *Nature* (Mar. 2017).

[143] K. A. Ziesemer et al. “Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification”. *Sci. Rep.* 5 (Nov. 2015), p. 16498.

[144] C. Warinner et al. “A Robust Framework for Microbial Archaeology”. *Annu. Rev. Genomics Hum. Genet.* 18 (Aug. 2017), pp. 321–356.

[145] M. W. Pedersen et al. “Postglacial viability and colonization in North America’s ice-free corridor”. *Nature* 537.7618 (Aug. 2016), pp. 45–49.

[146] M. W. Pedersen et al. “Ancient and modern environmental DNA”. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370.1660 (Jan. 2015), p. 20130383.

[147] L. Parducci et al. “Ancient plant DNA in lake sediments”. *New Phytol.* 214.3 (May 2017), pp. 924–942.

[148] C. L. Weiß et al. “Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens”. *R Soc Open Sci* 3.6 (June 2016), p. 160239.

[149] J. Dabney, M. Meyer, and S. Pääbo. “Ancient DNA Damage”. *Cold Spring Harb. Perspect. Biol.* (May 2013), a012567.

[150] B. Shapiro and M. Hofreiter. “A Paleogenomic Perspective on Evolution and Gene Function: New Insights from Ancient DNA”. *Science* 343.6169 (Jan. 2014), p. 1236573.

[151] M. Kircher. “Analysis of high-throughput ancient DNA sequencing data”. *Methods Mol. Biol.* 840 (2012), pp. 197–228.

[152] S. Sawyer et al. “Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA”. *PLoS One* 7.3 (Mar. 2012), e34131.

[153] M. Pruvost et al. “Freshly excavated fossil bones are best for amplification of ancient DNA”. *Proc. Natl. Acad. Sci. U. S. A.* 104.3 (Jan. 2007), pp. 739–744.

[154] T. Särkinen et al. “How to open the treasure chest? Optimising DNA extraction from herbarium specimens”. *PLoS One* 7.8 (Aug. 2012), e43808.

[155] T. Lindahl and B. Nyberg. “Rate of depurination of native deoxyribonucleic acid”. *Biochemistry* 11.19 (Sept. 1972), pp. 3610–3618.

[156] M. Stiller et al. “Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA”. *Proc. Natl. Acad. Sci. U. S. A.* 103.37 (Sept. 2006), pp. 13578–13584.

[157] T. Lindahl and B. Nyberg. “Heat-induced deamination of cytosine residues in deoxyribonucleic acid”. *Biochemistry* 13.16 (July 1974), pp. 3405–3410.

[158] P. Brotherton et al. “Novel high-resolution characterization of ancient DNA reveals C> U-type base modification events as the sole cause of post mortem miscoding lesions”. *Nucleic Acids Res.* 35.17 (2007), pp. 5717–5728.

[159] B. E. Deagle, J. P. Eveson, and S. N. Jarman. “Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces”. *Front. Zool.* 3.1 (2006), p. 1.

[160] T. Lindahl. “Instability and decay of the primary structure of DNA”. *Nature* 362.6422 (Apr. 1993), pp. 709–715.

[161] M. Molak and S. Y. W. Ho. “Evaluating the impact of post-mortem damage in ancient DNA: a theoretical approach”. *J. Mol. Evol.* 73.3-4 (Oct. 2011), pp. 244–255.

[162] M. E. Allentoft et al. “The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils”. *Proceedings of the Royal Society of London B: Biological Sciences* 279.1748 (Dec. 2012), pp. 4724–4733.

[163] M. Brundin et al. “DNA binding to hydroxyapatite: a potential mechanism for preservation of microbial DNA”. *J. Endod.* 39.2 (Feb. 2013), pp. 211–216.

[164] G. W. Lundeen. “Preservation of paper based materials: Present and future research and developments in the paper industry”. *Allerton Park Institute (27th: 1981)* (1983).

[165] R. M. Gutaker et al. “Extraction of ultrashort DNA molecules from herbarium specimens”. *Biotechniques* 62.2 (Feb. 2017), pp. 76–79.

[166] M. M.-Y. Tin, E. P. Economo, and A. S. Mikheyev. “Sequencing degraded DNA from non-destructively sampled museum specimens for RAD-tagging and low-coverage shotgun phylogenetics”. *PLoS One* 9.5 (May 2014), e96793.

[167] P. Skoglund et al. “Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal”. *Proc. Natl. Acad. Sci. U. S. A.* 111.6 (Feb. 2014), pp. 2229–2234.

[168] L. Kistler. “Ancient DNA extraction from plants”. *Methods Mol. Biol.* 840 (2012), pp. 71–79.

[169] M. Kircher, S. Sawyer, and M. Meyer. “Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform”. *Nucleic Acids Res.* 40.1 (Jan. 2012), e3.

[170] H. Jiang et al. “Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads”. *BMC Bioinformatics* 15 (June 2014), p. 182.

[171] T. Magoč and S. L. Salzberg. “FLASH: fast length adjustment of short reads to improve genome assemblies”. *Bioinformatics* 27.21 (Nov. 2011), pp. 2957–2963.

[172] Arabidopsis Genome Initiative. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. *Nature* 408.6814 (Dec. 2000), pp. 796–815.

[173] D. Swarbreck et al. “The Arabidopsis Information Resource (TAIR): gene structure and function annotation”. *Nucleic Acids Res.* 36.Database issue (Jan. 2008), pp. D1009–14.

[174] Potato Genome Sequencing Consortium et al. “Genome sequence and analysis of the tuber crop potato”. *Nature* 475.7355 (July 2011), pp. 189–195.

[175] Tomato Genome Consortium. “The tomato genome sequence provides insights into fleshy fruit evolution”. *Nature* 485.7400 (May 2012), pp. 635–641.

[176] B. J. Haas et al. “Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans”. *Nature* 461.7262 (Sept. 2009), pp. 393–398.

[177] H. Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM” (Mar. 2013).

[178] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, NY, 2002. ISBN: 9781441930088.

[179] R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008.

[180] H. Jónsson et al. “mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters”. *Bioinformatics* 29.13 (July 2013), pp. 1682–1684.

[181] C. L. Weiß et al. “Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data”. *Elife* 4 (Nov. 2015).

[182] C. L. Weiß et al. “Mining ancient microbiomes using selective enrichment of damaged DNA molecules”. Jan. 2019.

[183] K. Prüfer and M. Meyer. “Comment on “Late Pleistocene human skeleton and mtDNA link Paleoamericans and modern Native Americans””. *Science* 347.6224 (Feb. 2015), pp. 835–835.

[184] L. Orlando, M. T. P. Gilbert, and E. Willerslev. “Reconstructing ancient genomes and epigenomes”. *Nat. Rev. Genet.* 16.7 (July 2015), pp. 395–408.

[185] R. E. Green et al. “The Neandertal genome and ancient DNA authenticity”. *EMBO J.* 28.17 (2009), pp. 2494–2502.

[186] J. Dröge and A. C. McHardy. “Taxonomic binning of metagenome samples generated by next-generation sequencing technologies”. *Brief. Bioinform.* 13.6 (Nov. 2012), pp. 646–655.

[187] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. *J. Mol. Biol.* 147.1 (Mar. 1981), pp. 195–197.

[188] S. F. Altschul et al. “Basic local alignment search tool”. *J. Mol. Biol.* 215.3 (Oct. 1990), pp. 403–410.

[189] S. F. Altschul et al. “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”. *Nucleic Acids Res.* 25.17 (Sept. 1997), pp. 3389–3402.

[190] D. E. Wood and S. L. Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments”. *Genome Biol.* 15.3 (Mar. 2014), R46.

[191] A. L. Delcher et al. “Alignment of whole genomes”. *Nucleic Acids Res.* 27.11 (June 1999), pp. 2369–2376.

[192] B. Buchfink, C. Xie, and D. H. Huson. “Fast and sensitive protein alignment using DIAMOND”. *Nat. Methods* 12.1 (Jan. 2015), pp. 59–60.

[193] I. M. Velsko et al. “Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research”. Feb. 2018.

[194] F. M. Key et al. “Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication”. *Trends Genet.* 33.8 (Aug. 2017), pp. 508–520.

[195] A. Herbig et al. “MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman”. Jan. 2016.

[196] Å. J. Vågene et al. “Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico”. *Nat Ecol Evol* 2.3 (Mar. 2018), pp. 520–528.

[197] R. Huebler et al. “HOPS: Automated detection and authentication of pathogen DNA in archaeological remains”. Feb. 2019.

[198] V. Slon et al. “Neandertal and Denisovan DNA from Pleistocene sediments”. *Science* (Apr. 2017), eaam9695.

[199] O. Smith et al. “Sedimentary DNA from a submerged site reveals wheat in the British Isles 8000 years ago”. *Science* 347.6225 (Feb. 2015), pp. 998–1001.

[200] International Wheat Genome Sequencing Consortium (IWGSC). “A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome”. *Science* 345.6194 (July 2014), p. 1251788.

[201] X. Xu and U. Arnason. "A complete sequence of the mitochondrial genome of the western lowland gorilla". *Mol. Biol. Evol.* 13.5 (May 1996), pp. 691–698.

[202] D. M. Church et al. "Modernizing reference genome assemblies". *PLoS Biol.* 9.7 (July 2011), e1001091.

[203] R. Schmieder and R. Edwards. "Quality control and preprocessing of metagenomic datasets". *Bioinformatics* 27.6 (Mar. 2011), pp. 863–864.

[204] J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, Dec. 2006. ISBN: 9780387400655.

[205] K.-V. Yuen. *Bayesian Methods for Structural Dynamics and Civil Engineering*. John Wiley & Sons, Feb. 2010. ISBN: 9780470824559.

[206] I. Olalde et al. "Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European". *Nature* 507.7491 (Mar. 2014), pp. 225–228.

[207] J. A. Chapman et al. "A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome". *Genome Biol.* 16 (Jan. 2015), p. 26.

[208] M.-T. Gansauge and M. Meyer. "Selective enrichment of damaged DNA molecules for ancient genome sequencing". *Genome Res.* 24.9 (Sept. 2014), pp. 1543–1549.

[209] P. Skoglund et al. "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe". *Science* 336.6080 (Apr. 2012), pp. 466–469.

[210] G. Renaud, U. Stenzel, and J. Kelso. "leeHom: adaptor trimming and merging for Illumina sequencing reads". *Nucleic Acids Res.* 42.18 (Oct. 2014), e141.

[211] P. S. Schnable et al. "The B73 maize genome: complexity, diversity, and dynamics". *Science* 326.5956 (Nov. 2009), pp. 1112–1115.

[212] M. Nolan et al. "Complete genome sequence of Streptosporangium roseum type strain (NI 9100 T)". *Stand. Genomic Sci.* 2.1 (2010), p. 29.

[213] H. Feil et al. "Comparison of the complete genome sequences of Pseudomonas syringae pv. syringae B728a and pv. tomato DC3000". *Proc. Natl. Acad. Sci. U. S. A.* 102.31 (Aug. 2005), pp. 11064–11069.

[214] Y. Kwak, B. K. Jung, and J.-H. Shin. "Complete genome sequence of Pseudomonas rhizosphaerae IH5 T (= DSM 16299 T), a phosphate-solubilizing rhizobacterium for bacterial biofertilizer". *J. Biotechnol.* 193 (2015), pp. 137–138.

[215] T. H. M. Smits et al. "Genome sequence of the biocontrol agent Pantoea vagans strain C9-1". *J. Bacteriol.* 192.24 (Dec. 2010), pp. 6486–6487.

[216] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". *Nat. Methods* 9.4 (Mar. 2012), pp. 357–359.

[217] H. Li. "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data". *Bioinformatics* 27.21 (Nov. 2011), pp. 2987–2993.

[218] H. Li. “Minimap2: pairwise alignment for nucleotide sequences”. *Bioinformatics* (May 2018).

[219] B. J. Knaus and N. J. Grünwald. “vcfr: a package to manipulate and visualize variant call format data in R”. *Mol. Ecol. Resour.* 17.1 (Jan. 2017), pp. 44–53.

[220] T. Jombart and I. Ahmed. “adegenet 1.3-1: new tools for the analysis of genome-wide SNP data”. *Bioinformatics* 27.21 (Nov. 2011), pp. 3070–3071.

[221] A. Bankevich et al. “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing”. *J. Comput. Biol.* 19.5 (May 2012), pp. 455–477.

[222] R. S. Harris. *Improved pairwise alignment of genomic DNA*. The Pennsylvania State University, 2007.

[223] M. Krzywinski et al. “Circos: an information aesthetic for comparative genomics”. *Genome Res.* 19.9 (Sept. 2009), pp. 1639–1645.

[224] S. Hohlfeld et al. *AliTV: Version 0.4.1*. 2016.

[225] K. Zaremba-Niedźwiedzka and S. G. E. Andersson. “No ancient DNA damage in Actinobacteria from the Neanderthal bone”. *PLoS One* 8.5 (May 2013), e62799.

[226] Austin Jeremy J. et al. “Problems of reproducibility – does geologically ancient DNA survive in amber–preserved insects?” *Proceedings of the Royal Society of London. Series B: Biological Sciences* 264.1381 (Apr. 1997), pp. 467–474.

[227] M. Stoneking. “Ancient DNA: how do you know when you have it and what can you do with it?” *Am. J. Hum. Genet.* 57.6 (Dec. 1995), pp. 1259–1262.

[228] R. Ward and C. Stringer. “A molecular handle on the Neanderthals”. *Nature* 388.6639 (July 1997), pp. 225–226.

[229] C. L. Weiß et al. “nQuire: a statistical framework for ploidy estimation using next generation sequencing”. *BMC Bioinformatics* 19.1 (Apr. 2018), p. 122.

[230] K. L. Adams and J. F. Wendel. “Polyploidy and genome evolution in plants”. *Curr. Opin. Plant Biol.* 8.2 (Apr. 2005), pp. 135–141.

[231] G. Blanc and K. H. Wolfe. “Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes”. *Plant Cell* 16.7 (July 2004), pp. 1667–1678.

[232] J. Ramsey and D. W. Schemske. “PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS”. *Annu. Rev. Ecol. Syst.* 29.1 (Nov. 1998), pp. 467–501.

[233] D. E. Soltis et al. “Autopolyploidy in Angiosperms: Have We Grossly Underestimated the Number of Species?” *Taxon* 56.1 (2007), pp. 13–30.

[234] S. P. Otto. “The evolutionary consequences of polyploidy”. *Cell* 131.3 (Nov. 2007), pp. 452–462.

[235] S. P. Otto and J. Whitton. "Polyploid incidence and evolution". *Annu. Rev. Genet.* 34 (2000), pp. 401–437.

[236] G. Bell. *The Masterpiece of Nature :: The Evolution and Genetics of Sexuality.* CUP Archive, 1982. ISBN: 9780856647536.

[237] L. Comai. "The advantages and disadvantages of being polyploid". *Nat. Rev. Genet.* 6.11 (Nov. 2005), pp. 836–846.

[238] P. W. Tooley and C. D. Therrien. "Cytophotometric determination of the nuclear DNA content of 23 Mexican and 18 non-Mexican isolates ofPhytophthora infestans". *Exp. Mycol.* 11.1 (Mar. 1987), pp. 19–26.

[239] Y. Harari et al. "Spontaneous Changes in Ploidy Are Common in Yeast". *Curr. Biol.* 0.0 (Mar. 2018).

[240] S. Venkataram et al. "Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast". *Cell* 166.6 (Sept. 2016), 1585–1596.e22.

[241] J. Peter et al. "Genome evolution across 1,011 Saccharomyces cerevisiae isolates". *Nature* (Apr. 2018).

[242] Y. Li et al. "Changing Ploidy as a Strategy: The Irish Potato Famine Pathogen Shifts Ploidy in Relation to Its Sexuality". *Mol. Plant. Microbe. Interact.* 30.1 (Jan. 2017), pp. 45–52.

[243] A. M. Selmecki et al. "Polyploidy can drive rapid adaptation in yeast". *Nature* 519.7543 (Mar. 2015), pp. 349–352.

[244] Z. Storchova and D. Pellman. "From polyploidy to aneuploidy, genome instability and cancer". *Nat. Rev. Mol. Cell Biol.* 5.1 (Jan. 2004), pp. 45–54.

[245] S. Dirihan et al. "Efficient analysis of ploidy levels in plant evolutionary ecology". *Caryologia* 66.3 (Sept. 2013), pp. 251–256.

[246] B. Chor et al. "Genomic DNA k-mer spectra: models and modalities". *Genome Biol.* 10.10 (Oct. 2009), R108.

[247] R. Chikhi and P. Medvedev. "Informed and automated k-mer size selection for genome assembly". *Bioinformatics* 30.1 (Jan. 2014), pp. 31–37.

[248] Y. O. Zhu, G. Sherlock, and D. A. Petrov. "Whole Genome Analysis of 132 Clinical Saccharomyces cerevisiae Strains Reveals Extensive Ploidy Variation". *G3* 6.8 (Aug. 2016), pp. 2421–2434.

[249] R. Augusto Corrêa Dos Santos, G. H. Goldman, and D. M. Riaño-Pachón. "ploidyNGS: visually exploring ploidy with Next Generation Sequencing data". *Bioinformatics* 33.16 (Aug. 2017), pp. 2575–2576.

[250] L. Scrucca et al. "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". *R J.* 8.1 (Aug. 2016), pp. 289–317.

[251] G. Celeux and G. Govaert. "Gaussian parsimonious clustering models". *Pattern Recognit.* 28.5 (May 1995), pp. 781–793.

[252] D. E. L. Cooke et al. “Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen”. *PLoS Pathog.* 8.10 (Oct. 2012), e1002940.

[253] S. Deorowicz et al. “KMC 2: fast and resource-frugal k-mer counting”. *Bioinformatics* 31.10 (May 2015), pp. 1569–1576.

[254] Z. Gompert and K. E. Mock. “Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis”. *Mol. Ecol. Resour.* (Feb. 2017).

[255] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. *Nature* 526.7571 (Oct. 2015), pp. 68–74.

[256] C. Bycroft et al. “Genome-wide genetic data on 500,000 UK Biobank participants”. July 2017.

[257] A. G. Clark et al. “Ascertainment bias in studies of human genome-wide polymorphism”. *Genome Res.* 15.11 (2005), pp. 1496–1502.

[258] V. Sousa and J. Hey. “Understanding the origin of species with genome-scale data: modelling gene flow”. *Nat. Rev. Genet.* 14.6 (June 2013), pp. 404–414.

[259] G. Lunter and M. Goodson. “Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads”. *Genome Res.* 21.6 (June 2011), pp. 936–939.

[260] D. Y. C. Brandt et al. “Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data”. *G3* 5.5 (Mar. 2015), pp. 931–941.

[261] T. Günther and C. Nettelblad. “The presence and impact of reference bias on population genomic studies of prehistoric human populations”. Dec. 2018.

[262] R. Poplin et al. “A universal SNP and small-indel variant caller using deep neural networks”. *Nat. Biotechnol.* 36.10 (Nov. 2018), pp. 983–987.

[263] A. Barlow et al. “Consensify: a method for generating pseudohaploid genome sequences from palaeogenomic datasets with reduced error rates”. Dec. 2018.

[264] V. Slon et al. “A fourth Denisovan individual”. *Sci Adv* 3.7 (July 2017), e1700186.

[265] S. Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses”. *Am. J. Hum. Genet.* 81.3 (Sept. 2007), pp. 559–575.

[266] I. Lazaridis et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. *Nature* 513.7518 (Sept. 2014), pp. 409–413.

[267] E. Willerslev et al. “Ancient biomolecules from deep ice cores reveal a forested southern Greenland”. *Science* 317.5834 (July 2007), pp. 111–114.

[268] T. F. G. Higham, R. M. Jacobi, and C. Bronk Ramsey. “AMS Radiocarbon Dating of Ancient Bone Using Ultrafiltration”. *Radiocarbon* 48.2 (2006), pp. 179–195.

[269] Z. Jacobs and R. G. Roberts. “Advances in optically stimulated luminescence dating of individual grains of quartz from archeological deposits”. *Evol. Anthropol.* Conference Proceedings 16.6 (Dec. 2007), pp. 210–223.

[270] M. G. Lorenz and W. Wackernagel. “Adsorption of DNA to sand and variable degradation rates of adsorbed DNA”. *Appl. Environ. Microbiol.* 53.12 (Dec. 1987), pp. 2948–2952.

[271] M. P. Greaves and M. J. Wilson. “The adsorption of nucleic acids by montmorillonite”. *Soil Biol. Biochem.* 1.4 (Nov. 1969), pp. 317–323.

[272] J. Haile et al. “Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor?” *Mol. Biol. Evol.* 24.4 (Apr. 2007), pp. 982–989.

[273] K. Andersen et al. “Meta-barcoding of ’dirt’ DNA from soil reflects vertebrate biodiversity”. *Mol. Ecol.* 21.8 (Apr. 2012), pp. 1966–1979.

[274] Z. Jacobs et al. “Testing of a single grain OSL chronology across the Middle to Upper Palaeolithic transition at Les Cottés (France)”. *J. Archaeol. Sci.* 54 (Feb. 2015), pp. 110–122.

[275] S. Talamo et al. “A radiocarbon chronology for the complete Middle to Upper Palaeolithic transitional sequence of Les Cottés (France)”. *J. Archaeol. Sci.* 39.1 (Jan. 2012), pp. 175–183.

[276] O. Jaillon et al. “The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla”. *Nature* 449.7161 (Sept. 2007), pp. 463–467.

[277] Y. Wan et al. “A phylogenetic analysis of the grape genus (Vitis L.) reveals broad reticulation and concurrent diversification during neogene and quaternary climate change”. *BMC Evol. Biol.* 13 (July 2013), p. 141.

[278] P. This, T. Lacombe, and M. R. Thomas. “Historical origins and genetic diversity of wine grapes”. *Trends Genet.* 22.9 (Sept. 2006), pp. 511–519.

[279] S. Myles et al. “Genetic structure and domestication history of the grape”. *Proc. Natl. Acad. Sci. U. S. A.* 108.9 (Mar. 2011), pp. 3530–3535.

[280] N. F. Miller. “Sweeter than wine? The use of the grape in early western Asia”. *Antiquity* 82.318 (Dec. 2008), pp. 937–946.

[281] P. E. McGovern et al. “Neolithic resinated wine”. *Nature* 381.6582 (June 1996), pp. 480–481.

[282] H. Barnard et al. “Chemical evidence for wine production around 4000 BCE in the Late Chalcolithic Near Eastern highlands”. *J. Archaeol. Sci.* 38.5 (May 2011), pp. 977–984.

[283] P. McGovern et al. “Early Neolithic wine of Georgia in the South Caucasus”. *Proc. Natl. Acad. Sci. U. S. A.* 114.48 (Nov. 2017), E10309–E10318.

[284] J.-F. Terral et al. “Evolution and history of grapevine (Vitis vinifera) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars”. *Ann. Bot.* 105.3 (Mar. 2010), pp. 443–455.

[285] D. Zohary. "The domestication of the grapevine Vitis vinifera L. in the Near East". *The origins and ancient history of wine*. Routledge, 2003, pp. 44–51.

[286] D. Zohary, M. Hopf, and E. Weiss. *Domestication of Plants in the Old World: The Origin and Spread of Domesticated Plants in Southwest Asia, Europe, and the Mediterranean Basin*. OUP Oxford, Mar. 2012. ISBN: 9780199549061.

[287] S. Myles et al. "Rapid genomic characterization of the genus vitis". *PLoS One* 5.1 (Jan. 2010), e8219.

[288] J. Sawler et al. "Genomics assisted ancestry deconvolution in grape". *PLoS One* 8.11 (Nov. 2013), e80791.

[289] G. Renaud et al. "freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers". *Bioinformatics* 29.9 (May 2013), pp. 1208–1209.

[290] E. Hodges et al. "Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing". *Nat. Protoc.* 4.6 (May 2009), pp. 960–974.

[291] Q. Fu et al. "DNA analysis of an early modern human from Tianyuan Cave, China". *Proc. Natl. Acad. Sci. U. S. A.* 110.6 (Feb. 2013), pp. 2223–2227.

[292] A. Peltzer et al. "EAGER: efficient ancient genome reconstruction". *Genome Biol.* 17 (Mar. 2016), p. 60.

[293] R. K. Jansen et al. "Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids". *BMC Evol. Biol.* 6 (Apr. 2006), p. 32.

[294] A. R. Quinlan and I. M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842.

[295] C. C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* 4 (Feb. 2015), p. 7.

[296] J. Graffelman and V. Moreno. "The mid p-value in exact tests for Hardy-Weinberg equilibrium". *Stat. Appl. Genet. Mol. Biol.* 12.4 (Aug. 2013), pp. 433–448.

[297] P. Skoglund et al. "Genetic evidence for two founding populations of the Americas". *Nature* 525.7567 (Sept. 2015), pp. 104–108.

[298] N. Wales et al. "The limits and potential of paleogenomic techniques for reconstructing grapevine domestication". *J. Archaeol. Sci.* 72 (2016), pp. 57–70.

[299] Y. Zhou et al. "Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication". *Proc. Natl. Acad. Sci. U. S. A.* 114.44 (Oct. 2017), pp. 11715–11720.

[300] Z. Liang et al. "Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses". *Nat. Commun.* 10.1 (Mar. 2019), p. 1190.

[301] E. R. Mardis. “The impact of next-generation sequencing technology on genetics”. *Trends Genet.* 24.3 (Mar. 2008), pp. 133–141.

[302] D. C. Koboldt et al. “The next-generation sequencing revolution and its impact on genomics”. *Cell* 155.1 (Sept. 2013), pp. 27–38.

[303] F. W. Allendorf, P. A. Hohenlohe, and G. Luikart. “Genomics and the future of conservation genetics”. *Nat. Rev. Genet.* 11.10 (Oct. 2010), pp. 697–709.

[304] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. “Prediction of total genetic value using genome-wide dense marker maps”. *Genetics* 157.4 (Apr. 2001), pp. 1819–1829.

[305] G. de Los Campos et al. “Whole-genome regression and prediction methods applied to plant and animal breeding”. *Genetics* 193.2 (Feb. 2013), pp. 327–345.

[306] Z. N. Kronenberg et al. “High-resolution comparative analysis of great ape genomes”. *Science* 360.6393 (June 2018).

[307] E. A. Boyle, Y. I. Li, and J. K. Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. *Cell* 169.7 (June 2017), pp. 1177–1186.

[308] C. Der Sarkissian et al. “Ancient genomics”. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370.1660 (Jan. 2015), p. 20130387.

[309] F. T. Bakker. “Herbarium Genomics: Plant Archival DNA Explored”. *Paleogenomics: Genome-Scale Analysis of Ancient DNA*. Ed. by C. Lindqvist and O. P. Rajora. Cham: Springer International Publishing, 2019, pp. 205–224. ISBN: 9783030047535.

Publication List

[1] **C. L. Weiß**, M.-T. Gansauge, A. Aximu-Petri, M. Meyer, and H. A. Burbano. "Mining ancient microbiomes using selective enrichment of damaged DNA molecules". *bioRxiv* (Jan. 2019), p. 397927.

[2] **C. L. Weiß**, M. Pais, L. M. Cano, S. Kamoun, and H. A. Burbano. "nQuire: a statistical framework for ploidy estimation using next generation sequencing". *BMC Bioinformatics* 19.1 (Apr. 2018), p. 122.

[3] M. J Ankenbrand, S. Pfaff, N. Terhoeven, M. Qureischi, M. Gündel, **C. L Weiß**, T. Hackl, and F. Förster. "chloroExtractor: extraction and assembly of the chloroplast genome from whole genome shotgun data". *The Journal of Open Source Software* 3 (Jan. 2018), p. 464.

[4] V. Slon, C. Hopfe, **C. L. Weiß**, F. Mafessoni, M. de la Rasilla, C. Lalueza-Fox, A. Rosas, M. Soressi, M. V. Knul, R. Miller, J. R. Stewart, A. P. Derevianko, Z. Jacobs, B. Li, R. G. Roberts, M. V. Shunkov, H. de Lumley, C. Perrenoud, I. Gušić, Ž. Kućan, P. Rudan, A. Aximu-Petri, E. Essel, S. Nagel, B. Nickel, A. Schmidt, K. Prüfer, J. Kelso, H. A. Burbano, S. Pääbo, and M. Meyer. "Neandertal and Denisovan DNA from Pleistocene sediments". *Science* (Apr. 2017), eaam9695.

[5] C. Warinner, A. Herbig, A. Mann, J. A. Fellows Yates, **C. L. Weiß**, H. A. Burbano, L. Orlando, and J. Krause. "A Robust Framework for Microbial Archaeology". *Annu. Rev. Genomics Hum. Genet.* (Sept. 2016).

[6] **C. L. Weiß**, V. J. Schuenemann, J. Devos, G. Shirsekar, E. Reiter, B. A. Gould, J. R. Stinchcombe, J. Krause, and H. A. Burbano. "Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens". *R Soc Open Sci* 3.6 (June 2016), p. 160239.

[7] F. Bemm, **C. L. Weiß**, J. Schultz, and F. Förster. "Genome of a tardigrade: Horizontal gene transfer or bacterial contamination?" *Proc. Natl. Acad. Sci. U. S. A.* 113.22 (May 2016), E3054–6.

[8] **C. L. Weiß**, M. Dannemann, K. Prüfer, and H. A. Burbano. "Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data". *Elife* 4 (Nov. 2015).

[9] **C. L. Weiß** and J. Schultz. "Identification of divergent WH2 motifs by HMM-HMM alignments". *BMC Res. Notes* 8 (Jan. 2015), p. 18.

Abbreviations

A,C,G,T,U	the nucleobases adenine, cytosine, guanine, thymine and uracil
DNA	Deoxyribonucleic acid
aDNA	ancient DNA
PCR	Polymerase Chain Reaction
qPCR	quantitative PCR
PTB	N-phenacyl thiazolium bromide
CTAB	Cetyl trimethylammonium bromide
UDG	Uracil-DNA glycosylase
UV	Ultraviolet
HTS	High Throughput Sequencing
SNP	Single Nucleotide Polymorphism
SAM	Sequence Alignment/Map
BAM	Binary Alignment Map
RG	Read Group
BED	Browser Extensible Data
VCF	Variant Call Format
GATK	Genome Analysis ToolKit
PCA	Principle Component Analysis
bp	base pair
nt	nucleotide
LCA	Lowest Common Ancestor
GSL	GNU Scientific Library
BLAS	Basic Linear Algebra Subprograms
GMM	Gaussian Mixture Model
GMMU	Gaussian Mixture Model with Uniform noise
EM	Expectation-Maximization
MVN	Mutivariate Normal
USDA	United States Department of Agriculture
NCBI	National Center for Biotechnology Information
ENA	European Nucleotide Archive
API	Application Programming Interface
GB	Gigabyte
RAM	Random Access Memory
CPU	Central Processing Unit

Supplementary Material

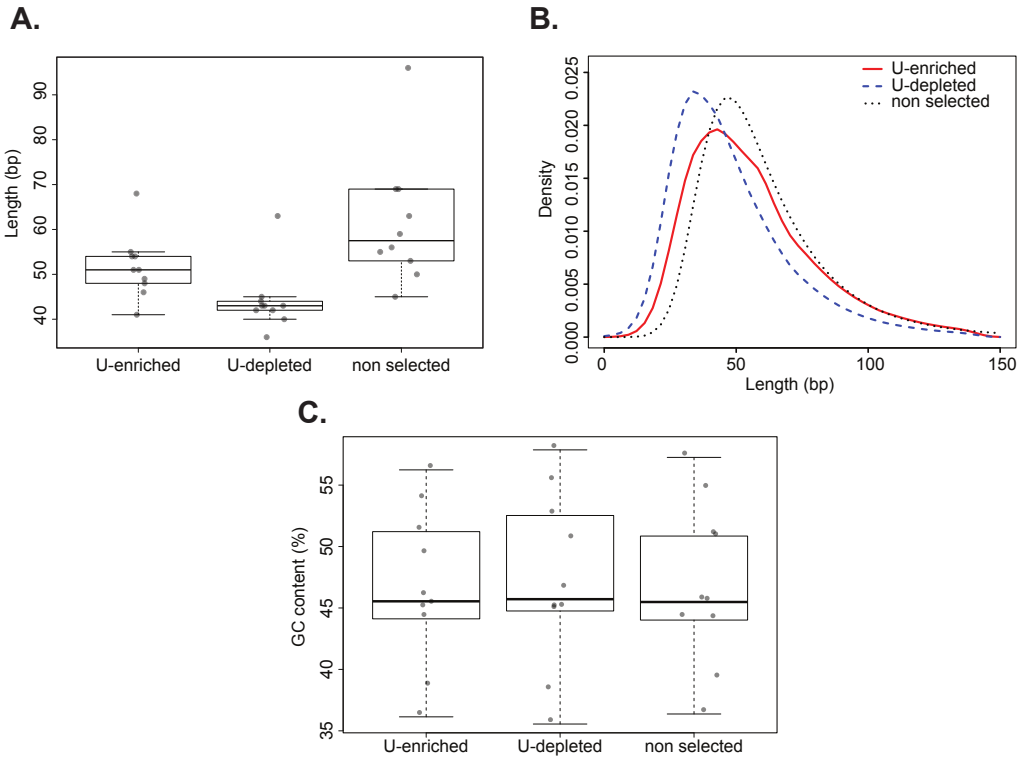

Figure S1: Length and GC content of plant historical specimens. **A.** Distributions of mean length for non-selected and U-selected libraries (U-enriched and U-depleted fractions). Median values are denoted as black lines and points show the original value for each individual sample. **B.** Length distribution of *Arabidopsis thaliana* sample NY1365375 for a non-selected and U-selected library (U-enriched and U-depleted fractions). **C.** Distributions of mean GC content for non-selected and U-selected libraries (U-enriched and U-depleted fractions).

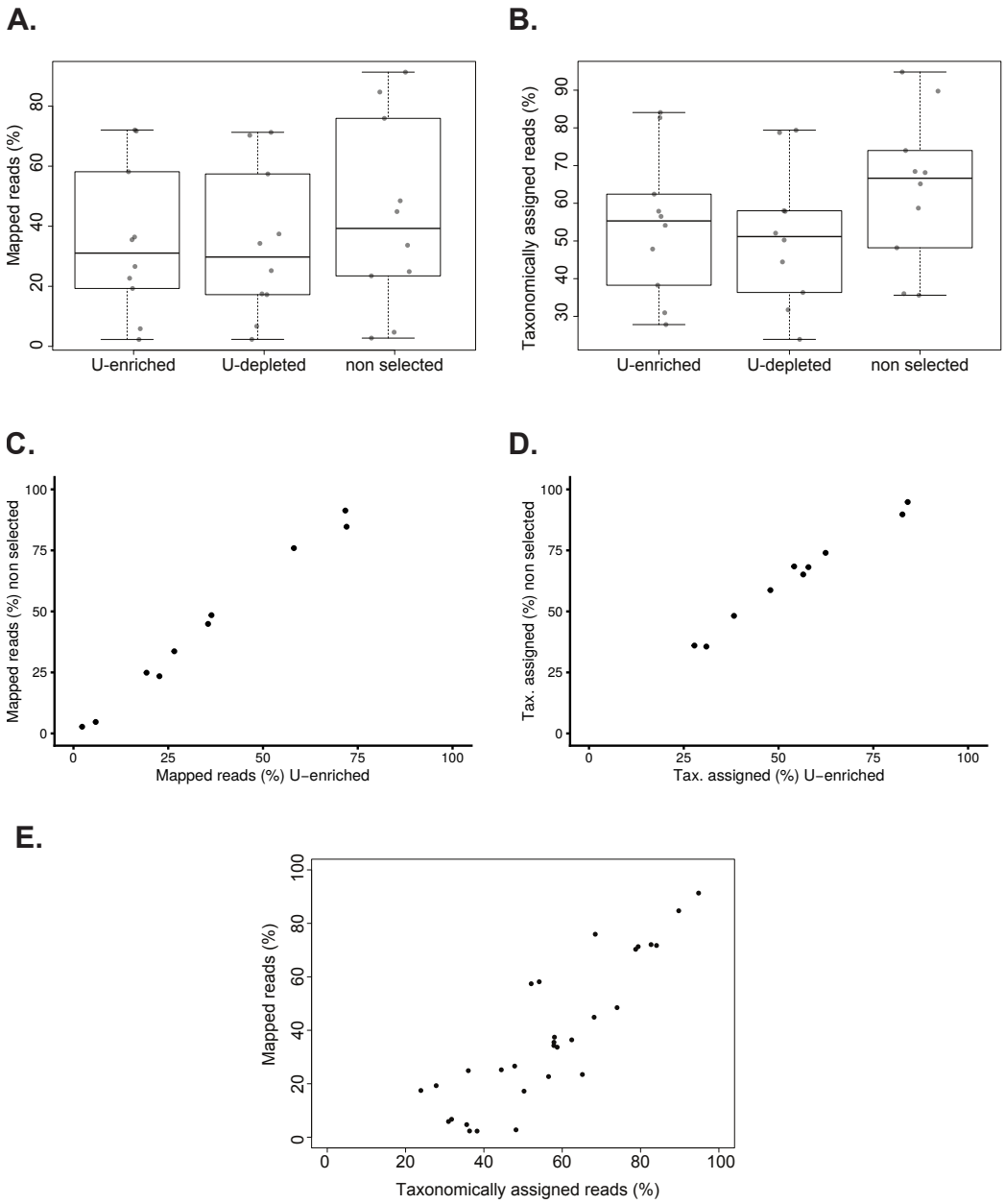

Figure S2: Mapped and taxonomically assigned reads of plant historical specimens. A. Distributions of percentage of mapped reads for non-selected and U-selected libraries (U-enriched and U-depleted fractions). **B.** Distributions of percentage of taxonomically assigned reads for non-selected and U-selected libraries (U-enriched and U-depleted fractions). **C.** Correlation of the percentage of mapped reads between the U-enriched and the non-selected library. **D.** Correlation of the percentage of taxonomically assigned reads between the U-enriched and the non-selected library. **E.** Relation between percentages of mapped and taxonomically assigned reads from U-selected libraries (U-enriched fraction).

Table S1: Provenance of herbarium samples used in Chapter 2.

ID	Country of origin	Collection year	Species	Source
BH0000052681	USA	1919	A. thaliana	1
BH0000061407	USA	1890	A. thaliana	1
BH0000061409	USA	1892	A. thaliana	1
BH0000061416	USA	1892	A. thaliana	1
BH0000061457	USA	1900	A. thaliana	1
BH0000061454	USA	1884	A. thaliana	1
BH0000061397	USA	1899	A. thaliana	1
BH0000061417	USA	1912	A. thaliana	1
BH0000061449	USA	1901	A. thaliana	1
BH0000061420	USA	1896	A. thaliana	1
BH0000061395	USA	1883	A. thaliana	1
BH0000061424	USA	1895	A. thaliana	1
BH0000061393	USA	1918	A. thaliana	1
BH0000061399	USA	1901	A. thaliana	1
BH0000061448	USA	1901	A. thaliana	1
BH0000061418	USA	1887	A. thaliana	1
BH0000061406	USA	1889	A. thaliana	1
OSU46488	USA	1917	A. thaliana	2
OSU13896	USA	1930	A. thaliana	2
OSU13900	USA	1934	A. thaliana	2
OSU54623	USA	1956	A. thaliana	2
OSU79944	USA	1911	A. thaliana	2
OSU183663	USA	1969	A. thaliana	2
OSU361885	USA	1930	A. thaliana	2
OSU150287	USA	1980	A. thaliana	2
OSU163761	USA	1981	A. thaliana	2
OSU364632	USA	1993	A. thaliana	2
UNC54051	USA	1945	A. thaliana	3
UNC25707	USA	1940	A. thaliana	3
UNC63978	USA	1910	A. thaliana	3
CT79391	USA	1904	A. thaliana	4
CT79409	USA	1929	A. thaliana	4
CT79389	USA	1975	A. thaliana	4
176849CFM	USA	1904	A. thaliana	5
531679CFM	USA	1922	A. thaliana	5
1507461CFM	USA	1952	A. thaliana	5
NY102365	USA	1903	A. thaliana	6
NY1365344	USA	1890	A. thaliana	6
NY888124	USA	1863	A. thaliana	6
NY1365363	USA	1888	A. thaliana	6
NY888144	USA	1866	A. thaliana	6
NY1365364	USA	1889	A. thaliana	6
NY888134	USA	1877	A. thaliana	6
NY1365332	USA	1890	A. thaliana	6
NY1365337	USA	1891	A. thaliana	6
NY1365370	USA	1897	A. thaliana	6
NY1365333	USA	1894	A. thaliana	6
NY1365374	USA	1896	A. thaliana	6
NY102364	USA	1904	A. thaliana	6

Continued on next page

Continued from previous page

ID	Country of origin	Collection year	Species	Source
NY1365375	USA	1897	A. thaliana	6
NY888141	USA	1879	A. thaliana	6
NY1365354	USA	1891	A. thaliana	6
NY888141	USA	1879	A. thaliana	6
NY1365354	USA	1891	A. thaliana	6
KM177509	England	1865	S. tuberosum	7
KM177517	Wales	1875	S. tuberosum	7
KM177514	Ireland	1847	S. tuberosum	7
KM177500	England	1845	S. tuberosum	7
KM177513	Ireland	1846	S. tuberosum	7
KM177548	England	1847	S. tuberosum	7
M-0182898	Germany	1863	S. tuberosum	8
M-0182906	Germany	1877	S. tuberosum	8
M-0182907	Germany	1875	S. tuberosum	8
M-0182896	Germany	1877	S. tuberosum	8
M-0182903	Canada	1896	S. tuberosum	8
M-0182904	Austria	1879	S. tuberosum	8
BM000777791	UK	1866	S. lycopersicum	9
BM000815937	UK	1737	S. lycopersicum	9
BM000849510	UK	1779	S. lycopersicum	9
M-0182897	USA	1876	S. lycopersicum	8
M-0182900	Germany	1873	S. lycopersicum	8

Sources:

1. Cornell Bailey Hortorium;
2. Ohio State University Herbarium;
3. University of North Carolina Herbarium;
4. University of Connecticut Herbarium;
5. Chicago Field Museum;
6. New York Botanical Garden;
7. Kew Royal Botanical Garden;
8. Botanische Staatssammlung München;
9. Natural History Museum, London;