Novel interactive methods for ancient and mitochondrial DNA analysis

Dissertation

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

vorgelegt von

M. Sc. Judith Neukamm

aus Aalen
Tübingen
2021

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

Tag der mündlichen Qualifikation:	12. November 2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Kay Nieselt
2. Berichterstatterin:	Prof. Dr. Dr. Verena J. Schünemann
3. Berichterstatter:	Prof. Dr. Daniel Huson

Dedicated to my parents
Brigitte & Berthold

Zusammenfassung

Das Aufkommen von Sequenzierungstechnologien der nächsten Generation hat zu einem Boom bei DNA-Sequenzierungsprojekten geführt. Solche Hochdurchsatz-Sequenzierungsmethoden erzeugen riesige Datenmengen, die vergleichend analysiert werden müssen, um Unterschiede zwischen verschiedenen Organismen oder Umweltbedingungen festzustellen. Außerdem wurde es möglich, den gesamten genomischen Inhalt einer Probe zu sequenzieren und nicht nur eine einzelne Art. Dies ermöglichte auch die Analyse von alten Proben. Diese Aufgaben erfordern effiziente Berechnungsmethoden, die genomische Daten aus mehreren Datensätzen auf automatisierte, effiziente und reproduzierbare Weise integrieren.

In dieser Dissertation wurden Beiträge zur Analyse metagenomischer Datensätze und spezifischer Genome moderner und alter Datensätze geleistet. Da alte DNA in der Regel stark geschädigt ist, wird neue oder angepasste Software benötigt, um den neuen Herausforderungen zu begegnen. Zunächst wird die Software DamageProfiler vorgestellt, die effizient die für alte DNA typischen Schädigungsmuster berechnet, wie z. B. die Häufigkeit von Basenfehlinkorporationen und die Fragmentlänge. Diese Merkmale werden zur Überprüfung des antiken Ursprungs der kartierten DNA-Fragmente verwendet und sollten in jede Analysepipeline für die Rekonstruktion alter Genome enthalten sein. Außerdem wird eine grafische Benutzeroberfläche zur Verfügung gestellt, um die Software auch für unerfahrene Benutzer zugänglich zu machen.

Die zweite vorgestellte Software ist mitoBench, die sich auf die Analyse moderner und alter vollständiger menschlicher mitochondrialer Genome konzentriert. Die Workbench bietet eine interaktive Möglichkeit zur Analyse und Visualisierung kompletter mitochondrialer Genome mit dem Schwerpunkt auf populationsgenetischen Anwendungen. Die gut kuratierte Datenbank, mit der die Workbench verknüpft ist, bietet einen mitochondrialen DNA-Referenzdatensatz, der aus hochwertigen Genomen und Metainformationen besteht.

Der letzte Teil dieser Arbeit befasst sich mit der Analyse von metagenomischen Datensätzen aus alten Proben. Es werden zwei Anwendungen der metagenomischen Analyse von bisher wenig untersuchten Geweben beschrieben: das Gewebe mumifizierter ägyptischer Individuen und eine menschliche Beinprobe eines Kleinkindes, die in Ethanol und Paraffin konserviert wurde. Neben dem Nachweis der mikrobiellen Zusammensetzung der Probe zeigen die Anwendungen die erfolgreiche Gewinnung von DNA aus verschiedenen Krankheitserregern und deren Genomrekonstruktion. Die Rekonstruktion der alten Genome von *Mycobacterium leprae*, Hepatitis-B-Virus, Variola-Virus und einer Reihe von oralen Krankheitserregern hat Beiträge zu deren geographischen Ausbreitung und Evolution erbracht.

Abstract

The advent of next-generation sequencing technologies has led to a boom in DNA sequencing projects. Such high-throughput sequencing methods generate vast amounts of data that need to be comparatively analyzed to identify differences between different organisms or environmental conditions. In addition, it became possible to sequence the entire genomic content of a sample rather than just a single species. Such new technologies also enabled the analysis of ancient samples. These tasks require efficient computational methods that integrate genomic data from multiple sources in an automated, efficient, and reproducible manner.

In this dissertation, contributions were made to analyzing metagenomic data sets and specific genomes of modern and ancient data sets.Since ancient DNA is usually highly damaged, new or adapted software is needed to meet the new challenges that arise. First, the software DamageProfiler is presented, efficiently calculating damage patterns typical for ancient DNA, such as base misincorporation frequency and fragment length. These features are used to verify the ancient origin of mapped DNA fragments and should be included in any analysis pipeline for ancient genome reconstruction. In addition, a GUI is also provided to make the software more accessible to inexperienced users.

The second software presented is mitoBench, which focuses on analyzing complete modern and ancient human mitochondrial genomes. The workbench provides an interactive way to explore and visualize complete mitogenomes, focusing on applications with population genetics. The well-curated database linked to the workbench provides a mitochondrial DNA reference data set consisting of high-quality genomes and meta information.

The final part of this thesis deals with the analysis of metagenomic data sets from ancient samples. Two applications of metagenomic analysis of previously under-studied tissues are described: the tissue of mummified Egyptian individuals and a human leg sample from an infant preserved in ethanol and paraffin. In addition to identifying the microbial composition of the samples, the applications show the successful recovery of DNA from various pathogens and subsequent genome reconstruction. Reconstruction of the ancient *Mycobacterium leprae*, hepatitis B virus, and Variola virus genomes has provided meaningful contributions to their geographical spread and insights into their evolution.

Acknowledgements

First of all, I would like to thank my two supervisors Prof. Dr. Kay Nieselt, group leader of the Integrative Transcriptomics Group at the University of Tübingen, and Prof. Dr. Dr. Verena J. Schuenemann, group leader of the Paleogenetics Group at the University of Zurich. Both Kay and Verena have supported me during my time as a Ph.D. student and have allowed me to work on a variety of exciting projects that have, not least, enabled me to write this thesis. I would also like to thank Prof. Dr. Daniel Huson for reviewing my work and Prof. Dr. Stephan Ossowski for participating on my committee.

Several marvelous people made it possible for the work to be completed. I thank all my colleagues at the Integrative Transcriptomics Group and the Paleogenetics Group at the University of Tübingen and the Paleogenetics Group the University of Zürich. I thank all of you for the many great moments, conversations, and Raclette and BBQ events. Having such wonderful colleagues and inspiring work environments has genuinely enriched my time.

Moreover, I would also like to thank Alexander Peltzer, Marie Gauder, Donikë Sejdiu, and Christian Urban for stimulating discussions (both on and off-topic), ideas, and their feedback on this thesis. A big thanks also goes to Abagail Breidenstein for reading my thesis.

I would like to thank all collaborators on my projects, especially mitoBench and metagenomic analysis of mummified Egyptian individuals, for their contributions, fruitful discussions, and assisting in the conception and completion of this thesis.

I also want to thank my partner Patrick for his support, especially during the last year of my Ph.D. You make my dreams come true!

Finally, I would most like to express my deepest gratitude to my family and friends, and this goes far beyond my doctoral studies and the writing of this dissertation.

Contents

List of Figures

List of Tables

List of Abbreviations

A	Adenine
API	Application programming interface
aDNA	Ancient DNA
BCE	Before common era
bp	Base pairs
C	Cytosine
CE	Common era
CI	Continuous integration
CLI	Command line interface
DNA	Deoxyribonucleic acid
dNTP	Nucleoside triphosphate
dsLib	Double-stranded library protocol
EAGER	Efficient ancient genome reconstruction
F_{st}	Fixation index, a measure of population differentiation
G	Guanine
GATK	Genome analysis toolkit
GUI	Graphical user interface
HBV	Hepatitis B virus
HOPS	Heuristic operations for pathogen screening
IGV	Integrative genome viewer
INDELs	Insertions and deletions
LGM	Last glacial maximum
MALT	MEGAN alignment tool
MDS	Multidimensional scaling
MRCA	Most recent common ancestor
mtDNA	Mitochondrial DNA
NGS	Next-generation sequencing
NUMT	Nuclear mitochondrial DNA segments

PE	Paired-end (sequencing)
PCA	Principal component analysis
PCR	Polymerase chain reaction
rCRS	Revised Cambridge Reference Sequence
RSRS	Reconstructed Sapiens Reference Sequence
SBS	Sequencing by synthesis
SE	Single-end (sequencing)
SNP	Single nucleotide polymorphism
ssLib	Single-stranded library protocol
T	Thymine
TMA	Terminal maternal ancestor
U	Uracil
UDG	Uracil-deglycosylase (treatment)
VARV	Variola virus
VCF	Variant call format

CHAPTER 1

Introduction

The research field of ancient DNA (aDNA) explores the genetic background of archaeological and paleontological events such as human migration, animal domestication, the evolution of various species, and the dispersal and extinction of past species. It started in 1984 with a single 229 base pairs (bp) long sequence from the quagga, an extinct member of the horse family [1]. One year later, the first human genetic material from a mummified Egyptian individual was published [2]. This publication was the "kick-off" of this new field of research, and numerous publications appeared with high expectations about extracting DNA from such ancient samples. However, results such as the retrieval of DNA from dinosaurs [3] or preserved in amber [4–7] turned out to be contamination from modern and microbial DNA [8–10]. These data were disregarded because on one hand DNA molecules cannot survive from several millions of years ago and, on the other hand, sequencing technologies and analyses methods were at the time not sufficiently developed to deal with the specific properties of aDNA. The polymerase chain reaction (PCR) method, traditionally used in aDNA research, could amplify a limited number of specific DNA targets simultaneously when using multiplex assays. Additionally, only DNA fragments matching the specific targets could be identified. As a result, damaged DNA fragments, which are crucial for the authentication of aDNA, were lost, and the ancient origin of the mapped reads could not be verified. Since then, the field was instantly growing, side by side with the development of new sequencing technologies and DNA enrichment approaches. With the introduction of the next-generation sequencing (NGS) technology, it became possible to combine amplification and sequencing of up to several billions of individual DNA library templates at a time. Most importantly, this technology also can sequence short and damaged DNA molecules, which is well suited to the characteristics of aDNA. Lastly, with every year the cost of producing NGS data decreases and becomes more reasonable.

Together with the development of new software and precautions in the laboratory workflow, such as a sterile work environment, surface cleaning, and the sample treatment with UV radiation, intense work on our species' past and their migration and admixture began in earnest during the past two decades [11–14]. Besides anatomically modern humans, the genomes of archaic hominids including Neanderthals and Denisovans [15–21] have been successfully sequenced. The analysis of human DNA usually follows a standard pipeline [22], for which established software exists [23]. The reconstructed genomes can then be compared with a relevant data set of other modern and ancient (human) genomes.

Due to the improvements in sequencing technologies and bioinformatics analysis methods, the research field of DNA analysis also extended into entire microbial communities and specific pathogens. One of the most significant projects in this area is the human microbiome project [24] to understand the full range of human genetic and physiological diversity. This analytical approach has also already reached the field of aDNA, and projects investigating the ancient human oral or gut microbiome have been carried out [25, 26]. Extensive work has also been conducted to understand the evolution of various pathogens, their genomic structure, and their links with the host. One example concerns research on *Mycobacterium leprae*, the causative agent of leprosy. In 2013, the first reconstructed ancient genomes from victims of the leprosy epidemic in the middle ages were published [27]. With a follow-up study in 2018 [28], Schuenemann and colleagues expanded their data set of ancient genomes. Also, they confirmed their initial findings that the genome of *M. leprae* is relatively unchanged over time. However, the origin of this pathogen still remains unresolved, and more ancient strains from other regions of the world are needed to shed more light on this aspect.

Bacterial pathogens were not the only focus of ancient DNA research, but also viruses, including the hepatitis B virus (HBV) [29–31] and Variola virus - the causative agent of smallpox [32, 33]. Analysis of these pathogens is challenging in part because of the recombinant components of the genomes [34–37]. While these regions are already well-handled in other species, for example, the bacterium *Treponema pallidum* [38], there is generally no accepted method to address these regions in the analysis of HBV.

1.1 Contributions of this thesis

This work is based on the research conducted by the Group for Integrative Transcriptomics and the Group for Paleogenetics at the University of Tübingen (Germany), and the Group for Paleogenetics at the University of Zürich (Switzerland). It was carried out in collaboration with the Max Planck Institute for the Science of Human History in Jena (Germany) and other international research groups. It focuses on the analysis of next-generation sequencing data from ancient DNA samples to provide and improve the methodology for aDNA analyses for both human and microbial genomes. By increasing the existing data set and performing in-depth studies with adapted methods, the goal is to better understand various organisms such as humans, bacteria, and viruses. The prerequisite for this research field is reliable data obtained from samples from multiple sources. When reconstructing the genomes of (ancient) DNA samples of either single or many organisms, as it is possible from a metagenomic sample, every step needs computational support.

1.1.1 DamageProfiler

Although laboratory methods and sequencing technologies continue to improve, it is still necessary to investigate and verify the ancient origin of the obtained DNA fragments. Methods developed for this purpose use the characteristic features of aDNA (Section 2.4) to distinguish between modern and ancient DNA. This is, among others, based on post-mortem damage that is characteristic for ancient DNA and leads to specific damage patterns [39–41]. Existing methods [42, 43] for evaluating these patterns are inefficient and not user-friendly. They also rely on different software and R packages, which makes maintenance difficult. To efficiently calculate damage patterns in next-generation sequencing data, this thesis presents DamageProfiler. While methods exist for either the analysis of single-organism mapping files or the analysis of metagenomic samples [42, 43], DamageProfiler provides a valuable feature by combining these two cases of application and address the lack of user-friendliness and efficiency.

1.1.2 mitoBench

With the advent of ancient DNA research, the study of human evolution has gained intensity, and genetic variation in mitochondrial DNA (mtDNA) is an established target for analysis. Although human nuclear genomes have become more accessible through modern sequencing technologies, the use of mtDNA in population genetics remains highly informative due to its small genome size and easier availability. For instance, mtDNA has been used to reconstruct demographic events that occurred in Europe during the period before the Last Glacial Maximum (LGM) and after the LGM [44, 45], or to trace demographic changes that have shaped mtDNA variation in past and present populations [46–48]. The array of mtDNA analysis tools usually are based on different file formats and often require manual intervention for downstream analysis steps, which can be quite cumbersome and lead to an increased risk of error. This issue also could not be solved with analysis tools such as Arlequin [49] or the newer MitoSuite [50]. In addition, the collection of a comparative data set is still challenging. Even though ancient and modern mtDNA databases are currently available, e.g. EMPOP [51], MitoMap [52], HmtDB [53], mtDB [54], and AmtDB [55], there is no database available that provides modern and ancient data and especially their accompanying meta-information. Moreover, they typically include only partially overlapping subsets of all available samples, sometimes with varying degrees of missing information. In addition, the growing number of complete mitogenomes makes it difficult to keep up with the available information. To address these issues, we present mitoBench with its main components: the workbench and the database. With the workbench, a graphical user interface, the user can perform basic human population genetic analyses and visualize the results on a selected data set. The database provides a unique opportunity to easily assemble a well-curated, high-quality comparative data set of modern and ancient mtDNA genomes.

1.1.3 Pathogen evolution

In addition to humans, uncovering the origin and evolution of pathogens have also boomed with the study of ancient DNA, but in many cases these questions remain unresolved. Despite such advances in the field of ancient DNA, work with difficult material, such as mummified tissue from ancient Egyptians, remains challenging. Such materials are fraught with questions regarding their overall DNA preservation due to the hot climate and high humidity in many tombs, as is the case in Egypt, and the chemicals used in embalming techniques, as well as questions regarding possible contamination of the DNA, recovered [56, 57]. In addition, methods used for modern DNA have limited applicability due to low coverage or damage to the DNA, and thus, must be newly developed or adapted. This work helps to shed light on questions about the origin, diversity, and genomic structure of HBV, Variola virus, and *M. leprae*. Moreover, it is the first metagenomic investigation of various mummified tissue from ancient Egypt and biological material fixed in ethanol and embedded in paraffin.

1.2 Outline

This dissertation is divided into six chapters. Following the introduction in Chapter 1, a biological and computational background is given in Chapter 2, providing a basic understanding of the topics covered in this dissertation. Chapter 3 introduces DamageProfiler, a software to determine damage patterns in high-throughput next-generation sequencing data. These damage patterns are essential to verify the ancient origin of aDNA fragments. The steadily increasing amount of sequencing data requires the need for a more efficient, intuitive, and widely applicable software. In Chapter 4, mitoBench is described, a user-friendly workbench to investigate modern and ancient human mtDNA. The workbench is connected to a comprehensive database of several thousands of complete modern and ancient human mitochondrial genomes, which can be loaded into the workbench and used as a comparison data set for various types of analysis. This part is followed by two ancient DNA studies in Chapter 5, describing the investigation of the metagenomic content of mummified Egyptian individuals, as well as the genome reconstruction of three pathogens: *M. leprae*, HBV, and Variola virus. The thesis concludes with Chapter 6, in which the software development and analysis work is summarized and discussed with an outlook on potential improvements of the introduced tools and methods.

CHAPTER 2

Background

This chapter provides the biological and computational basis for a better understanding of this work. The first part will introduce the reader to the history of sequencing technologies focusing on next-generation sequencing, which facilitated a major advancement in ancient DNA research. Then, the genetic variation at the DNA level is presented, followed by introducing mitochondrial DNA. Next, an overview of the specific characteristics of ancient DNA is provided, laying the groundwork to understand the facilities and importance of differentiating ancient and modern DNA as described later. Finally, the chapter closes with a presentation of the methods used in this thesis to analyze and reconstruct metagenomic communities and single genomes and highlight the bioinformatic challenges in this field.

2.1 DNA sequencing

The research field of DNA sequencing was built upon the initial developments after the first determination of the DNA structure by Watson and Crick in 1953 [58]. Twelve years later, the group led by Robert Holley was able to determine the order of the nucleotide acids for the first time [59]. The breakthrough of DNA sequencing followed in 1977 with so-called 'Sanger Sequencing' [60]. This was the beginning of automated sequencing, which was the predominant technique for the following three decades. It led to the sequencing of even larger genomes, culminating with the Human Genome Project [61, 62], which is still the world's largest biological collaboration project and was completed in 13 years at the cost of nearly $3 billion. This opened the door for large-scale sequencing projects to investigate human sequence variation; however, Sanger sequencing was too labor-intensive,

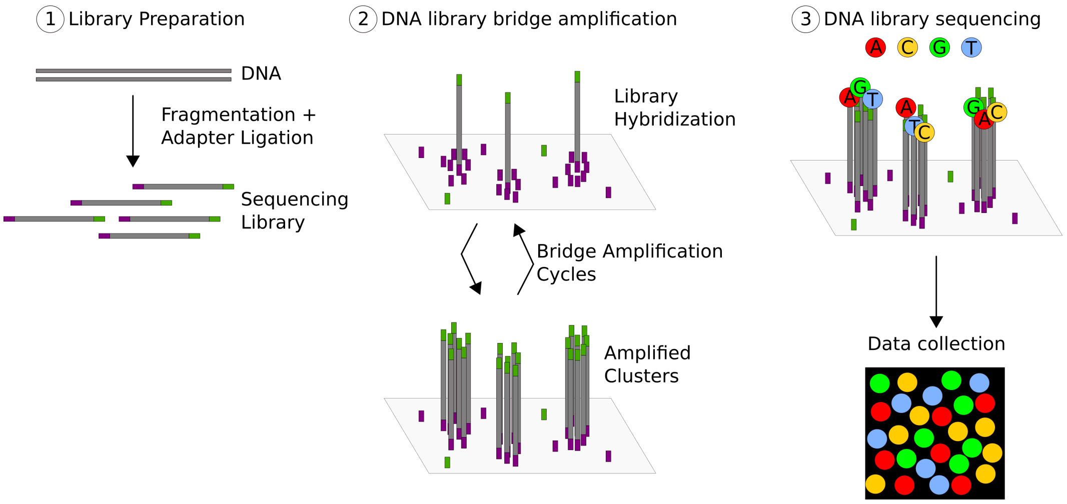


Figure 2.1: Main steps of Illumina Sequencing: (1) Library preparation, (2) DNA library bridge amplification, and (3) DNA library sequencing applying Sequencing-by-Synthesis (SBS). The figure is adapted from [74].

time-consuming, and expensive for these types of projects. In 2004, the National Human Genome Research Institute started a program to reduce the cost of sequencing the entire genome to $1000 in 10 years [63]. In this setting, massively parallel sequencing was introduced and became known as next-generation sequencing (NGS) technology. These techniques mostly employ sequencing by synthesis (SBS), which combines cycles of biochemistry and imaging. For SBS, three main strategies were developed. The pyrosequencing approach involves the discrete, stepwise addition of each deoxynucleotide (dNTP), which releases pyrophosphate that is tracked with a luciferase assay [64]. A second strategy uses the specificity of DNA ligases to attach fluorescent oligonucleotides to templates in a sequence-dependent manner [65, 66]. Eventually, however, a third approach was established involving stepwise, polymerase-mediated inclusion of fluorescent-tagged deoxynucleotides [67, 68].

The sequencing technologies developed further, and third-generation methods, such as PacBio SMRT [69] and Oxford Nanopore [70], are progressing rapidly. These methods are specialized in producing longer reads [71]. However, these methods still have a much higher error rate compared to NGS [72, 73], which can negatively impact downstream processing, such as genome reconstruction. For this reason and due to the short length of ancient DNA molecules, this work focuses on Illumina's state-of-the-art sequencing methodology, which is currently the favored method in aDNA research.

Illumina sequencing can be divided into three main components: (1) the library preparation, (2) DNA library bridge amplification, and (3) DNA sequencing, as illustrated in Figure 2.1.

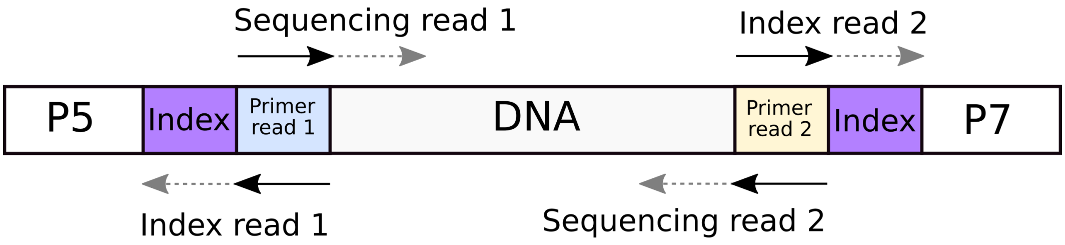


Figure 2.2: Schematic overview of fully realized library molecule indicating the sequencing direction for reads 1 and 2 and the index reads 1 and 2; P5 and P7 are the anchor sequences during read 1 and read 2, respectively.

Sequencing library preparation

Following DNA extraction, DNA molecules are converted into DNA libraries. Usually, DNA molecules have to be broken down into smaller fragments beforehand, but due to the high degree of postmortem fragmentation, this is not necessary for aDNA [75]. Two common ways of preparation are the double-stranded (dsLib) [75] and the single-stranded (ssLib) [76] library protocol. While the former has already been in use for many years, the latter was introduced in 2013 and promised to perform better on highly degraded DNA. Each strand of the fragment is considered independently. Since the ssLib protocol is not applicable for all aDNA projects due to high costs and lower throughput, the dsLib protocol was established as the standard method. With improvements to the ssLib protocol in 2017 and 2020 [77, 78], it became viable for more aDNA research groups.

The following section will focus the dsLib protocol as the projects discussed here only employed this method. Using this protocol, DNA sequencing adapters are attached to both ends of the DNA fragment to build a library for NGS sequencing. A fully realized adapter sequence contains several specialized parts (Figure 2.2 for a schematic representation): an anchor sequence, with which the fragment attaches to the flow cell; a primer sequence, which functions as the binding site for the sequencing primer and the optional index; an optional index sequence used for bioinformatic separation of multiplexed samples after sequencing; and an optional second primer sequence for a possible second read and index.

The assignment of a second primer and index also reduces costs, as multiple libraries can be sequenced together. A unique index combination is used per library, which helps avoid cross-contamination with other samples that are simultaneously sequenced. Only correct and fully formed molecules are processed further.

Bridge amplification and sequencing

Illumina sequencing is also described as sequencing by synthesis (SBS), of which the main steps are illustrated in Figure 2.1. Once the molecule is linked to the flow cell via complementary sequences, a bridge amplification process amplifies each fragment and produces clonal clusters. These are necessary to increase the

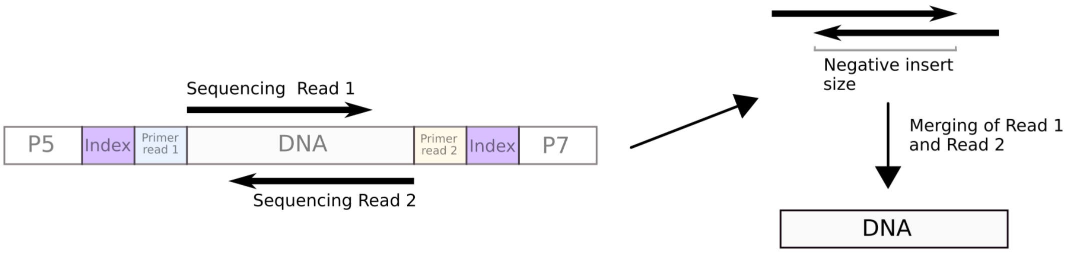


Figure 2.3: Paired-end sequencing resulting in a forward and reverse read. Due to the short length of the DNA fragments, the reads can be merged based on the negative insert site. This results in a longer DNA fragment with overall higher quality.

fluorescent signal in the latter sequencing reaction. SBS itself begins with the extension of the first sequencing primer to produce the first read. With each cycle, four tagged nucleotides compete for the addition to the growing chain. Only one is incorporated, based on the sequence of the template. After adding the individual nucleotides, the clusters are stimulated by a light source and a characteristic fluorescence signal is emitted. The emission wavelength, along with the signal intensity, determines the base call. Hundreds of millions of clusters are sequenced in a massively parallel process, and state-of-the-art sequencing machines, e.g., Illumina's NovaSeq6000 sequencing system, can produce up to 6TB of 2×250bp read length in less than two days [79].

Paired-end & single-end Sequencing

Illumina sequencing can be distinguished in single-end (SE) and paired-end (PE) sequencing, which can be used independently of the library preparation protocol. While SE sequencing reads the DNA fragment from only one end to the other, PE sequencing sequences the fragment from both ends. If paired-end sequencing is applied to short fragments, it results in a negative insert size, meaning that there is an overlap between two sequencing reads originating from a single DNA fragment (Figure 2.3). For aDNA analysis, this is an advantage; read 1 and read 2 are merged based on the overlap using adapted methods that construct a consensus call by taking the base with higher quality, thus improving the overall quality of the data set. PE sequencing also provides advantages with modern applications, as it enables the detection of genomic rearrangements and repetitive regions within a genome, as well as the ability to detect insertion-deletion variants [80].

2.2 Genetic variation

Genetic variation is defined as the differences in DNA sequences between individuals or populations of individuals. These differences are the key to insights into the genetic history of the human species. Variation is caused by multiple sources,

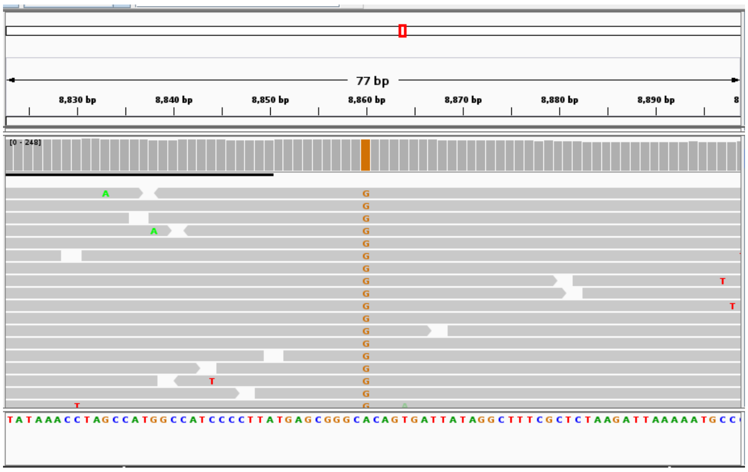

77 bp
8,830 bp
8,840 bp
8,850 bp
8,860 bp
8,870 bp
8,880 bp
8,890 bp
[0 - 248]
TATAAACCTAGCCATGGCCATCCCCTTATGAGCGGGCACAGTGATTATAGGCTTTCGCTCTAAGATTAAAAATGCC

Figure 2.4: Screenshot of IGV [81], showing reads mapping to a reference genome with a SNP call at position 8860. While the reference is showing a 'A', all fragments mapping to the reference have a 'G' at that position.

including mutations and genetic recombination. With advances of the NGS sequencing technologies (Section 2.1), studies on a population-scale have become possible.

2.2.1 Single nucleotide polymorphisms

Of particular interest to population, genetics are single nucleotide polymorphisms (SNPs), variants at a single base pair position. SNPs play an essential role in explaining disease susceptibility and phenotypic differences between populations and represent the only visible genetic marker. Therefore, SNPs are often inspected by hand with the help of appropriate software, such as the Integrative Genomics Viewer (IGV) [81] shown in Figure 2.4. SNPs are evolutionary variations and accumulate over time. The differences of individual genomes and groups of individuals are therefore defined by their set of shared and unique SNPs. They can be used to determine the placement of newly sequenced genomes within the known diversity of already sequenced genomes.

For diploid organisms, combinations of base pairs can be defined at any genetic locus. These combinations are called genotypes. Combinations of genetic variants located on the same chromosome are called haplotypes. A haplotype can be specific to an individual, a population, or even a species and plays an important role in disease and population genetics. For example, sickle-cell anemia, β-thalassemia, and cystic fibrosis result from SNPs [82–84]. Another example is the International HapMap Project, where individual combinations of SNPs are used as genetic markers [85].

2.2.2 Insertions and deletions

Another form of genetic variation is insertions and deletions (INDELs), which are typically measuring from 1 to 10,000 base pairs in length [86, 87]. Large-scale changes in DNA are chromosomal inversions and translocations. An inversion is a chromosome rearrangement, where a part of the chromosome is inverted end-to-end. In contrast, translocations are a phenomenon that results in an unusual rearrangement of chromosomes. A particular case of translocations are NUMTs (nuclear mitochondrial DNA segments) [88], which are limited to the human genome and are segments of the mtDNA that can be found in the nuclear genome, for example, on chromosome 1. One of the challenging areas of molecular biology and bioinformatics is the detection of variation. This is further complicated by false-positive variants, caused e.g. by sequencing errors or bases that are chemically modified, as is often the case for aDNA (Section 2.4) [41, 89]. Especially the analysis of DNA modifications resulting in more extensive structural changes is limited by short-read sequencing.

2.2.3 Genetic Recombination

The exchange of genetic material between different organisms is called recombination [90], which leads to new gene and feature combinations. While mtDNA is not affected by recombination (Section 2.6), it can be observed in nuclear DNA, e.g. driven by sexual reproduction [91], and abundant bacteria and virus species, such as the *Treponema pallidum* bacterium [92, 93] and hepatitis B virus [34].

2.3 Measures and visualizations of genetic distance

When studying species and populations, it automatically leads to the question: how similar or different is the newly reconstructed genome to other strains of the same or closely related species, and how does it behave on a population scale? This can be answered by using several methods to measure the distances within and between populations and can be used to investigate humans, as well as bacterial and viral species. The following paragraphs focus on the most commonly used methods in population genetics and are also applied during the analysis of the projects presented in this work.

2.3.1 F_{st} value

There are two ways of measuring genetic distances: the distance within and between populations [94]. The distance within a population describes the genetic diversity H, which defines the degree of relatedness of individuals within a population. With this, the distance is expressed by the number of bases that are different between the DNA sequences of the considered individuals. The distance between populations is also derived from the genetic diversity H and is based on the partitioning of the total genetic variance into components within and between populations. Formally, the total genetic diversity is defined as H_T, which assumes that all individuals considered are from the same population. The average genetic diversity within groups is denoted as H_S and obtained by calculating H for each group individually and taking the average value. Taking H_T and H_S together by subtracting the average within-group diversity from the total diversity, then dividing by the total diversity (Formula 2.1), resulting in a measure of genetic distance, called F$_{st}$.

$$F_{st} = \frac{H_T - H_S}{H_T} \tag{2.1}$$

This value can range from 0 to 1, where 0 means that the populations are identical, and 1 indicates that they do not share any genetic material. The F_{st}value is usually visualized by a heatmap, where each row and column represents one population. The genetic distance between the populations is highlighted in colors, where white represents F_{st}=0 (genetically identical) and dark blue F_{st}=1 (no genetic material shared). However, the display of a distance matrix quickly reaches gargantuan proportions (e.g., for 50 populations, there would be 1225 distance values). While probably only a few people can see any pattern in a large matrix, most need a different kind of visualization - the following presents two frequently used possibilities: Principal component analysis plots and trees.

2.3.2 Principal component analysis

A similar approach follows the principal component analysis (PCA). The PCA is a way to extract strong patterns from large and complex data sets. The essence of the data is captured in a few principal components, which convey most of the variation in the data set. The PCA reduces the number of dimensions without selecting or discarding them. Instead, it constructs principal components that focus on variation and account for the multiple influences of the dimensions. Such influences can be traced back from the PCA plot to determine what generates the differences between clusters. For mtDNA, a PCA represents the data points and the haplogroups that explain the composition of the considered samples best. Besides a distance matrix, a PCA can be generated from a matrix of gene frequencies, depending on the input data set.

2.4 Characteristics of ancient DNA

Besides the DNA recovered from modern individuals, it is also possible to retrieve DNA from fossils or other ancient material because DNA can survive in dead organisms under favorable conditions for thousands of years. The preservation of the DNA can be influenced, e.g., by temperature, humidity, or pH value [89, 95, 96]. While many parts of DNA analysis are the same for ancient and modern DNA, there are some aspects that need to be considered and also require adaptation to existing methods and the development of new software. The aDNA research is faced with different challenges in comparison to modern DNA due to specific characteristics, which can also be used to authenticate the ancient origin of DNA fragments and are elaborated in this section.

Living organisms keep their DNA intact using different repair mechanisms [39]. Immediately after death, these cease to work, and the DNA starts to degrade. Ideal conditions can slow down the degradation. However, it can not be stopped completely. This form of natural degradation breaks down the DNA fragments into tiny pieces [97], resulting in an average fragment length of 40-60bp [57]. Not only does the size of the DNA fragments decreases, but also the amount of DNA. The endogenous DNA, meaning the DNA that belongs to the species of interest, is usually very low, often less than 3% of the total DNA contained in a sample [98, 99]. As a result, most of the DNA belongs to other sources, mainly environmental DNA (e.g., DNA from bacteria, viruses, fungi, etc.) or contamination from modern human DNA introduced by people handling the sample during excavation, archival, or laboratory work.

In addition to the decreasing quantity and fragment length, ancient DNA is also affected by chemical damage that changes the shape of the DNA [41]. Hydrolytic attacks on the phosphate backbone of the molecule, e.g., destabilizes the DNA strand [100]. Other effects mostly lead to gaps between subsequent bases in the DNA strand, double-strand breaks, and cross-links [39, 40]. However, the most common damage has the effect of cytosine (C) to uracil (U) base substitution (Figure 2.5), which predominately affects the single-stranded overhanging ends of the fragmented double-stranded DNA and can be observed to varying degrees. As uracil is read as thymine (T) during DNA replication, this results in cytosine to thymine base misincorporations in the sequenced aDNA fragments. This base substitution can be observed with increasing frequency towards the 5'end of the fragment and accumulates over time. The complementary base substitution of guanine (G) by adenine (A) can be seen at the 3'end.

2.5 Analysis of (ancient) NGS data

The analysis of ancient DNA differs from the analysis of modern DNA due to the post-mortem DNA modifications, resulting in highly fragmented and damaged

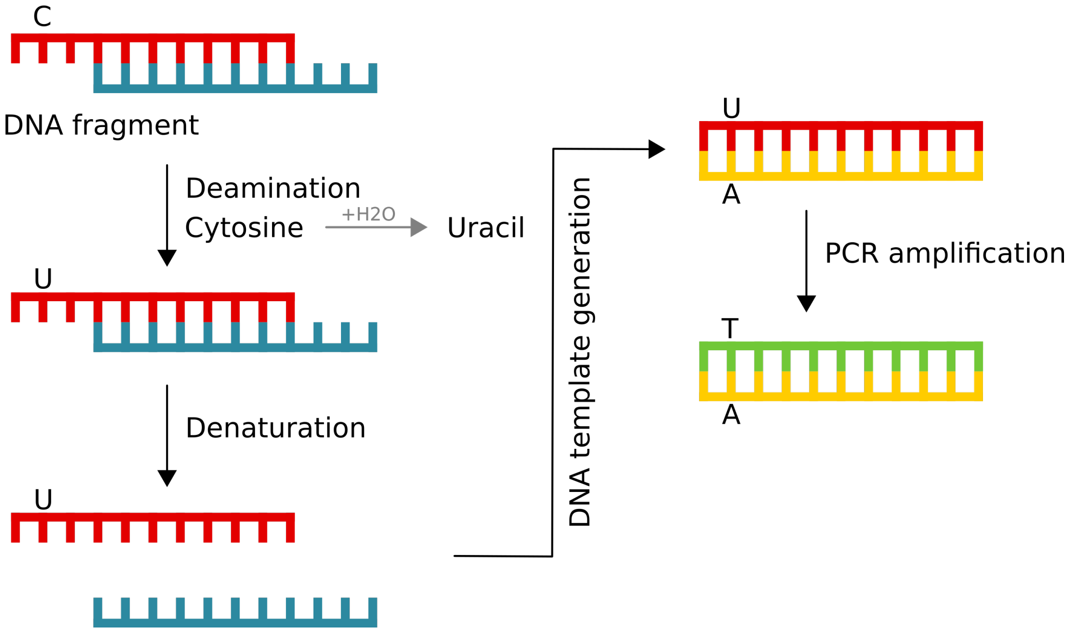


Figure 2.5: Conversion process of cytosine to uracil. Cytosine becomes uracil by methylation, which is read off as adenine during DNA template generation. In PCR amplification, the adenine is complemented by thymine.

DNA. Specific methods have been developed, and parameters were adapted to perform metagenomic analysis, ancient genome reconstruction, and the authentication of ancient reads. The methods that were used in this work are presented in the following paragraphs.

2.5.1 Metagenomic analysis

The metagenomic analysis describes the assessment of all genetic material in a sample, regardless of its origin. This analysis is based on DNA fragments from next-generation sequencing, which undergo several preprocessing steps. After adapter removal, all reads are filtered for length and, if paired-end sequencing was done, the reads are merged. Then, the resulting reads are trimmed to a certain quality. Subsequently, the resulting DNA fragments are mapped against a reference database, set up by the user, and consist of various species. Due to the increasing amount of sequencing data, even for old samples, and many available reference genomes, the mapping process is very challenging. To solve these problems efficiently, the software MALT (MEGAN alignment tool) [101] was developed. By adapting the mapping parameters, it can also be used for ancient DNA.

Based on the taxonomic assignment, it is possible to determine the microbial communities within the samples by comparing them with already known communities. Available methods, e.g., SourceTracker2 [102], compare the microbial composition of the input sample to a set of reference microbial communities and determine their percentage. Unclassified species are marked as 'unknown'.

2.5.2 Pipeline for ancient genome reconstruction

Another focus of aDNA research is the reconstruction of a specific genomes and usually includes quality assessment, read preprocessing, mapping, and genotyping. Since the execution of the individual steps by hand is very error-prone, time-consuming, and challenging to reproduce, the EAGER pipeline [23] provides a fully automated way of mapping reads from high-throughput sequencing against a specific reference genome. EAGER is optimized for, but not limited to, the reconstruction of ancient genomes and therefore allows the preparation of ancient genomes and the modern comparative data set simultaneously. It follows the GATK (Genome analysis toolkit) best practice guidelines [103] and the parameters tested best for ancient DNA are set as default [22, 104].

The pipeline consists of three components preprocessing, read mapping, and genotyping. The preprocessing step is based on the raw sequencing data and includes quality trimming, read merging, and length filtering. Then, the resulting reads are mapped against a reference genome, taking into account various parameters such as mapping quality and several mismatches. After assessing statistics to verify the ancient origin of the reads, variants can be called to be later able to reconstruct the genome. Moreover, the variants can be categorized according to their effects, and coding effects can be predicted based on the genomic locations.

2.6 Human population genetics using mitochondrial DNA

Besides the nuclear genome, all species having mitochondria also have an additional genome present in their cells: mitochondrial DNA [105]. MtDNA is a double-stranded, circular, closed covalent molecule of 16569bp that has been completely sequenced [106]. Due to its small size, maternal inheritance [107], high copy number [108, 109], high mutation rate [110], and lack of recombination and population-level variability [111, 112], mtDNA is a popular resource in population genetics given its ability to facilitate analyses. Also, the use of mtDNA is not limited to humans but can also be applied to other species [113–115]. One feature of mtDNA is the classification into haplogroups, which are defined by a group of haplotypes (Section 2.2) shared by individuals. However, this division is based on known variation. If new variants are found, this can potentially change the mtDNA tree or confuse naming and ordering.

In this work, the focus is on human mtDNA and its analysis. Since only mtDNA was affordably retrievable from ancient samples for a long time, many studies on human population genetics are based on it [46, 116–118], and still are [119–121]. Among others, these studies found that haplogroups, which are

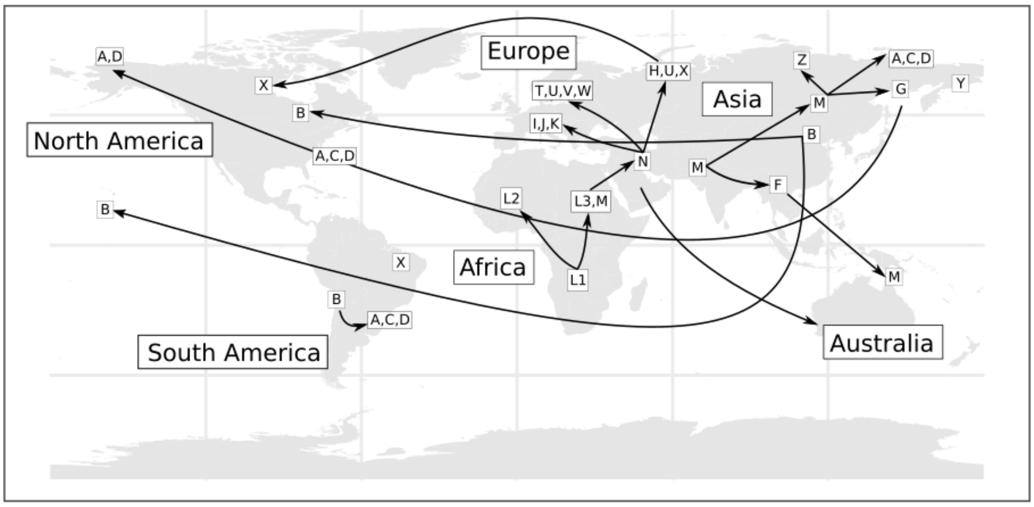


Figure 2.6: A haplotype map showing the migration of haplogroups worldwide (adapted from `http://www.gbpec.ac.in/research/HgsDb/mtdna.html`).

defined by the genetic variation of mtDNA, can be assigned to specific regions of the world. Based on human mitochondrial haplogroup research, maps have been created that show the distribution of various haplogroups throughout the world (Figure 2.6). These maps contain information about the order in which SNPs have appeared over time and provide a variety of interesting information, such as the ethnic composition of a population and the history of past human migrations to different parts of the world. The groups are named from A to Z in the order of their discovery, and subgroups are defined by adding numbers and letters (e.g., N1a is a sub-haplogroup of N1, which belongs to haplogroup N). All information is collected and displayed in PhyloTree [122], which is the only systematical collection of haplogroups publicly available to date. The root of this tree represents the so-called *Mitochondrial Eve*, which stands for the matrilineal most recent common ancestor (MRCA) of all currently living humans [123–125]. In other words, the *Mitochondrial Eve* defined the most recent woman, from whom all living humans originate in an unbroken line solely through their mothers and the mothers of those mothers, going back until all lines converge.

2.7 Ancient pathogen genomics

Ancient pathogen genomics is a scientific field concerned with the study of pathogen genomes. These are typically obtained from ancient human, plant, or animal remains. Ancient pathogens are microorganisms, some of which are now extinct, that have caused several epidemics and deaths worldwide in past centuries. Typical sources of ancient DNA are the remains (bones and/or teeth) of victims of the pandemics caused by these pathogens.

With the development of NGS and appropriate computational methods, the analysis of ancient pathogen genomes became possible. The process begins with the extraction of DNA from ancient samples, followed by NGS library construction and (if necessary) subsequent capture-based screening (Section 2.1). Computational tools are used to map NGS-derived reads against a single- or multi-genome reference. Alternatively, metagenomic profiling or taxonomic assignment methods from shotgun NGS reads can be used (Section 2.5).

By analyzing ancient pathogen genomes, researchers can understand the evolution of modern microbial strains that can potentially cause new pandemics, or outbreaks [126]. Analysis of aDNA is performed using bioinformatics tools and molecular biology techniques to compare ancient pathogens with modern descendants, providing phylogenetic information about these strains. This provides a detailed look at microbial evolution, with the goal of better understanding the interactions between pathogens and their hosts on an evolutionary time scale. It also uncovers the origins of various pathogens and unravels the genetic processes involved in their epidemic emergence in human populations. It has succeeded in reliably identifying the pathogens of past pandemics. The first such genome, published in 2011, was that of the notorious bacterial pathogen of *Yersinia pestis*, the causative agent of plague [127]. Subsequent analyzes achieved a more precise determination of its genome, and further insights into the evolutionary history [128, 129]. Other examples include *M. leprae* [27, 28], HBV [29–31], and Variola virus [32, 33, 130, 131], which are presented in more detail in Chapter 5.

The analysis of sequence data obtained by NGS is based on the same computational approaches used for modern DNA (Section 2.5), with some special features. For example, damage patterns characteristic of ancient DNA must be assessed and accounted for (Section 2.4). Other genomic features complicate further analysis. In the case of HBV or *Treponema pallidum*, genetic recombination [34] occurs. This interferes with phylogenetic analysis by leading to incorrect phylogeny, thus interfering with the analysis of the evolution of the pathogen. While this problem can be addressed, for example for *Treponema pallidum* [38, 132], there is no solution yet for HBV and further research is needed. Another problem is the diversity of genomic strains available. To construct a complete picture of the evolution and spread of a pathogen, strains from different ages and geographical regions are needed.

CHAPTER 3

DamageProfiler: Fast damage pattern calculation for ancient DNA

Text and figures in this chapter were adapted with modifications from our work published in Bioinformatics [133].

3.1 Introduction

The use of damage patterns in ancient DNA analysis is still the most applied method to verify the ancient origin of the mapped DNA fragments. Whit these patterns, it can be distinguished between modern and ancient DNA, e.g., to identify modern human DNA contamination. The patterns rely on specific aDNA characteristics, which are a short fragment length [57], preferential strand breaks at purine bases [41], and increasing frequency of cytosine (C) to thymine (T) base substitutions towards the end of the fragment [41] (Section 2.4). Besides accessing the content of modern human DNA in ancient samples, for which a variety of appropriate tools already exists such as schmutzi [134], contaMix [118], DICE [135], and ANGSD [136], the authentication of sequenced samples is a central step in the aDNA analysis and still offers potential for improvement in terms of efficiency and usability. Methods such as mapDamage2 [42] and PMDTools [43] are typically used to assess the damage patterns in next-generation sequencing data. Due to advances in sequencing technology, the number of sequenced DNA increases, and the costs reduce. This also opens up DNA analysis to less experienced users and requires user-friendly and intuitive software. Furthermore, the methods need to be efficient in order to process large mapping files. Unfortunately, the tools available are only command-line-based and inefficient or not applicable when large quantities of data need to be analyzed.

While the focus of aDNA research was limited to single genomes for many years, new techniques and methodologies have made it also possible to now study the genomic entirety of an ancient sample, e.g., a specific body part or environment [25, 26, 137, 138]. Consequently, there is a need for suitable software that can assess the damage patterns of several species simultaneously and provide detailed results for each species and suitable overview files. This is one of the targets of the existing software HOPS (Heuristic Operations for Pathogen Screening) [139], which is connected to the metagenomic mapping software MALT [101]. HOPS runs on the MALT output file and calculates specific statistics that can be used to verify the ancient origin of DNA fragments. However, it cannot be executed on mapping files from software other than MALT, which also demonstrates the need for an efficient, independent, and user-friendly software, combining the evaluation of damage patterns from single as well as metagenomic mapping files.

In this thesis, DamageProfiler has been developed to fill these gaps. It evaluates the read length distribution and the accuracy of the mapped reads and the frequency of base substitutions within the assigned reads, both aforementioned characteristic features of aDNA. The results are stored in various file formats to make them easily accessible for further processing. DamageProfiler can process metagenomic mapping files independently of the underlying mapping software and generate damage patterns for a user-defined list of species. The result is a detailed per-species report and an informative overview file with all species. To facilitate the use of DamageProfiler, a graphical user interface is provided in addition to its use as a command-line application to simplify the configuration and to examine the results interactively. This allows even inexperienced users to explore their samples without in-depth computational knowledge easily. In addition, detailed online documentation supports the understanding of the individual functions provided by DamageProfiler. Finally, a log file is automatically generated in which every parameter of the analysis is recorded to reproduce the results later. The following sections describe how the damage patterns are calculated by DamageProfiler, the development (Section 3.3), the testing and deployment (Section 3.4), and the functionality of the resulting application (Section 3.5) in more detail.

3.2 Calculation of damage patterns

Various statistics are calculated to assess the degree of degradation of the mapped next-generation sequencing reads. These include the frequency of base substitutions per position of the DNA fragment, the fragment length distribution, and the Hamming distance between the DNA fragments and the reference genome.

The first mentioned base exchanges occur by chemical modifications postmortem and mainly involve cytosines happening at the 5' and 3' end of the DNA fragment (Section 2.4). The frequency of each type of substitution is determined by comparing the base of each position of the DNA fragment with the corresponding base of the reference. These are stored as a table and then visualized as a line chart, where each type of base substitution corresponds to a line. The most common types of base substitutions are C to T and G to A, which are also specially highlighted in the visualization.

The differences of the DNA fragment compared to the reference genome are calculated using the Hamming distance, which is the number of positions between two strings of the same length where the corresponding symbols are different. The result is visualized as a histogram.

For the length distribution of DNA fragments, each mapped fragment's bases are determined, and the resulting distribution is plotted as a histogram.

3.3 Development

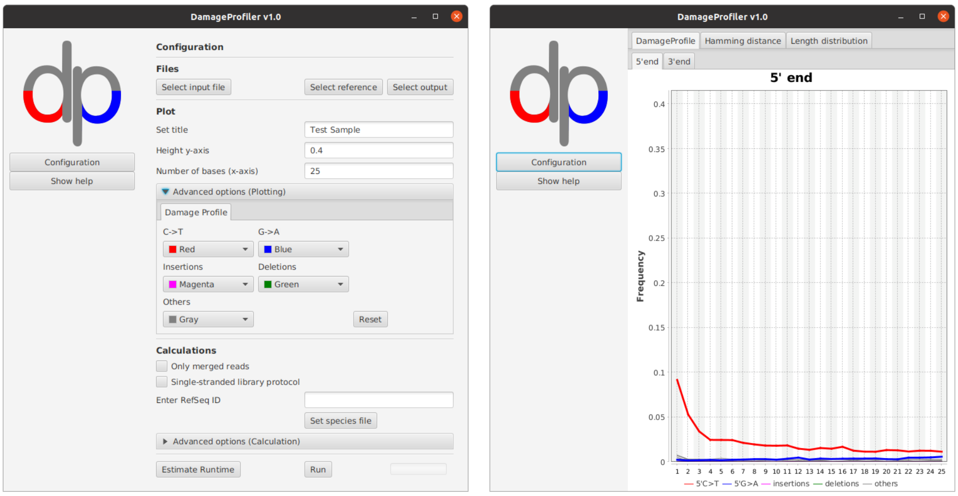
DamageProfiler is written in the Java programming language and the GUI is developed in JavaFX. It has been developed to be platform-agnostic and the source code is freely available on GitHub[1]. The software can be run via a jar file that is available on GitHub and is available as a Bioconda package [140]. All parameters are explained in the online documentation[2] and are briefly described within the help text of the tool itself.

DamageProfiler can be started and configured with the command line or the graphical user interface. While the command line provides the possibility of a fast configuration of all parameters and eases the integration into pipelines and the execution on clusters, e.g., when running more extensive jobs, the GUI allows even inexperienced and computer-illiterate people to use it easily. It is designed to be user-friendly and allows the easy configuration of the run and the visualization and interactive exploration of the results.

The main window of the GUI is divided into two parts: the navigation panel on the left and the main panel on the right side (Figure 3.1a). After starting DamageProfiler, the user is asked to specify the required parameters for input, followed by the possibility of setting additional parameters for the output directory and visualization, such as a title or specific colors. The help page, which can be accessed via the navigation panel, offers more information about each parameter, and a detailed description including examples can be found in the online documentation. After setting all required parameters, the *Run* button will be enabled, through which the analysis can be started. The status of the calculations is visualized with the status bar, next to the *Run* button. After completion, the results are visualized directly in the main window (Figure 3.1b), and the user can navigate through the damage plot, length distribution, and the visualization of the edit distance. The results are also directly saved in various file formats at the specified

[1] https://github.com/Integrative-Transcriptomics/DamageProfiler

[2] https://damageprofiler.readthedocs.io/en/latest/



Set title: Test Sample
Height y-axis 0.4
Number of bases (x-axis): 25

C->T: 🟥 Red
G->A: 🟦 Blue
Insertions: 🟪 Magenta
Deletions: 🟩 Green
Others: ▨ Gray

☐ Only merged reads
☐ Single-stranded library protocol
Enter RefSeq ID:

(a) Configuration of DamageProfiler.

(b) Visualization of the results.

Figure 3.1: GUI of DamageProfiler. The main window is used (a) to configure the run, which can then be started directly from the GUI and (b) to visualize the results.

location. If the sample is metagenomic and multiple species need to be considered, the results are displayed as separate tabs per species.

3.4 Testing & deployment

Software development always confronts the developer with the challenge of creating clean and working code. Various appropriate software and services exist and can be put together according to the needs of the software developer. Especially when developing more complex software and toolboxes, it is crucial to ensure the correct behavior of each part. While unit tests solely validate the functionality of specific methods, it is also necessary to test more complex processes. For this, developers use continuous integration (CI) to ensure consistency by automatically (re)building the software during development and applying integration tests.

The developing process of DamageProfiler is kept simple as shown in Figure 3.2. The elementary methods are tested using JUnit 5[3] to ensure that all functionality is working correctly. Each code change is reassessed using the continuous integration service Travis CI[4]. Only a configuration file needs to be specified. It is then used to build the code and test it automatically after every code change is submitted to the GitHub repository, and feedback is provided immediately. After successful testing using JUnit 5 and Travis CI, an executable file is created using the Gradle[5] build tool. Relying on Gradle ensures that all used software libraries are up-to-date

[3] `https://junit.org/junit5/`
[4] `https://travis-ci.org`
[5] `https://gradle.org/`

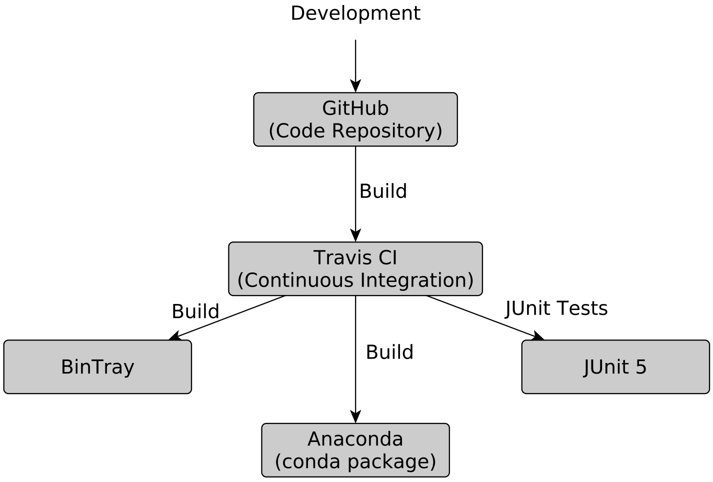


Figure 3.2: DamageProfiler. The code is organized in an online repository. Each push triggers a validation of the code by Travis CI. The software is then built using gradle and is available as anaconda package and at BinTray. JUnit 5 tests are used to verify the functionality of the methods.

and compatible with the source code. Finally, the resulting binaries are uploaded to Bintray[6], a free software distribution service, including a version tag. Bintray also keeps previous versions of DamageProfiler downloadable, which allows the user to (re)run analyses with all so far published versions. In addition, the release of a new version also triggers the automated build of an Anaconda package, which eases the installation process for the user.

3.5 Resulting application

DamageProfiler is a Java application for fast calculation and visualization of damage patterns in ancient DNA, which are typically used to verify the ancient origin of aDNA fragments [15, 96]. Based on the provided input file, the tool calculates the length distribution of all mapped reads, the number of bases by which read and reference differ (by calculating the edit distance), and the damage profile. The latter is the most commonly used feature for authentication of aDNA. It describes the frequency of misincorporation of cytosine to thymine bases per position of the fragment and the resulting complementary misincorporation of guanine to adenine bases. In addition, various statistics, base frequencies, and composition of the sample and reference are provided in text format.

DamageProfiler can be applied to single-reference and metagenomic mapping files (Section 3.5.2). Both can be used with the CLI and GUI, which allows a simple

[6] `https://bintray.com/`

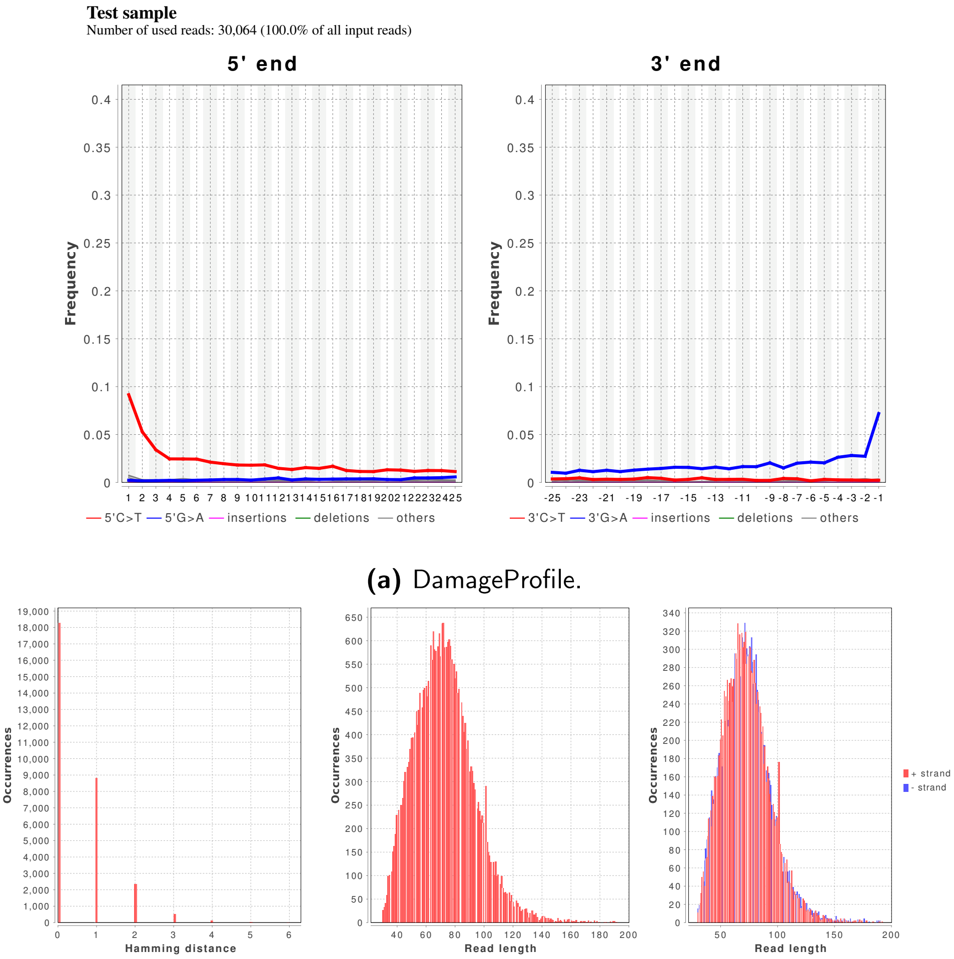
configuration and subsequent interactive exploration of the results. This section provides details about the input and output options and describes all features of DamageProfiler.

3.5.1 Input and output files

DamageProfiler accepts a mapping file in SAM, BAM, or the newer CRAM format, which are the standard mapping file formats. For each record in the mapping file that maps against the reference genome, the reference sequence belonging to the particular fragment is calculated based on the MD tag and CIGAR string of the specific record using the Java library htsjdk [141]. The MD tag encodes mismatched and deleted reference bases, while the CIGAR string is a compressed format of describing an alignment. As a result, the reference file is not required to assess the damage patterns. This is particularly helpful if the input file is a multi-reference mapping file. If the record does not provide the MD and CIGAR information, the reference file must be an additional input parameter for recalculating the values. Based on the mapped DNA fragment information and the corresponding reference sequence, the damage patterns can be calculated and are finally provided in different file formats. This includes PDF and SVG format (Figure 3.3), which simplifies further processing, as well as TXT and JSON for advanced downstream analysis. All values on which the visualizations are based and other statistics, e.g., base frequencies and compositions, are provided as text files. The length distribution and base substitution frequencies are also stored in JSON format, a standard standardized file format, which eases the parsing of the results when integrating the output into an automated report generation tool, such as MultiQC [142]. The final result files can either be inspected directly or within the GUI, which provides an interactive exploration of the results.

3.5.2 Analysis of metagenomic samples

DamageProfiler also accepts multi-reference mapping files, which can be considered a combined single-reference mapping file, such as the case for metagenomic studies. In this case, a list of the species of interest is required and needs to be provided by the user (Figure 3.4a). A complete damage patterns calculation will be conducted for each species in this list, including all output files as described above. The GUI also represents the result of each species in a separate tab, allowing the user to compare and switch between the different damage patterns quickly (Figure 3.4b). In addition, a summary file in PDF format is provided that contains an overview of the damage patterns of all species in the list (Appendix, Figure A.1).

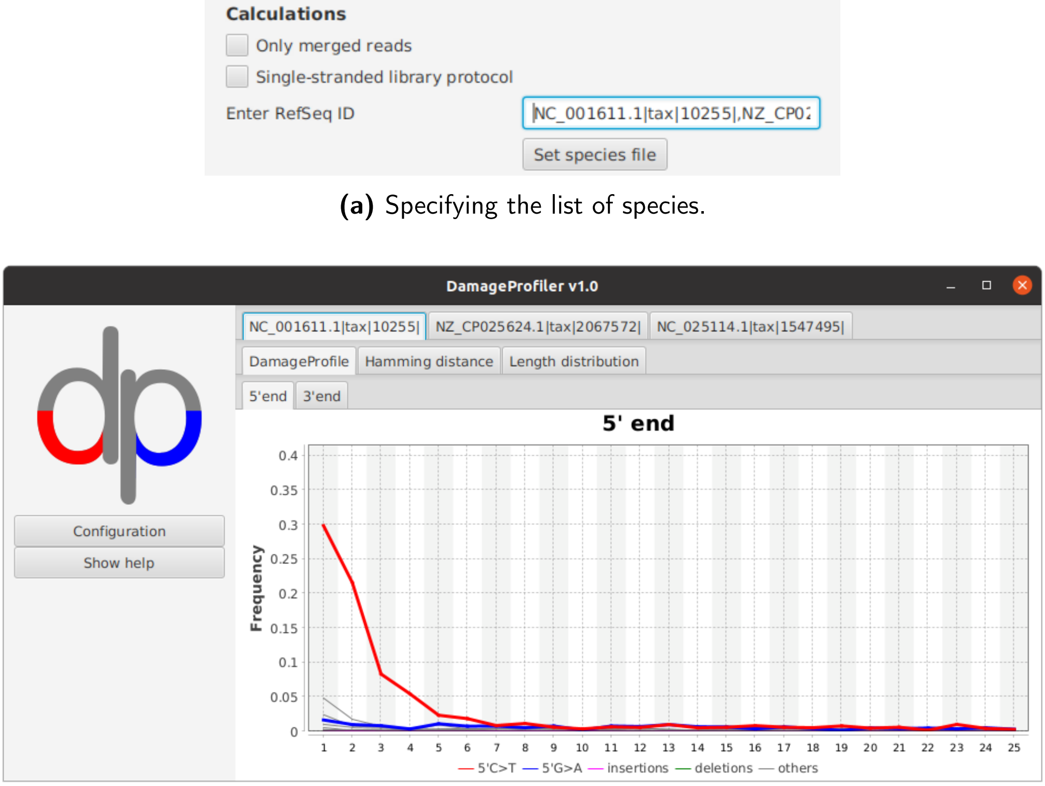


(a) DamageProfile.

(b) Hamming distance.

(c) Read length.

Figure 3.3: Graphical output of DamageProfiler. (a) The frequency of base misincorporations on the DNA fragment. The x-axes are the position in the DNA fragment from the 5' and 3' end, respectively. The frequency is given on the y-axis. (b) Hamming distance between DNA fragment and the reference genome. The x-axis shows the number of bases that differ, the number of DNA fragments is given on the y-axis (c) Read length distribution of all mapped DNA fragments. The length is given on the x-axis, the number of reads on the y-axis. The left part shows all DNA fragments. On the right figure, the strand direction is distinguished.



Calculations

□ Only merged reads

□ Single-stranded library protocol

Enter RefSeq ID: NC_001611.1|tax|10255|,NZ_CP0

(a) Specifying the list of species.

(b) Representation of the result of the metagenomic analysis.

Figure 3.4: Metagenomic analysis with DamageProfiler. (a) The user needs to specify a list of species for which the damage patterns have to be calculated. This can be either entered directly or provided as a text file. (b) The damage patterns of each species are represented in a single tab.

3.5.3 Managing various library preparation protocols

For library preparation, two different protocols exists: the single-stranded (ssLib) [76] and the double-stranded (dsLib) [75] library protocol (Section 2.1). Until recently, the dsLib protocol was commonly used for aDNA analysis, as the ssLib was too expensive to be applied in most ancient DNA labs. Because of the reduction of costs by the use of less expensive chemicals and improvements of the protocol [77, 78], it also became applicable for labs with a more limited budget. As the two protocols work differently, DamageProfiler has been adapted to both methods. In particular, the DNA fragments prepared using the ssLib protocol no longer have the G to A substitutions at the 3' ends. This makes their visualizations redundant and is therefore excluded from the damage plot when selecting the *ssLib*-option.

3.5.4 Runtime improvement

Due to the increasing amount of genomic data, the software must run in a reasonable time. Therefore, this was also one of the main improvements of DamageProfiler compared with existing software. Java libraries providing the required functionality were used whenever possible to achieve an increase in the runtime. For example, DamageProfiler relinquishes the manual reconstruction of the reference sequence of a specific record but uses the htsjdk library [141]. In addition to the time reduction, this also diminishes the actual code and prevents associated errors.

3.5.5 Runtime estimation

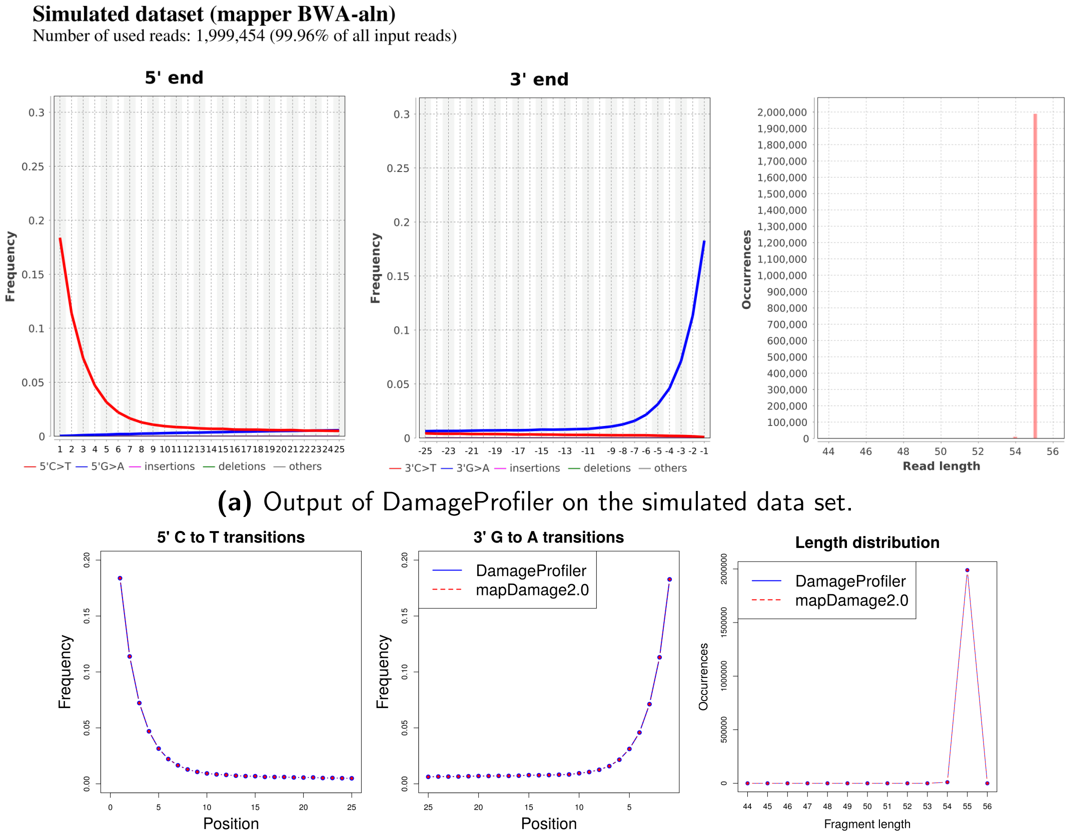
Along with the increasing amount of sequencing data, the size of the BAM files also becomes larger. This also increases the runtime of DamageProfiler (Table 3.1), which makes the GUI unsuitable for large files. DamageProfiler ascertains the size of the input file and the number of read operations before execution and informs the user if the use of the command line version is recommended due to the file size. This message also provides the complete command to run the analysis via the command line.

3.6 Evaluation

3.6.1 Correctness of damage patterns

To ensure the correctness of DamageProfiler, it was evaluated on a simulated data set and a previously published sample [119].

Gargammel [143] was used to simulate an Illumina HiSeq 2500 data set with 1 million fragments and a fragment length of 55bp. A deamination pattern with 18%, 11%, and 7% C-to-T misincorporation frequencies at the first, second, and



(a) Output of DamageProfiler on the simulated data set.

(b) Base misincorporation pattern calculated with mapDamage2 and DamageProfiler.

(c) Length distribution calculated with mapDamage2 and DamageProfiler.

Figure 3.5: Evaluation of the correctness of DamageProfiler. DamageProfiler executed on a simulated data set, resulting in the expected fragment length distribution and base substitution frequency (a). Moreover, DamageProfiler and mapDamage2 [42] were executed on the same data set. The damage profile (b) and the length distribution (c) yield identical results.

third positions at the 5' end was simulated. The G-to-A misincorporation frequency at the 3' end was set identically. Applying DamageProfiler to this simulated data set results in identical misincorporation frequencies (18%, 11%, and 7% misincorporation frequencies) and expected read length of 55bp (Figure 3.5a).

The previously published data set [119] was used as input to DamageProfiler and mapDamage2 [42]. The results were compared and yielded identical damage patterns (Figure 3.5b,c), demonstrating the consistency of DamageProfiler with existing similar software.

3.6.2 Independence of mapping software

To demonstrate that DamageProfiler runs on any mapping file, regardless of the software used to create it, one sample was processed using various mapping software. The commonly used tools Bowtie2 version 2.3.4.1 [144], BWA aln version 0.7.17 [145], BWA mem version 0.7.17 [146], and MiniMap2 [147] were used to map the publicly available sequencing data from an Egyptian mummy (sample JK2134), dated to 776-569 cal BCE [119], against the human mitochondrial genome (RefSeq ID NC_012920.1). For this purpose, the adapter was first removed from the raw data and overlapping paired-end reads were fused. Then, each mapping software was used to align the sequencing reads against the reference genome independently. For BWA aln, the parameters best suited for ancient DNA were used (n=0.01, o=2, l=1024) [22, 104]. Bowtie2 was run with the parameters suggested by EAGER for ancient DNA (–end-to-end, –very-sensitive) [23]. MiniMap2 was run with the option to perform mapping for short reads (-x sr). BWA mem was evaluated with default parameters. If supported by the corresponding software, the MD tag was also calculated. This tag contains information about the mismatching positions of the dataset and eliminates the need for the reference file when calculating the corresponding reference of a dataset. The resulting mapping files were not processed further to allow for an unbiased comparison of each mapping software. A comparison of the results is shown in Figure 3.6. In detail, the frequency of base mismatches, the length distribution, and the edit distance are compared. While the values for the base substitution frequencies vary slightly, the length distribution and the edit distance distribution are stable across all methods.

3.6.3 Runtime evaluation

To evaluated the efficiency of DamageProfiler, the runtime was compared with mapDamage2 [42] and PMDTools [43], the currently most used software for damage pattern calculation. Hence, all tools were executed on five mapping files of different size, host, and target species ([14, 28, 114, 119], Table 3.1). Each software was run ten times under identical computational conditions and the average of all runs gives the resulting runtime. To directly compare the runtime of all tools concerning the file size, a runtime factor was calculated by dividing the average runtime of DamageProfiler by the average runtime of mapDamage2 and PMDTools, respectively. It could be demonstrated that DamageProfiler is on average 5.5 times faster than mapDamage2 and 174.6 times faster than PMDTools. Moreover, an increase in speed can be specifically observed for larger mapping files (Table 3.1, Figure 3.7). It is worth mentioning that PMDTools did not get any results on large BAM files within 24h (Motala12, Loschbour), so the run was aborted.

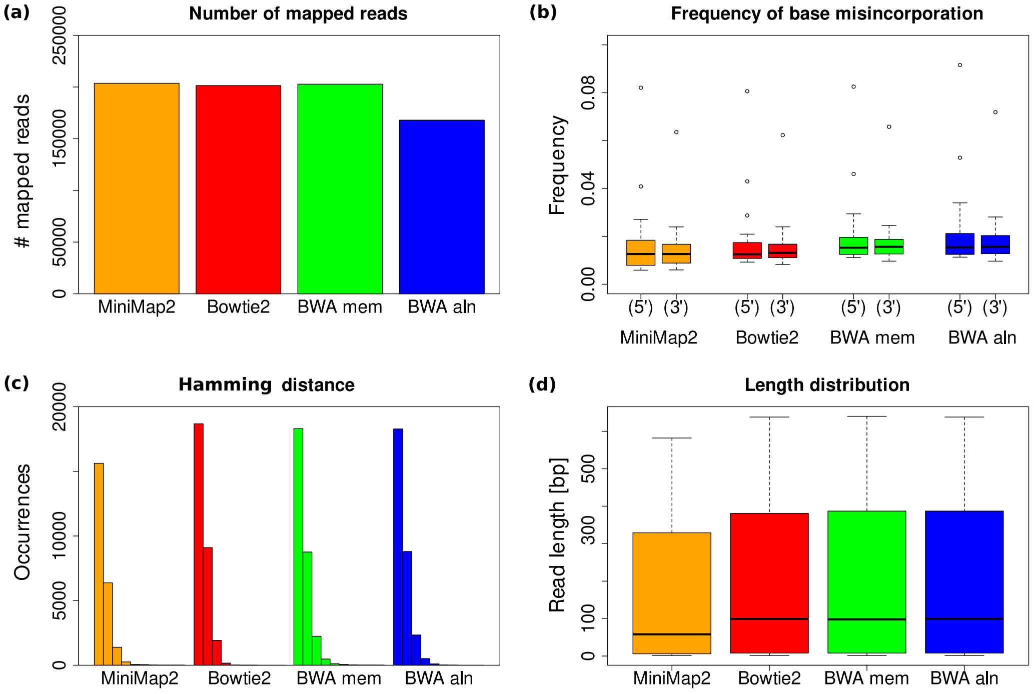


Figure 3.6: Comparison of different mapping software. DamageProfiler applied to mapping results of four different mapping software (MiniMap2 [147], Bowtie2 [144], BWA mem [146], and BWA aln [145]). The resulting (a) number of mapped reads (barchart), (b) the frequency of the specific base misincorporations (boxplot), (c) the edit distance (histogram), and (d) the length distribution (boxplot) are compared.

Table 3.1: Runtime comparison between mapDamage2 (MP2), PMDTools (PMDT), and DamageProfiler (DP). Five published data set [14, 28, 114, 119] of different size were used to evaluate the runtime. The runtime is given in seconds and the file size in megabytes. The factor is calculated by dividing the runtime of DamageProfiler by the runtime of mapDamage2 and PMDTools, respectively.

Sample	**Ref. genome**	**Mean Cov.**	**Size BAM**	**Runtime in [s]**			**Factor**	
		in [X]	**in [mb]**	**MD2**	**PMDT**	**DP**	**MD2 vs. DP**	**PMDT vs. DP**
TU7 (SAMN11897366)	NC_011112.1	70.9	0.7	3.3	49.8	2.0	1.7	24.9
JK2134 (SAMEA4461508)	NC_012920.1	131.9	1.5	4.2	122.1	2.1	2.0	58.1
Jorgen533 (SAMEA102120418)	NC_002677.1	9.6	21.6	28.3	2,689.5	6.1	4.6	440.9
Motala12 (SAMEA2697132)	GRCh37.p13	2.4	5,500.0	7,603.6	-	803.7	9.5	-
Loschbour (SAMEA2697124)	GRCh37.p13	22.0	50,700.0	81,132.7	-	8,410.2	9.6	-

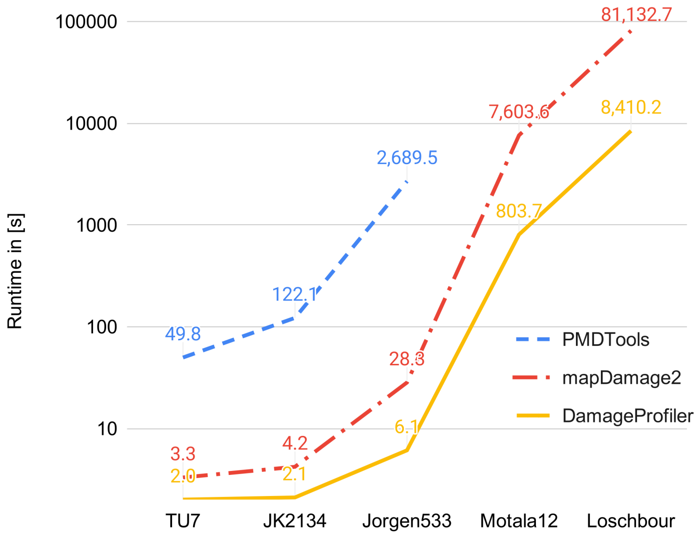


Figure 3.7: Runtime of DamageProfiler in comparison to mapDamage2[42] and PMDTools [43] with respect to the file size. The x-axis refers to the five samples of different sizes, the runtime [s] is log-scaled and given on the y-axis.

3.7 Discussion

DamageProfiler is a fast and user-friendly software to investigate damage patterns in ancient DNA from next-generation sequencing platforms, which runs independently of the mapping software used. It calculates aDNA-specific damage patterns, like length distribution, edit distance, and base substitution frequency, which are the commonly used features for aDNA authentication [15, 96]. The independence of the mapping software was tested by generating mapping files with widely used software; namely, BWA mem [146], BWA aln [145], Bowtie2 [144], and MiniMap2 [147]. While Bowtie2 and especially BWA are well-established mapping software in aDNA research, MiniMap2 is not suitable for samples with low coverage as it can generate sub-optimal alignments for regions of low complexity [147]. This limits its use for aDNA samples. Nevertheless, MiniMap2 has been included, as it may become more applicable for aDNA in the future. In addition, where available, the mapping parameters tested for aDNA were used for all mapping algorithms. No further editing of the mapping files was performed, as it is intended to demonstrate that DamageProfiler can process the direct output of each mapping tool. It could be observed that the number of reads used for the damage pattern calculations differs slightly between the different mapping algorithms (Figure 3.6). This is most likely due to the settings of the various mapping algorithms, for example, the number of mismatches or different weights that influence longer reads to be preferential. The differences are marginal, though, and have no significant impact on the resulting damage patterns.

DamageProfiler was run on a simulated data set to demonstrate the reliability of the results, which reproduced the expected results. Moreover, it was executed on four already published data sets [14, 28, 114, 119]. To verify the results of DamageProfiler, mapDamage2 [42] was run on the same mapping files, yielding identical results. However, in this way, the calculations could only be verified for statistics that are provided by both tools. The intersection of the functionality includes the length distribution, the base substitution frequency, and various statistics. The remaining functionality implemented in DamageProfiler, such as the edit distance, was tested and verified using JUnit tests.

In comparison with mapDamage2 [42], DamageProfiler does not offer the rescaling option. This functionality adjusts quality scores of likely damaged positions in the reads. A new BAM file is created where the quality values for misincorporations that are likely to be due to old DNA damage are scaled down according to their original qualities in the reads and their damage patterns. As the main focus of DamageProfiler lies in the fast and intuitive investigation of damage patterns, this option has not been implemented. Still, it could be considered as an additional feature in future versions. The unique ability of DamageProfiler to also handle metagenomic mapping files of arbitrary mapping software makes it also applicable in pipelines that are targeting both, single genome reconstruction and metagenomic analysis, such as nf-core/eager [148].

Finally, it could be shown that DamageProfiler is faster than the existing software [42, 43] for determining damage patterns. This is due to improvements through suitable packages and can even be improved for larger input files through a multi-threaded implementation. Furthermore, it could be shown that the runtime factor increases with increasing file size.

CHAPTER 4

mitoBench: Novel interactive methods and repository for population genetics of human mitochondrial DNA

4.1 Introduction

Besides nuclear DNA, a different extra-chromosomal genome can be found in mitochondria. Mitochondrial DNA (mtDNA), a circular, double-stranded molecule, was first detected in 1963 when Margit and Sylvan Nass intended to study some mitochondrial fibers that appeared to be related to DNA regarding fixation, stabilization, and staining behavior [149]. Eighteen years later, the first complete human mtDNA sequence was published [106]. This was the beginning of mtDNA success, which became the *workhorse* of human population genetics. The breadth of publications about human evolution based on mtDNA in the last forty years also show the high impact and usability of this molecule. Its important role in evolutionary biology is due to its unique characteristics, such as the small size (16,569bp), high copy number [108, 109], high mutation rate [110], and lack of recombination and variability at a population level [111, 112]. Due to its rapid rate of evolution, mtDNA in particular has played a central role in understanding human population genetics and discerning the movement of our species across the planet. Another important feature of mtDNA is its haploid nature, which differs from its nuclear counterpart. This means that the coalescence of neutral genes will be positively correlated with the effective population size of the species [150, 151]. From the perspective of population genetics theory, mtDNA evolves faster than nuclear DNA [152], which allows the analysis of divergence times in a higher resolution, within taxa or even species. Therefore, mtDNA has been widely used for population mapping, evolutionary and phylogenetic studies, species identification by DNA barcoding, diagnosis of various pathologies, and forensic medicine

[153, 154]. Conversely, there are several drawbacks to be aware of when working with mtDNA: it is partially contained in other parts of the nuclear genome as non-coding sequences [155] and accumulates mutations freely. These parts are called NUMTs (nuclear mitochondrial pseudogenes). They can cause issues, especially when enriching for mtDNA [154, 156] as these translocated mitochondrial sequences in the nuclear genome have the potential to be amplified in addition to, or even instead of, the authentic mtDNA target sequence. Such mis-amplifications can seriously affect population genetic and phylogenetic analyses. For instance, variations in NUMTs are misreported as mitochondrial mutations in patients [88], as for example, a NUMT on chromosome 1 has been associated with low sperm motility [157] and cystic fibrosis [158].

Scientists in paleogenetics have been able to study the ancestry of extinct populations by using ancient DNA (aDNA), and are thus able to retrieve information that archaeologists alone may not be able to discover. Major insights into the migration and admixture of humans have been made [14, 44, 46, 159] and have been connected to artifacts found by archaeologists [160]. For a long time, mtDNA was thought-to-be the genetic information most likely to be obtained from ancient individuals in quantity suitable for more detailed analysis [40]. With improvements in capture and enrichment methods, sequencing technologies, and bioinformatic tools, progress was made to retrieve complete nuclear genomes. Though nuclear DNA allows a deeper insight into human population genetics [161–163], it is often not accessible or possible to extract enough DNA. Thus, mtDNA is still used in this field of research, among other reasons, because of its easier accessibility in ancient samples and lower costs associated with the extraction and analysis of mtDNA.

Taken together, this has resulted in a precious collection of mitochondrial genomes over the decades. These data are not centrally accessible but often have to be compiled from individual publications, which is cumbersome and can lead to inconsistencies or errors.

Most analytical methods used to compare populations based on genetic difference, F_{st}-metric (Section 2.3), are independent of the DNA's source, i.e. of nuclear or mitochondrial origin. Existing software offers a wide range of analysis options for mtDNA [49, 164, 165]. However, in most cases, they are not compatible with standard genomic data file formats or only provide a specific output file format. As a result, population genetic analyses of mitogenomes can be cumbersome and error-prone, as input files have to be generated by hand and often converted to be used by different software. After more than four decades of research, no publicly accessible database brings together the still-growing body of modern and ancient human mitochondrial genomes in a single, well-curated, and easily accessible database. Instead, public data collections are focusing on either ancient [55] or modern data [53] only, are specialized on different topics such as forensics [51] or mitochondrial diseases [54, 166], or are simply kept private and not accessible or usable to the scientific community. Furthermore, private collections are often stored as Microsoft Excel spreadsheets (XLSX), CSV, or TXT files.

Especially when opening files with Excel, errors can creep in because numbers and gene names are automatically converted into dates and vice versa, as Ziemann *et al.* have demonstrated [167]. All this together leads to an immense amount of redundant work in preparing and collecting information and requires user-friendly, system-independent software capable of processing and converting all formats into each other and performing initial population genetic analyses. One of the goals of this chapter of the thesis is to create a central database to compile a data set for further analysis easily. For this purpose, a consortium consisting of international experts in the field of mtDNA was established.

To overcome these hurdles and provide a platform for the mitogenome community for all publicly available human mitogenomes, mitoBench has been developed within the scope of this thesis. The software is platform-agnostic and consists of a workbench and a comprehensive, quality-controlled database. The workbench provides a graphical user interface (GUI), offering analysis and visualization features focusing on population genetic applications. The GUI is kept simple to make it suitable for users without prior knowledge of computational methods. Additional information such as metadata and summary statistics, focusing as well as visualization methods, are provided. In addition, mitoBench also offers file conversion and compatibility tools to connect the workbench with existing analysis software such as BEAST [168, 169] and Arlequin [49]. Moreover, it is possible to directly execute HaploGrep2 [164] within the workbench. The analyses which have been implemented include basic statistical methods such as F_{st}-statistics and principal component analysis (PCA).

The quality-controlled database contains an extensive collection of complete and published mtDNA reference data and comprehensive metadata, most critical for sorting and filtering. All settings and kinds of analyses and visualizations performed within a mitoBench session are stored in a human-readable log file for better reproducibility.

This chapter describes the methodological basics, ideas, and concepts of mitoBench, its components, the workbench, and the database. Finally, analysis of a previously published data set consisting of ancient and modern mitochondrial sequences from Umbrians was repeated to demonstrate the reliability and reduction in using mitoBench.

4.2 Concept and components of mitoBench

4.2.1 The idea of mitoBench

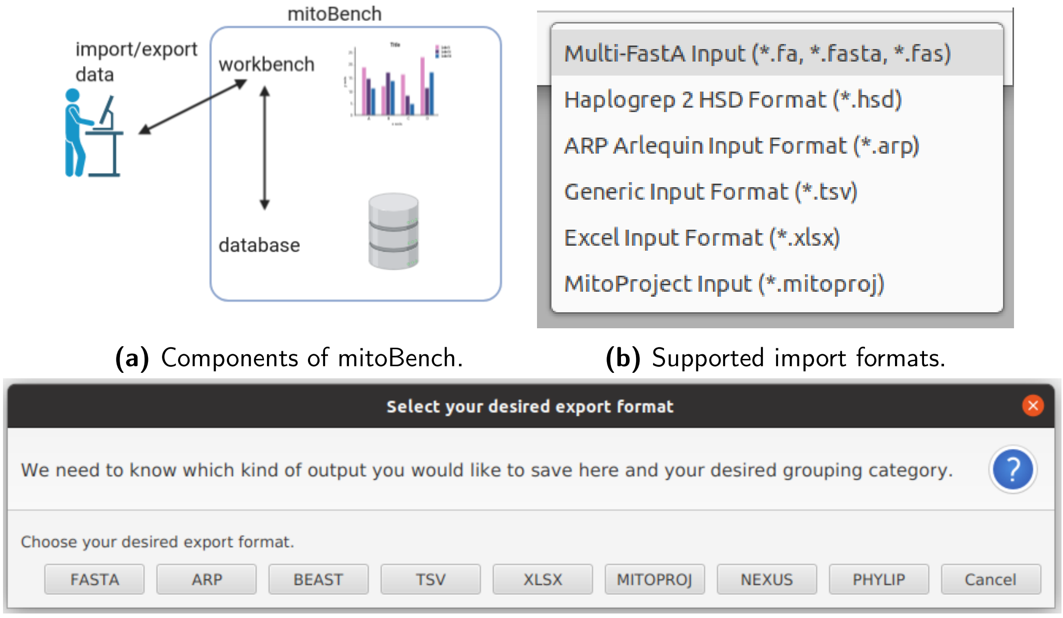
Collecting comparative data sets and mtDNA analysis is still arduous and error-prone and often needs to be done manually. MitoBench was developed to provide a convenient and user-friendly software to convert between file formats, combining genomic data with meta-information, and offering access to a public database of complete modern and ancient human mtDNA genomes. The initial idea of mitoBench was formed during the analysis of the data set published by Schuenemann *et al.* [119] of 151 Egyptian mummies, and the conceptual design was previously presented in the dissertation of Alexander Peltzer [170]. In analyzing the data mentioned above, the researchers were faced with systematically collecting and storing meta-information and converting between file formats to use different analysis software. Intensive research and discussion with several international mtDNA research groups from different fields revealed that this appears to be a common problem within the human genetics community. Therefore, the 'mitoBench Consortium' was founded to address these problems. The consortium comprises research groups from Italy, Germany, the United Kingdom, Austria, Russia, and Switzerland, coming from human population genetics, forensics, and genetic epidemiology. Based on topic-specific discussions among the groups, the decision was made to implement software providing an intuitive and easy way to work with mtDNA genome data and the associated meta-information. Moreover, at the beginning of the project, the need for a central database was recognized, which aims to collect public available modern and ancient human mitochondrial genomes, together with meta-information necessary for human population analyses. Therefore, the main requirements for mitoBench were postulated as follows:

- Being user-friendly
- Combining files of different file formats
- Exporting data into various file formats for further downstream processing
- Analyzing and visualizing the imported data together with their meta-information to enable the user a first investigation of the data
- Containing, or being compatible with, various state of the art tools, e.g. HaploGrep2 [164] or Arlequin [49]
- Accessing a central, well-curated, and high-quality database containing published modern and ancient complete human mitochondrial genomes

The main components of mitoBench are the workbench and the database (Figure 4.1a). The user can easily import and export data in various file formats (Figure 4.1b and 4.1c) and get access to a database of several thousand complete human mitochondrial genomes. Moreover, the workbench also acts as an analysis platform for data from the database and private data. In the following sections, the main functionality of both components is described in more detail.

4.2.2 The workbench

The workbench of mitoBench is the component with which the user mainly interacts. Therefore, it is designed to be as user-friendly and intuitive as possible. The


mitoBench
import/export
data
workbench
database
Multi-FastA Input (*.fa, *.fasta, *.fas)
Haplogrep 2 HSD Format (*.hsd)
ARP Arlequin Input Format (*.arp)
Generic Input Format (*.tsv)
Excel Input Format (*.xlsx)
MitoProject Input (*.mitoproj)
FASTA
ARP
BEAST
TSV
XLSX
MITOPROJ
NEXUS
PHYLIP
Cancel

(a) Components of mitoBench.

(b) Supported import formats.

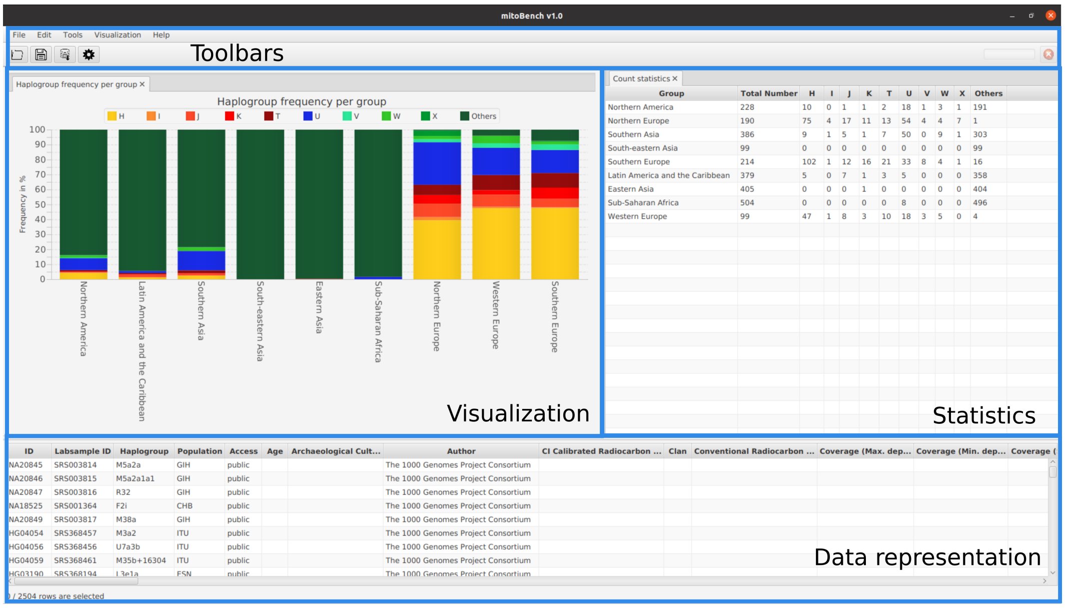
Select your desired export format

We need to know which kind of output you would like to save here and your desired grouping category.

Choose your desired export format.

(c) Supported export formats.

Figure 4.1: Overview of mitoBench. (a) MitoBench comprises a workbench and a database. The workbench serves as interface to access and download data form the database. In addition, the workbench also provides analysis methods for mtDNA and a number of different visualizations. Multiple file formats are provided for data (b) import and (c) export.



Group	Total Number	H	I	J	K	T	U	V	W	X	Others
Northern America	228	10	0	1	1	2	18	1	3	1	191
Northern Europe	190	75	4	17	11	13	54	4	4	7	1
Southern Asia	386	9	1	5	1	7	50	0	9	1	303
South-eastern Asia	99	0	0	0	0	0	0	0	0	0	99
Southern Europe	214	102	1	12	16	21	33	8	4	1	16
Latin America and the Caribbean	379	5	0	7	1	3	5	0	0	0	358
Eastern Asia	405	0	0	0	1	0	0	0	0	0	404
Sub-Saharan Africa	504	0	0	0	0	0	8	0	0	0	496
Western Europe	99	47	1	8	3	10	18	3	5	0	4

ID	Labsample ID	Haplogroup	Population	Access	Age	Archaeological Cult...	Author	CI Calibrated Radiocarbon ...	Clan	Conventional Radiocarbon ...	Coverage (Max. dep...	Coverage (Min. dep...	Coverage (
NA20845	SRS003814	M5a2a	GIH	public			The 1000 Genomes Project Consortium						
NA20846	SRS003815	M5a2a1a1	GIH	public			The 1000 Genomes Project Consortium						
NA20847	SRS003816	R32	GIH	public			The 1000 Genomes Project Consortium						
NA18525	SRS001364	F2i	CHB	public			The 1000 Genomes Project Consortium						
NA20849	SRS003817	M38a	GIH	public			The 1000 Genomes Project Consortium						
HG04054	SRS368457	M3a2	ITU	public			The 1000 Genomes Project Consortium						
HG04056	SRS368456	U7a3b	ITU	public			The 1000 Genomes Project Consortium						
HG04059	SRS368461	M35b+16304	ITU	public			The 1000 Genomes Project Consortium						
HG03190	SRS368194	L3e1a	ESN	public			The 1000 Genomes Project Consortium						

) / 2504 rows are selected

Figure 4.2: The mitoBench main window is divided into three sub-windows: Visualization, Statistics, and Data representation. The toolbar at the top of the window allows easy and fast navigation.

workbench is composed of three main windows: data representation, visualization, and statistics (Figure 4.2). The data representation window displays the data imported to the workbench, either public data from the database or/and data from a private collection. Data from the personal collection can be imported via an import template or the various file import functions of the workbench. After formatting the private data according to mitoBench's standards, the data can easily be imported. The visualization window contains all figures that were made based on the (selected) data from the data representation window. Each visualization is stored in a separate tab with the possibility of switching between figures.. After launching mitoBench, the statistics window shows a welcome page, referencing the documentation, and various database statistics, such as the number of modern and ancient samples or publications covered. This window hosts various statistics and setting dialogues, which are separated into tabs. The user can navigate through all functions of mitoBench using the menu bar at the top of the main window. An additional toolbar below, which contains individual buttons for the main functions, enables faster operation.

Data import

Data can be uploaded to the workbench either via direct file upload of various file formats or simply by using the drag-and-drop option in the data display window. Alternatively, the data can also be imported from the database. For the former, mitoBench supports the data import from various commonly used file formats, such as FASTA, TSV, HSD, ARP, and XLSX. In doing so, the sequence and the associated meta-information do not necessarily have to be submitted as a single file or in the same file format. In this way, the results of different previously run, such as Arlequin [49] or HaploGrep2 [164], can be combined by the workbench. The only requirement is to use identical sequence identifiers across all analyses, which act as a key per sample to merge all information later. In addition to this option, it is possible to import meta-information, including the sequence. This can be done either via a template accessible online or via a generic TSV file created by the user, provided the specific format requirements are met. The combined information is displayed in the data window of the workbench.

Furthermore, data can also be imported directly from the database, which users can easily access through the workbench (Section 4.2.5). The two options for data import can also be combined: In addition to the user's own (private) data, which can be imported via file upload, the database can be loaded into the workbench. This allows users to put their data in context with previously published data quickly.

Data completion and validation

An essential aspect of mitoBench is high data quality, as this is the prerequisite for meaningful analyses. To ensure high data quality of the data imported into the workbench or loaded into the database, mitoBench offers two methods: Data validation and data completion. Both services can either be started within the workbench or downloaded as separate tools from the corresponding GitHub repositories.

First, the data can be validated. This step is recommended after adding new data to the workbench and obligatory if data is uploaded to the database. The data validation ensures that all values have the correct format, e.g., String, Integer. Furthermore, it also ensures compliance with specific standards for particular attributes. For example, only defined values or ranges are allowed for some features, e.g., for geographical descriptors such as continent or country, haplogroups, or publication details.

Second, mitoBench offers the function *Complete Information*. This includes the automated calculation of particular values, such as the percentage of N's in the mtDNA sequence or defining a user alias. It also involves the integration of geographical locations. It first checks the correctness of the specified country or continent and then completes the information. This means that others can be assessed with appropriate Java libraries if only one geographical location is given. For example, if the latitude and longitude are given for the sample, the corresponding country and continent are automatically filled.

Data visualization

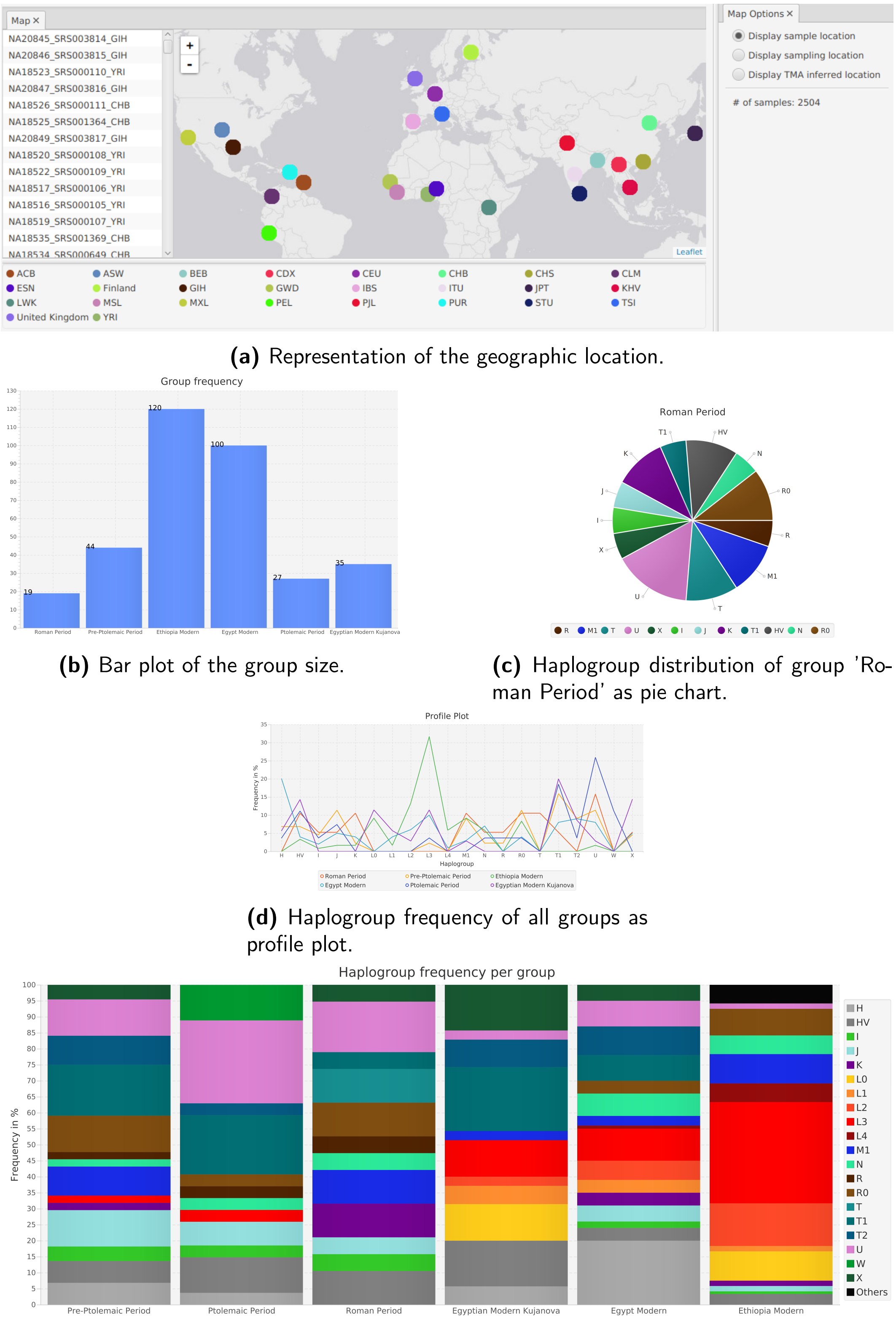
Regardless of the method used, the data can be visualized and analyzed further after a successful import. The analysis and interactive visualization methods provided by mitoBench result from discussions among the mitoBench Consortium and represent the most commonly used analyses and visualizations.

The visualizations are implemented using JavaFX, as is the case for the entire User Interface. JavaFX provides a modern scenegraph API, supporting 2D and 3D, and media and web controls. This allows the set up of powerful interactive plots, helping the user better understand their data. The underlying information for most of the visualizations is the haplogroup, which is determined with HaploGrep2 [164]. The haplogroup distribution of the considered sample set can be visualized in various ways. If other groups are now defined, it is also possible to display the haplogroup frequency of each group. In addition, the geographical location of the samples can be displayed on a map. Furthermore, the haplogroups assigned to the investigated samples can be shown as a tree. This tree organizes the haplogroups according to the classification of PhyloTree [122].

As the most interest in population genetic studies often compares different groups or populations, mitoBench provides a grouping option. This allows a grouping of the data based on an arbitrary attribute, for example, *sample country* or *population*. This grouping will be kept for all further analyses but also can be changed by the user.

An important descriptor of mtDNA and therefore used for many visualizations and analyses is the haplogroup. The representation of haplogroups has a tree-like structure, which allows a description of a group of haplogroups by the node that contains the respective group. This summarizing node is called macro-haplogroup. For example, L1 would be the macro-haplogroup for all haplogroups that are included in the branch L1 according to PhyloTree [122]. These macro-haplogroups are useful for visualizations, as it allows the summary of haplogroups and clear visualization. Additionally, the macro-groups are characteristic of particular geographic regions. To help the user with the set up of an appropriate list, mitoBench provides pre-built lists for various geographic regions in the world that were calculated based on the data available in the database, together with already published knowledge about the occurrences of haplogroups in different geographic areas (Figure 4.4). These lists can be selected at each visualization analysis. In addition, the user can also specify his list. To avoid errors when re-typing this list for each analysis, mitoBench offers the option to select this list only once, and it will be automatically applied for each analysis. Still, it can be adapted to each part of the analysis.

An overview of all possible visualizations is shown in Figure 4.3. A first investigation of the samples can be the geographic location of the sample or sampling origin, or the location of the terminal maternal ancestor (TMA) (Figure 4.3a). Groups are indicated using different colors. The number of samples per location is visible as a popup when hovering over the symbol at the corresponding location. To get more information about the groupings, the group size can be visualized as a bar plot (Figure 4.3b). For a more detailed investigation of the haplogroup composition per group, the haplogroup distribution can be investigated in differ-



(a) Representation of the geographic location.

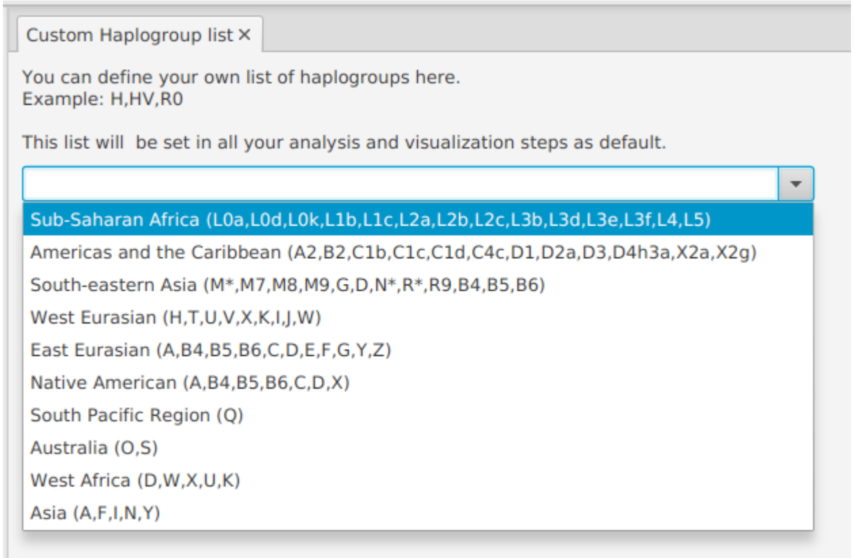
(b) Bar plot of the group size.

(c) Haplogroup distribution of group 'Roman Period' as pie chart.

(d) Haplogroup frequency of all groups as profile plot.

(e) Haplogroup distribution as stacked bar chart.

Figure 4.3: Visualizations within mitoBench.



You can define your own list of haplogroups here.
Example: H,HV,R0

This list will be set in all your analysis and visualization steps as default.

Sub-Saharan Africa (L0a,L0d,L0k,L1b,L1c,L2a,L2b,L2c,L3b,L3d,L3e,L3f,L4,L5)
Americas and the Caribbean (A2,B2,C1b,C1c,C1d,C4c,D1,D2a,D3,D4h3a,X2a,X2g)
South-eastern Asia (M*,M7,M8,M9,G,D,N*,R*,R9,B4,B5,B6)
West Eurasian (H,T,U,V,X,K,I,J,W)
East Eurasian (A,B4,B5,B6,C,D,E,F,G,Y,Z)
Native American (A,B4,B5,B6,C,D,X)
South Pacific Region (Q)
Australia (O,S)
West Africa (D,W,X,U,K)
Asia (A,F,I,N,Y)

Figure 4.4: Predefined haplogroup lists that are specific for a certain geographic region. The selected list will be kept for all following analyses.

ent ways and levels. A pie chart, for example, provides a detailed overview of the haplogroup composition of each group (Figure 4.3c). In contrast, the profile plot (Figure 4.3d) emphasizes the number of samples with a specific haplogroup within a group, allowing a direct comparison. The profile plot is connected to the statistics table on the right part of the workbench, which provides additional statistics, and hovering over a row in the table highlights the corresponding line in the visualization. A commonly used visualization for population genetic analyses is the stacked bar chart of haplogroups (Figure 4.3e). Each group is represented as a bar reflecting the haplogroup distribution of all mitochondrial genomes in the corresponding group. The flexible order of the bars allows sorting, for example, according to a time period or geographic distance. This provides an intuitive way of comparing the distributions between the different groups.

Data analysis and functionality

MitoBench provides fundamental analysis methods used in population genetics to assess the genetic distance between defined groups of individuals, or individuals themselves, such as F_{st}-value and principal component analysis (PCA). The distance measure of these calculations is usually based on the sequence itself or the haplogroups derived from it, both provided by mitoBench.

Two of the most common analyses of the genetic proximity of individuals or groups are the F_{st} and PCA. The F_{st} metric, also called fixation index, is a value based on the number of bases that differ between mtDNA sequences. The index can range from 0 to 1, where 0 means complete sharing of genetic material, and 1 the opposite (Section 2.3). The results are finally displayed in the visualization window as a heat map, which is commonly provided after an F_{st}-analysis (Figure 4.11a).

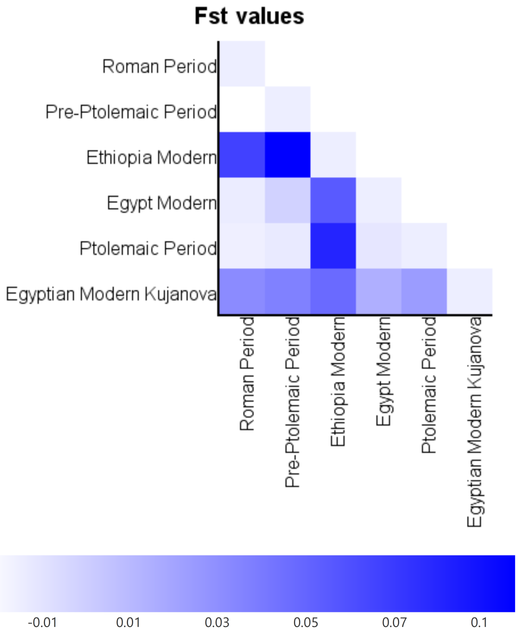
In contrast to the F_{st}-value, the PCA is based on the haplogroup distribution within the investigated set of groups. Generally speaking, it is used to reduce the dimensionality of a larger data set and to identify the specific variables or combinations of variables that would explain the statistical variance seen in the data [171] (Section 2.3). The PCA implemented in mitoBench is based on the haplogroup frequencies present in the different groups. Therefore, the more representative the data set is, the better resolution will be achieved. The closer the groups are placed, the closer the genetic relationship is, or in other words, the more similar is the haplogroup distribution within the groups. In mitoBench, the PCA analysis is calculated using an appropriate Java library and visualized as a scatter chart; each dot represents a group. The dots are labeled with the corresponding group name (Figure 4.11b).

Data export

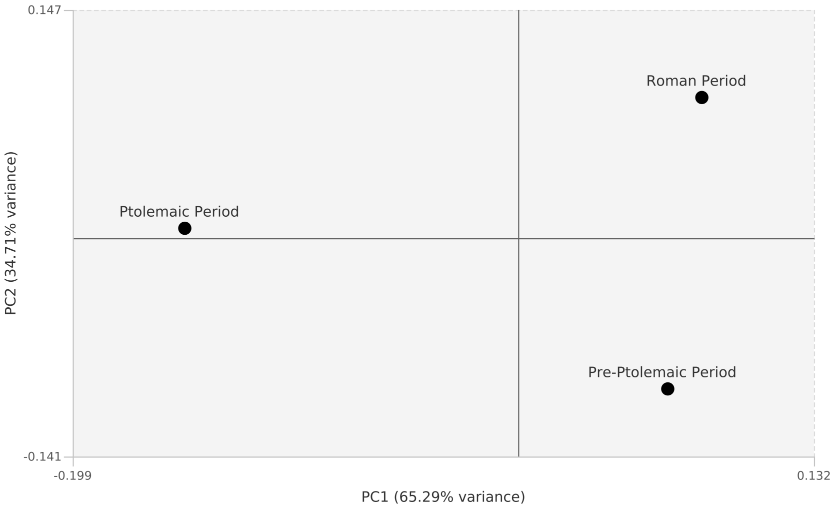
One of the main intentions of developing mitoBench was to facilitate the conversion between different file formats and get a first overview of a data set. Therefore, exporting the data alone or combined with (selected) meta-information for further analysis is crucial. The export function of mitoBench offers multiple file-formats; an overview is illustrated in Figure 4.1c. Plain sequences can be downloaded in FASTA format or, together with all meta-information, as more generic formats such as TSV and XLSX. In addition, mitoBench also offers file formats that can directly be used as input for downstream analysis tools. This includes the ARP format (input format for Arlequin [49]), BEAST format (containing a multiple sequence alignment with the dating information included to the header as needed for BEAUti [172]), or HSD format, which can be loaded into HaploGrep2 [164] for more details about the haplotypes occurring within the sequence. Although mitoBench does not directly support calculating the phylogeny on a given set of samples, a multiple sequence alignment can be calculated with MAFFT [173]. The alignment can then be downloaded in either FASTA or PHYLIP format, which is often required by software investigating phylogenetic relationships, for example, RAxML [174], or PhyML [175].

Besides the data export, each figure can be downloaded as a high-resolution PNG file. Since the style of visualization often varies greatly depending on the information to be highlighted or the preferences of the scientist, mitoBench also offers the export of the values on which the graphs are based. This allows the user to design the visualization with any software without having to perform additional complex calculations.

Finally, it is possible to save all settings and data as a project file (MITOPROJ) to continue the analysis.



(a) Result of F_{st}analysis.



(b) Result of PCA analysis.

Figure 4.5: Graphical representation of the analysis results. (a) The F_{st}analysis is visualized as heatmap. The darker the color (blue), the higher the F_{st}value. (b) The PCA represents each group as dot, labeled with the corresponding group name. By default, the first and second principal component are visualized.

4.2.3 Software testing and documentation

The software was tested during the development using JUnit5 and TextFX [176], which simulates user interactions, such as clicks, and tests the expected behaviors. This ensures quality checks are already at an early stage of development and prevents the introduction of errors that can hardly be fixed later on. Additionally, the import and export functionality is tested automatically whenever the application is built using Travis CI's service.

Furthermore, mitoBench comes with comprehensive documentation, provided by the ReadTheDocs platform (`https://mitobench.readthedocs.io/`). The manual contains detailed documentation about the provided analyses and visualizations and exemplary data and a FAQ. The documentation can be assessed online or downloaded in HTML, PDF, or ePUB format. Moreover, links to other web pages, including the glossary of the database attributes and the list of publications included in the database, are provided.

4.2.4 Database

While analyzing a set of samples, e.g., a population from a specific site, is already interesting by itself, more background and information about the samples can be gained by comparison with other populations or groups. For this, an extensive and informative database is needed, containing the mtDNA sequence itself, together with describing meta-information. Usually, researchers collect these databases over time, keeping them private, and often maintain and update the text or Excel sheets by hand. On the one hand, this is sensitive to errors, and on the other hand, very time-consuming and done redundantly among different research groups. To get one step closer to easier maintenance, access, and sharing of collected and published mtDNA and meta-information, the mitoBench database was designed.

In the following, the conceptual design of the database is introduced, and features for data processing and the standard data import workflow.

Database design

The initial design of the database was developed together with Alexander Peltzer and is already described in his dissertation [170]. Only an overview of the initial setup is given here, and the focus is on the necessary adjustments for the final database. First, all members of the mitoBench consortium submitted an exemplary data set of their private data collections to understand the database requirements. The submitted data were usually Excel sheets containing 100-200 samples. The evaluation of the submitted data revealed several attributes that were present in all groups. This resulted in a number of mandatory features that are required for each submission to the database. Therefore, additional meta-information can be specified. Based on the sample data submitted, the data tables were abstracted and stored in multiple table layouts, as shown in Figure 4.6. PostgreSQL was used

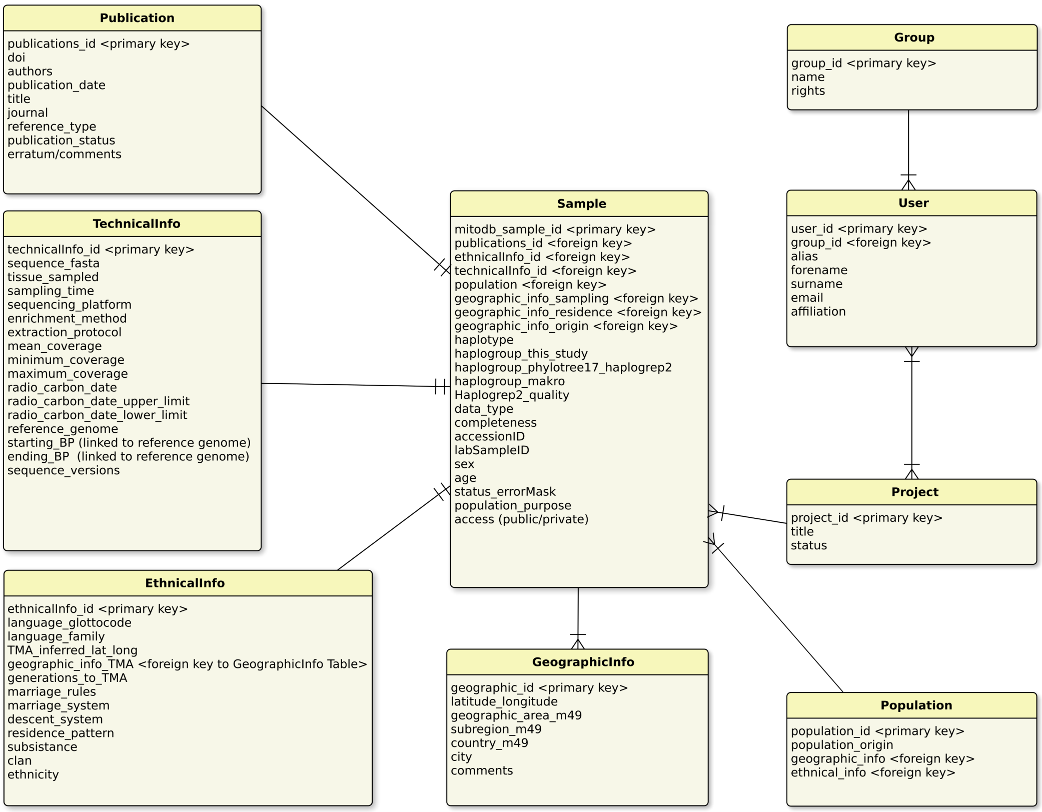


Figure 4.6: Initial database layout showing the available single tables, as published by Peltzer [170].

to build and design the database, allowing theoretically simultaneous access by hundreds of users with thousands of queries. Moreover, this also allows the use of more than one backend server so that further scaling of the database layout is possible in the future, if necessary.

The original layout consists of a total of nine tables (Figure 4.6). The center is formed by table 'Sample', containing mainly mandatory information and information directly derived from the mtDNA - the remaining eight tables store user information and further information about the sample background. In addition to making query time as efficient as possible, this layout was also chosen to automate the manual work of data curation and quality checking as much as possible. Therefore, the design decisions define only low-level quality checks such as the sequence quality assigned by HaploGrep2 [164].

During the implementation of the database, various adaptations had to be made to the design, which is explained in detail below, and the final attributes are visualized in Figure 4.7. First, an adaptation had to be made to handle identical mtDNA sequences submitted with different publications but the same accession ID. Since the database was initially indexed by accession ID, this would not allow entries with the same identifier. Therefore, the mitoBenchID has been introduced and used as an index for the database table. In addition, the functionality to get private access to a specific project, which can be shared with collaboration partners, was ultimately not implemented. Therefore, the tables 'Groups' and 'Project', as well as the user password, were removed. In addition, all tables have been combined into one table for easier handling. Moreover, attributes were added to describe ancient samples better. More minor adaptations were made on attribute level, such as the addition and exclusion of various fields that turned out to be necessary and dispensable, respectively.

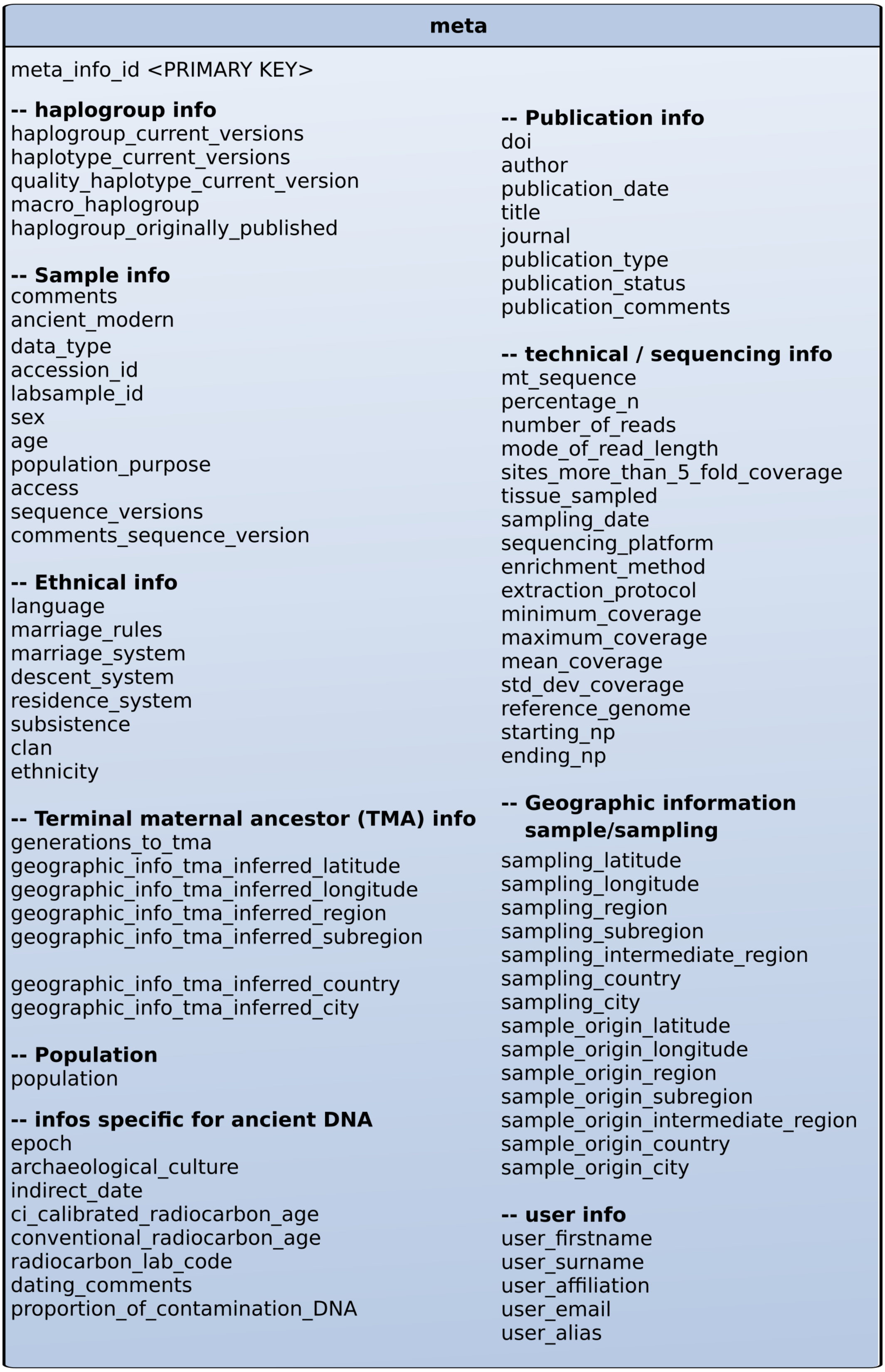
Standardization efforts

Standards for data storage and curation are critical for the rapid implementation of developed applications. While investigating the exemplary data sets, no standards or common formatting were detectable among the research groups. As part of the mitoBench project, we have therefore implemented procedures to provide certain parts of the database with standardized data types to ensure that all samples in the database are homogeneously formatted. One example is the use of geographic descriptors, which are based on the classification of M49 [177], city and country names according GeoNames [178] and OpenStreetMap [179]. As far as language classification is concerned, we use Glottolog [180] as the source of language categories, while we adhere to ISO standards for all other columns. To ensure data entry consistency, mitoBench also integrates other controlled vocabularies into the database. These can be viewed online and are to be used to create the metadata files.

Sample processing, quality checks, and data import into database

One of the main goals of mitoBench is to provide a comprehensive database that contains well-curated and high-quality data. To this end, an initial data set has been built, consisting of 26,522 modern and 1,505 ancient sequences, covering 153 countries and five continents (as of today, September 5, 2021). With the release of mitoBench, additional genome data can not only be uploaded by members of the consortium but also by other researchers, with the aim that any user can easily submit data via the workbench. A template to submit the meta-information is provided and accessible online via the documentation (`shorturl.at/kwST8`), as well as a detailed description of the required information (`shorturl.at/nBHQ7`). The corresponding mtDNA sequences, which must be mapped against the revised Cambridge Reference Sequence (rCRS) reference, can be submitted as separate files in FASTA format. These data will then automatically be validated, completed, and quality-checked to keep the manual work as low as possible.

To ensure the upload of only high-quality data, we implemented specific quality measures. First, no sequences with more than 1% missing data are allowed. This rule applies both to modern and well as ancient samples. Second, the quality of the sequence itself is rated by the quality provided by HaploGrep2 [164] and provided as



meta

meta_info_id <PRIMARY KEY>

-- haplogroup info
haplogroup_current_versions
haplotype_current_versions
quality_haplotype_current_version
macro_haplogroup
haplogroup_originally_published

-- Sample info
comments
ancient_modern
data_type
accession_id
labsample_id
sex
age
population_purpose
access
sequence_versions
comments_sequence_version

-- Ethnical info
language
marriage_rules
marriage_system
descent_system
residence_system
subsistence
clan
ethnicity

-- Terminal maternal ancestor (TMA) info
generations_to_tma
geographic_info_tma_inferred_latitude
geographic_info_tma_inferred_longitude
geographic_info_tma_inferred_region
geographic_info_tma_inferred_subregion

geographic_info_tma_inferred_country
geographic_info_tma_inferred_city

-- Population
population

-- infos specific for ancient DNA
epoch
archaeological_culture
indirect_date
ci_calibrated_radiocarbon_age
conventional_radiocarbon_age
radiocarbon_lab_code
dating_comments
proportion_of_contamination_DNA

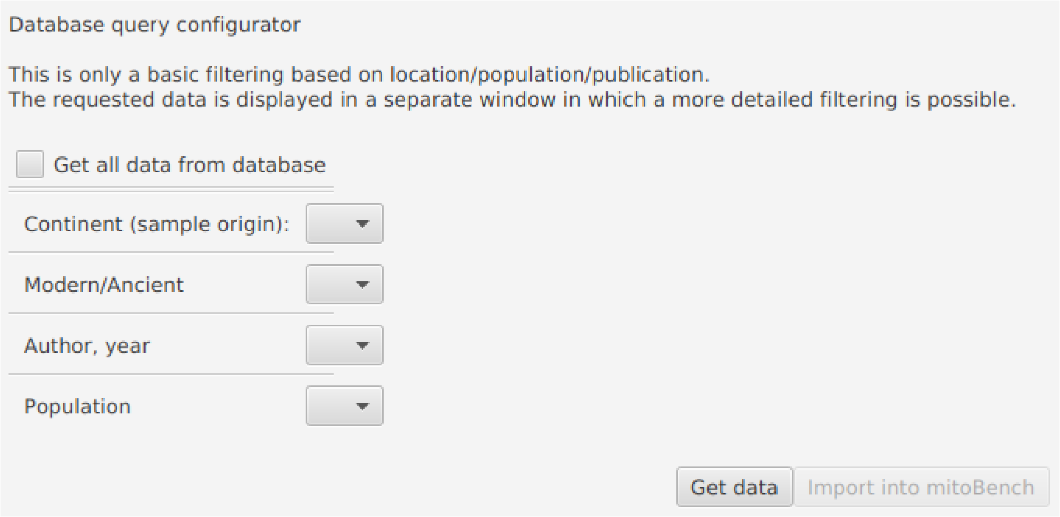
-- Publication info
doi
author
publication_date
title
journal
publication_type
publication_status
publication_comments

-- technical / sequencing info
mt_sequence
percentage_n
number_of_reads
mode_of_read_length
sites_more_than_5_fold_coverage
tissue_sampled
sampling_date
sequencing_platform
enrichment_method
extraction_protocol
minimum_coverage
maximum_coverage
mean_coverage
std_dev_coverage
reference_genome
starting_np
ending_np

-- Geographic information sample/sampling
sampling_latitude
sampling_longitude
sampling_region
sampling_subregion
sampling_intermediate_region
sampling_country
sampling_city
sample_origin_latitude
sample_origin_longitude
sample_origin_region
sample_origin_subregion
sample_origin_intermediate_region
sample_origin_country
sample_origin_city

-- user info
user_firstname
user_surname
user_affiliation
user_email
user_alias

Figure 4.7: Final database layout showing the available attributes.



Database query configurator

This is only a basic filtering based on location/population/publication.
The requested data is displayed in a separate window in which a more detailed filtering is possible.

□ Get all data from database

Figure 4.8: Database accession window allowing the user to configure the database search to download the requested data.

information to the user. Lastly, several attributes are mandatory. These files were decided as essential for any kind of mtDNA analysis by the mitoBench Consortium and include the accession ID, the data type, the author, publication status, user contacts, and the mtDNA sequence as at least one geographic information, such as sample or sampling location.

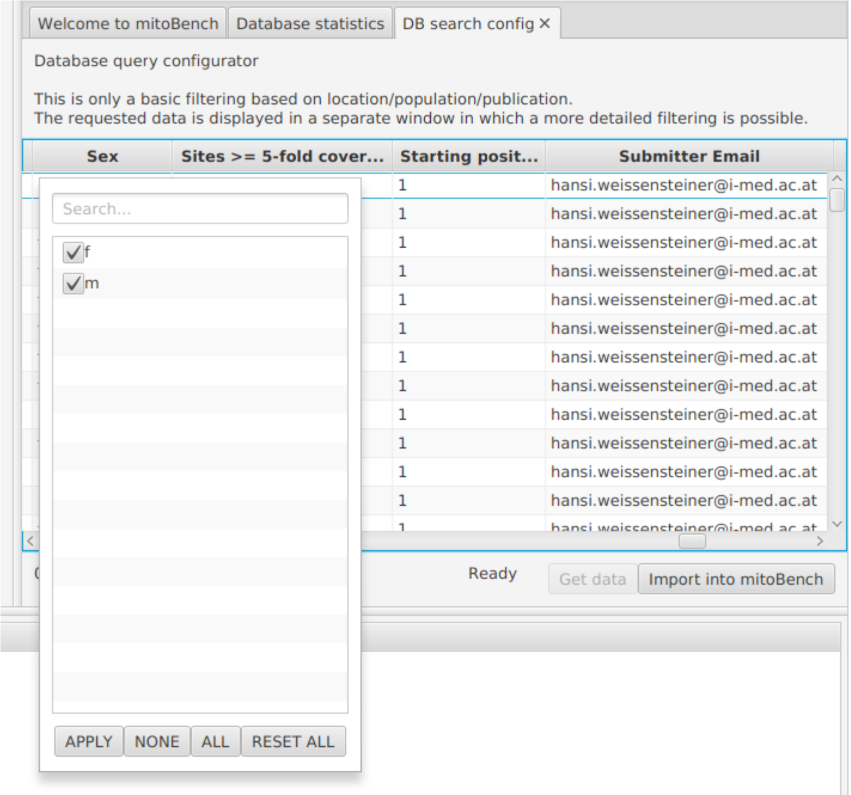
It is currently standard and recommended in the aDNA community to upload the raw sequencing data to a common data repository, e.g., SRA or ENA, within the publication process. This offers the possibility to process the data with its quality thresholds and the fast-changing state-of-the-art software. However, it also required additional pipelines to provide mtDNA sequence data imported into the mitoBench database finally. For this purpose, a pipeline has been implemented by Alexander Hübner using Snakemake [181] to provide mtDNA sequences and additional statistics in a standardized form. This pipeline can be found online and publicly available in the GitHub Code repository[1].

4.2.5 Interaction between workbench and database

The database can be accessed directly via the workbench without registration. The database is located and maintained by the IT department of the Max Planck Institute for the Science of Human History to ensure long-time support. However, access to everything on the Institute's network is protected by firewalls, so direct access to the database is impossible. This was solved with the HTTP client library Unirest[2], an interface between the client and the server within a network. The

[1] https://github.com/alexhbnr/mitoBench-ancientMT

[2] https://www.baeldung.com/unirest



Database query configurator

This is only a basic filtering based on location/population/publication.
The requested data is displayed in a separate window in which a more detailed filtering is possible.

Sex	Sites >= 5-fold cover...	Starting posit...	Submitter Email
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at
		1	hansi.weissensteiner@i-med.ac.at

Search...

☑ f
☑ m

Figure 4.9: After downloading data from the database, further filtering can be done using the column headers.

data request by the user is transformed and conveyed to the REST API, which collects the queried data from the database and sends it back to the client. In addition, access rights can be defined, and specific commands, such as 'DELETE', are not allowed. Requests can be formulated with the help of the database request window (Figure 4.8). This offers filtering based on the most informative columns of the database, such as population, geographic region, the publication (first author, year, and title), and modern or ancient samples. Additionally, it is also possible to download the entire database. After importing the data to the workbench, they can be filtered further within the workbench. Each column has a filter property implemented in the column header and selects and deselects specific values. For example, after downloading all data from the 1000 Genome Project, the data can be filtered further on sex (Figure 4.9).

4.3 Evaluation

To ensure the reliability of mitoBench, in particular, the correct operation of the workbench and the database, a previously published study by Modi *et al.* [182] was reanalyzed to reproduce the results. This study was chosen as most of the analysis and visualization features provided by mitoBench, such as $F_{st}$and PCA analysis based on modern and ancient complete mtDNA and various graphical representations of the haplogroup distribution, are covered. Modi *et al.* [182] examined the mtDNA variation of 19 ancient pre-Roman and 538 modern mitochondrial genomes, covering entire Umbria. However, only 198 modern mt genomes could be used for the mitoBench evaluation here, as only the hypervariable regions were available for the remaining sequences. Their study concludes that most individuals could be assigned to West Eurasian haplogroups, revealing a high mtDNA diversity in Umbria. These lineages in the region are quite homogeneous, except haplogroup K and J, which comprise the highest frequency (17%) in southern Umbria and 30% of the present inhabitants of the present inhabitants of the eastern area, respectively.

The main steps of their analysis were the determination of the haplogroups of the newly sequenced mitochondrial genomes using HaploGrep2 [164] and visualizations based on them. Moreover, they investigated the genetic distance between the ancient and modern samples using F_{st} and PCA analysis. While some further analyses were performed within the study of Modi *et al.* [182], e.g., calculations based on the hypervariable regions and phylogenetic analyses, these steps can be performed within mitoBench and also lead to the same conclusions.

In the following, the individual steps are described with which the results of Modi *et al.* [182] can be reproduced with the mitoBench. Based on the data of this study that can already be downloaded from the database, the haplogroup determination and possible visualizations and analyses based on it are shown. Furthermore, it will be shown how additional genomes can be easily downloaded for further analysis using the database configurator.

4.3.1 Data collection

First, the fasta sequences of 191 modern and 19 ancient mitochondrial genomes were downloaded. As the study is already published, this could be done via mitoBench, filtering for publication. However, only 16 ancient mtDNA sequences of this study were available, as three genomes did not fulfill the quality criteria of mitoBench. These sequences were added manually to reproduce the study as closely as possible to the original publication. Furthermore, six additional samples were added from other publications [183, 184], which were also analyzed by Modi *et al.* [182]. This resulted in a data set of 198 modern and 19 ancient mitochondrial genomes.

4.3.2 Data analysis and visualization

After collecting all genomes, the haplogroups were determined using HaploGrep2 [164]. The results are consistent with the haplogroups published by Modi *et al.* (Appendix, Table A.4). The haplogroup distribution was visualized with one pie chart each for the modern and ancient data set (Figure 4.10a), as already done by Modi *et al.* [182]. Additionally, the haplogroup distribution of Umbria's geographic regions are visualized (4.10b).

The genetic distance within the studied modern Umbrian population was investigated by calculating the F_{st} values between North (NU), West (WU), East (EU), South (SU), Central (CU), and Central-East Umbria (CEU). The results show the genetic distance of the inhabitants from eastern Umbria to the other groups (Figure 4.11a). The distribution is consistent with Modi *et al.*, showing that the individuals were mainly assigned to western haplogroups, having an overall homogeneous distribution. Notable are haplogroup K, predominantly present in southern Umbria, and haplogroup J, comprising the highest frequency in Southern Umbria.

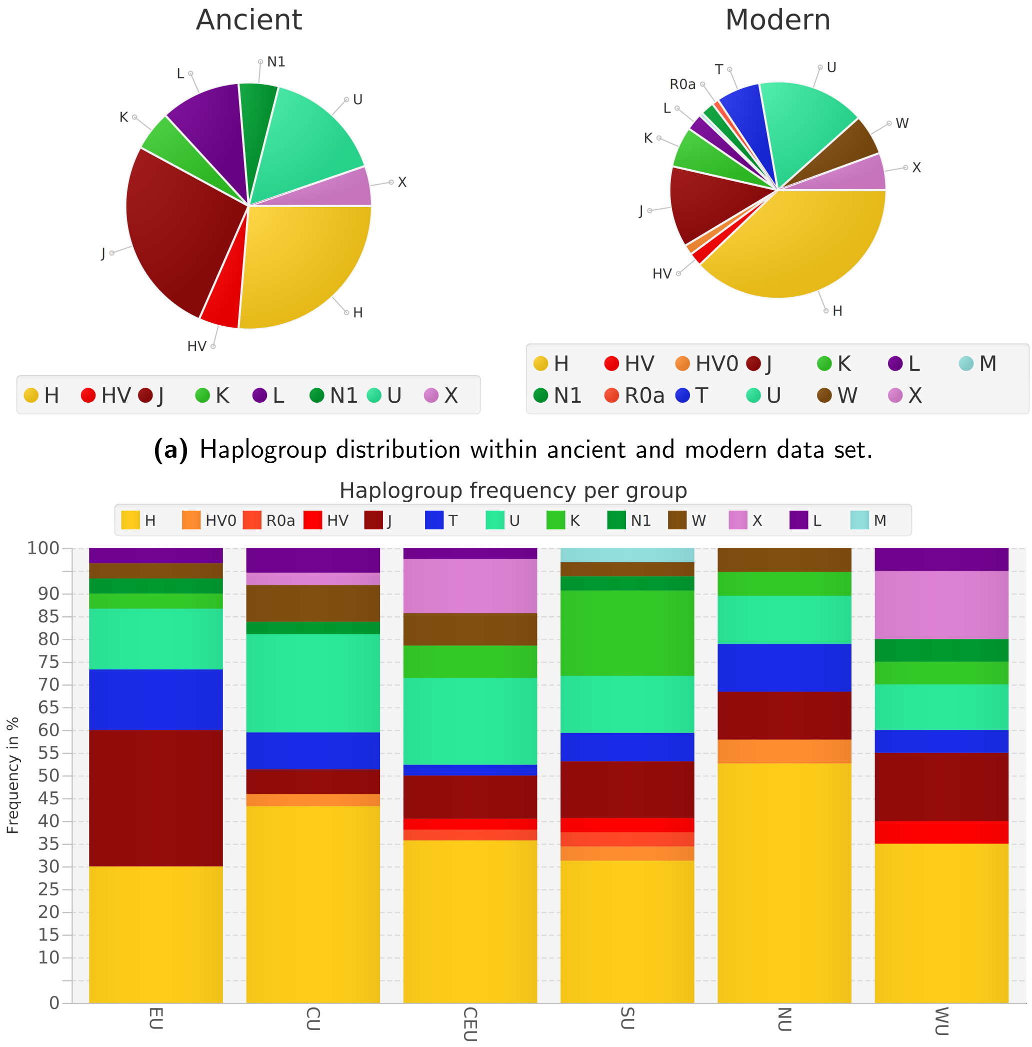
Finally, a PCA analysis was carried out based on the modern samples from the six Umbrian sub-regions and a Eurasian data set. For this purpose, Modi *et al.* suggests a data set of 15,650 individuals, covering 58 populations. However, most of the data is only given as the hypervariable region of the mtDNA, and therefore cannot be used with mitoBench. Therefore, a comparable data set was compiled using the mitoBench database configurator. Based on the list of countries used by Modi *et al.*, complete modern mitochondrial genomes were selected if available. This resulted in a final data set of 8,059 modern samples from 37 geographic regions. The result is congruent with the published ones, clearly showing the affinity of different parts of Umbria with the typical Mediterranean population, only Eastern Umbria clusters together with Eastern European countries (Figure 4.11b).

4.3.3 File download for downstream analysis

All further analyses conducted in the publication are based on the hypervariable regions of the mitochondrial genome and can therefore not be executable with mitoBench, or are not provided. However, mitoBench offers the export of the data in various file formats, which eases further downstream analyses. Examples are the export in ARP format for in-depth human population genetics analyses using Arlequin [49], in BEAST format to estimate the effective population size or the divergence times using BEAST2 [168], or PHYLIP format for phylogenetic investigations. Furthermore, all haplogroup statistics, such as the haplogroup frequency, can be downloaded and processed further.

4.4 Discussion

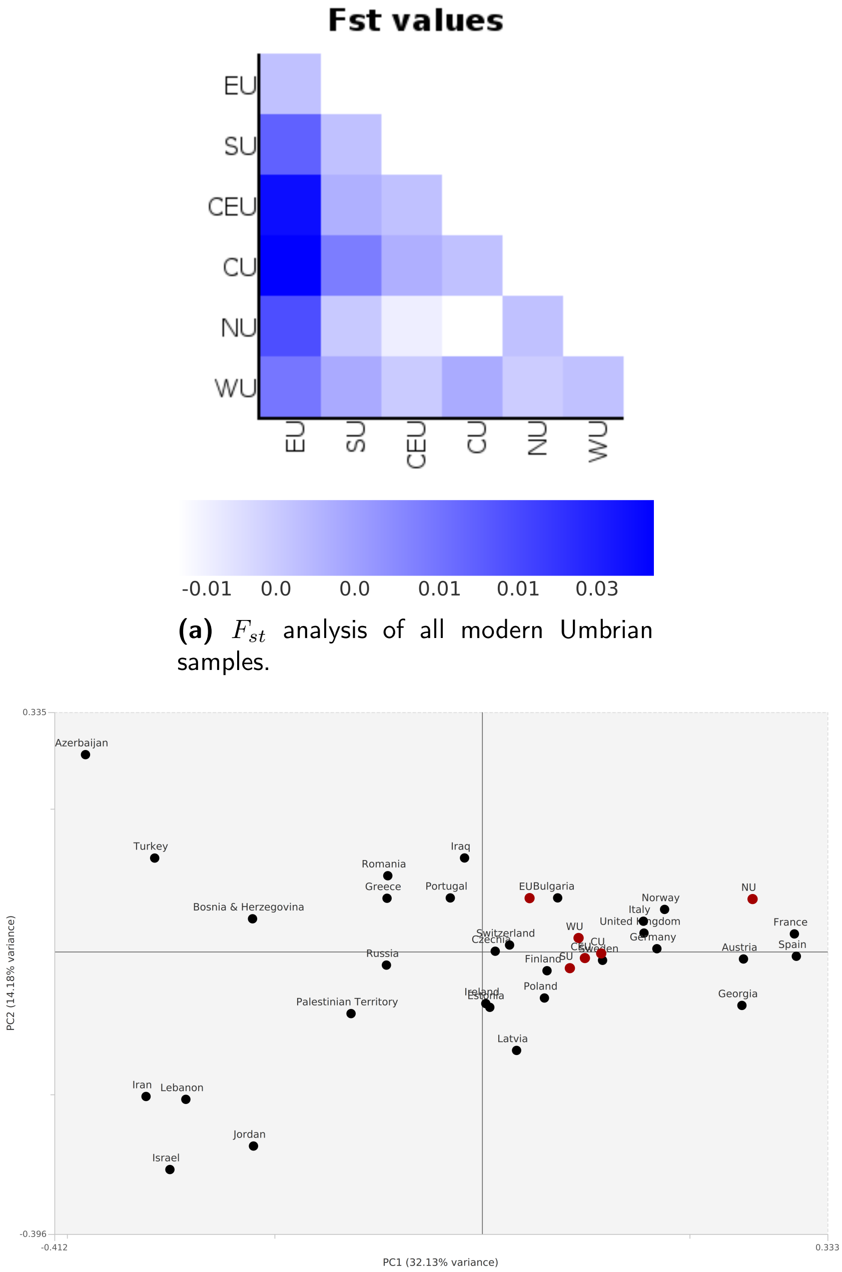
This chapter introduced mitoBench, a user-friendly workbench and well-curated database with several thousand complete modern and ancient human mitochondrial genomes. The main features used to achieve ease of use include drag-and-drop data import, interactive visualizations, and a clear design. With mitoBench, we



(a) Haplogroup distribution within ancient and modern data set.

(b) Haplogroup distribution of Umbria's geographic regions.

Figure 4.10: Evaluation of mitoBench by applying the provided visualizations on an already published data set [182]. The haplogroup distributions within the modern and ancient data set are shown using various visualization methods provided by mitoBench. The visualization as pie chart is shown in (a); (c) displays the haplogroup distributions within the geographic regions of Umbria (Grouping: North Umbria (NU), West Umbria (WU), South Umbria (SU), Central Umbria (CU), Central East Umbria (CEU), and East Umbria (EU)) as stacked bar chart.



(a) F_{st} analysis of all modern Umbrian samples.

(b) PCA plot based on a data set of all modern Umbrian samples (red) and comparative Eurasian genomes.

Figure 4.11: Evaluation of mitoBench by applying the provided analysis methods on an already published Umbrian data set [182], grouped by location: North Umbria (NU), West Umbria (WU), South Umbria (SU), Central Umbria (CU), Central East Umbria (CEU), and East Umbria (EU). (a) Result of the F_{st} analysis of this data set. (b) PCA analysis based on an Eurasian data set (black dots), including the Umbrian data (red dots). The analysis was performed on the haplogroup frequencies, derived from complete mitochondrial genomes.

intended to generate software that can be used to investigate human complete mitochondrial genomes, file conversions, and, most importantly, collect comparative data set from the database. Combining modern and ancient data in one central database and the directly connected work platform offers a unique opportunity for mtDNA analysis. Comparable existing databases focus either exclusively on ancient data, such as AmtDB [55], or on mitochondrial pathology in human diseases, such as mitoDB [166] and mtDB [54], which also do not offer further analysis methods. Other databases, such as Hmtdb [53], was created to support population genetics and mitochondrial disease studies. While Hmtdb provides a large amount of published complete human mitochondrial genomes (51,302 complete mtDNA genomes, accessed July 14, 2021), only a limited amount of meta-information is available, such as sex, age, or geographic origin. Information beyond this, such as language, is not available. Neither is it possible to download the genomes with their meta-information; only the genome alignment is available. Moreover, its toolbox, MToolBox, focuses on calculating and investigating variant positions and does not provide any further downstream analyses and visualizations besides haplogroup assignment and functional annotation.

This database intends to offer researchers a central mtDNA repository with easy access to published modern and ancient complete mitochondrial genomes alongside all-important contextual meta-information. It also provides the possibility to compare own and private data with this reference data set of ancient and modern genomes, which to our knowledge is the only such database, and allows users to perform typical analytical procedures much faster, more reliably, and more conveniently than before.

Compared to most existing data collections, mitoBench focuses on complete mitochondrial genomes. For successful upload, the chosen genomes fulfill high-quality criteria. Moreover, mitoBench focuses on population genetics and provides topic-specific relevant meta-information in addition to the high-quality mtDNA sequence itself. In addition, the genomes can be called up directly from the workbench, linked with metadata, and processed further without file formatting. This allows the user to include unpublished data in the analysis without registration easily. To ensure reproducibility, each analysis step and all settings can be viewed in the log file. Also, the upload of data is entirely automatic, given the constantly growing amount of data.

In summary, mitoBench is a concept that provides researchers with state-of-the-art methods to perform integrative and reproducible analysis tasks in the context of mitochondrial population genetics. Future enhancements to the current version will further improve the capabilities of both applications, the workbench, and the database.

CHAPTER 5

The detection of pathogens in various tissues using metagenomic screening methods

5.1 Introduction

With the introduction of next-generation sequencing and the resulting reduction in sequencing costs, it has become possible to sequence metagenomic samples on a larger scale, not only for modern but also for ancient samples. Over the last years, the microbiome of different body parts, such as the human gut [185], the oral cavity [186], or skin [187], have come into focus. As the microbial composition in modern samples differs between a healthy and diseased state, this information can also be transferred to ancient samples and provide insights into the health status of individuals in the past. Although ancient samples are often poorly preserved, several studies could successfully reconstruct the ancient microbiome of different body sites [137, 138, 188], and especially dental calculus has proven to be an excellent source for human [25, 189, 190] and microbial DNA [25, 191] in ancient samples. Along with this analysis, DNA from pathogens is also often detected and reconstructed. This chapter describes two applications of metagenomic analysis of so far little attended tissue: the tissue of mummified Egyptian individuals (Section 5.2) and a human leg sample preserved in ethanol and paraffin (Section 5.3). These applications also demonstrate the successful retrieval of DNA from various pathogens and their genome reconstruction. Finally, the chapter concludes with a discussion of the individual applications and an overall summary.

5.2 Application 1: 2,000 year old pathogen genomes from Egyptian mummified individuals

Text and figures in this chapter were adapted with modifications from our work published in BMC Biology [192].

5.2.1 Introduction

With the work on the Tyrolean Iceman and the accompanying reconstruction of a 5300-year-old *Helicobacter pylori* genome [193], the analysis of mummified remains came to the fore. Archaeological tissues offer the possibility of studying the metagenomic composition of various soft tissues in order to detect specific pathogens, as well as reconstructed genomes. In contrast to natural or spontaneous mummification, anthropogenically mummified individuals are intentionally preserved, which is a typical burial rite practiced in Ancient Egypt. Besides this unique way to preserve the dead, it is a scientifically interesting place to study. Its location on the Isthmus of Africa and the extensive trade links with the Levant, sub-Saharan Africa, and even European countries have resulted in the admixture of human populations and opportunities for the spread of pathogens. While several studies on ancient and modern human DNA were recently performed [119, 194], no systematic analysis of the metagenomic content of mummified tissue from mummified Egyptian individuals or possible pathogens has yet to be conducted. One of the reasons for this was the lack of appropriate sequencing technologies and bioinformatics methods to handle the specific characteristics of aDNA. Moreover, especially the analysis of ancient Egyptian mummified individuals was considered highly challenging, if not even impossible. The DNA is highly degraded due to the hot and humid conditions in many tombs and additionally destabilized by chemicals used during the embalming process [56, 57]. Recent advancements in sequencing technologies and analysis methods made it possible to extract DNA from such troublesome samples, as previously shown from several studies [119, 195]. However, the main focus of these investigations was on the human genetic content, while systematic metagenomic studies had yet to be performed.

The data set considered in this study includes 119 mummified individuals from the site Abusir el-Meleq, located 200km south of Cairo. One hundred thirty-three samples from different tissues were analyzed, namely bone, oral (tooth and calculus), and soft tissue. All samples were radiocarbon dated and fall between the First Intermediate (2196-2045 BCE) and the Roman Period (386-426 CE). While the human genetic content of 90 individuals in this data set was previously examined by Schuenemann *et al.* [119], this study focuses on the metagenomic compositions of the different tissues and periods, particularly the oral microbial composition, and whether such samples permit the assessment of pathogen composition as well as genome reconstruction.

5.2.2 Methods

This section provides a summary of the methods - a detailed description is available in Schuenemann *et al.* [119] and Neukamm *et al.* [192].

Sample information, extraction, and library preparation

As different tissues were available, 133 samples from 119 individuals were collected from bones, teeth, soft tissues, and dental calculus. All samples were processed in the cleanroom facilities at the University of Tübingen and at the University of Zurich, both of which are designed explicitly for ancient DNA processing [196].

For shotgun sequencing, the samples were processed as described in [119]. Double-stranded libraries with sample-specific dual barcodes were prepared using well-established protocols [197–199] as previously described in Schuenemann *et al.* [119].

Due to the positive signal of *M. leprae* in sample Abusir1630b, additional libraries for deeper sequencing, as well as UDG-treated libraries were prepared and sequenced on an Illumina HiSeq4000 platform with 2x75+8+8 cycles. The same applies to the samples Abusir1543s and Abusir1543b, where a hepatitis B virus (HBV) was detected. However, only non-UDG libraries were prepared.

To assess the ancient oral microbiome, five additional calculus samples were processed as described in [192]. After irradiation with UV light, the samples were ground, DNA sequencing libraries prepared, and sequenced on an Illumina NextSeq500.

Data preprocessing

The preprocessing step is based on the raw sequencing data provided as a FASTQ file. The quality is assessed using FastQC [200], which gives a general overview of the number of reads sequenced and the sequencing quality. Depending on the sequencing machine, DNA fragments of a fixed length are produced. As described in Section 2.1, two adapter sequences are attached to both ends of the fragment before sequencing. Due to the fragmented nature of aDNA, the adapters may be completely or partially sequenced; thus, AdapterRemoval2 [201] is used to remove adapters. Afterward, reads below a certain length are filtered. Finally, the reads from paired-end sequencing are merged if the overlap of the two reads is larger than 11 base pairs, which was also done using AdapterRemoval2. The resulting reads are then used for all subsequent analyses.

Metagenomic screening

After adapter trimming and read-merging, the merged libraries of all 133 samples were used for metagenomic analysis by mapping them against a reference database from GenBank consisting of all complete bacterial, viral, and archaeal genomes

(May 2018 version). The mapping and taxonomic analysis was performed using MALT [101] and further inspected and visualized with Megan6 [202]. MALT was executed with default parameters, except the following: Only reads with a minimum 85% identity (–minPercentIdentity) were considered as a possible match to the reference. Moreover, the minimum support parameter (–minSupport) was set to 5, i.e., only nodes with minimum support of five reads were kept. BlastN mode and SemiGlobal alignment were applied, and a top percent value (–topPercent) of 1 was set.

Endogenous DNA content

Based on the taxonomic assignment, the microbial communities within the samples were determined by comparing them to previously known communities using SourceTracker2 [102]. This method compares the bacterial composition in the input sample with a set of bacterial reference communities and determines the percentage of these present in the input sample. For assessing the microbial communities, modern oral and calculus [203–205] samples were used, in addition to a relevant soil sample from near Abusir al-Meleq [206], as soil samples from the site were unavailable or not collected during excavations.

Ancient genome reconstruction

For this project, complete genomes of three species of interest, namely *M. leprae*, HBV, and the human mitochondrial genome were reconstructed using the EAGER pipeline [23]. This pipeline provides a fully automated way of mapping reads from high-throughput sequencing against a specific reference genome. EAGER was built for - but is not limited to - the reconstruction of ancient genomes and therefore allows the preparation of ancient genomes and the modern comparative data set simultaneously. It follows the GATK best practice guidelines [103] and the parameters optimized for ancient DNA are set as default [22, 104]. The pipeline consists of three components: preprocessing, read mapping, and genotyping. As the preprocessing was previously completed as described above, this step is not detailed here, and the previously processed data were used as input for the genome reconstruction.

READ MAPPING The preprocessed sequencing data of all 133 were mapped against the three reference genomes: *M. leprae* TN chromosome (RefSeq ID: NC_002677.1), the mitochondrial genome of *Homo sapiens*, and HBV (Genbank ID: AY738142.1)) using CircularMapper [23] with a minimum quality score of 20 and a maximum hamming distance of $n = 0.2$. Mapping describes the process of comparing the preprocessed DNA fragments to the reference genome and identifying the position within the genome to which the DNA fragment belongs. The standard method in aDNA analysis is BWA aln [145]. This method works well for linear genomes, e.g., the nuclear genome. However, most genomes investigated in this thesis have a circular structure, e.g., bacterial genomes or the human mitochondrial genome. To also handle the circular structure, CircularMapper has been developed and is part of the EAGER pipeline [23]. This method uses BWA aln as underlying mapping software and can properly map DNA fragments to circular genomes, especially to the artificially introduced point of intersection.

ANCIENT DNA AUTHENTICATION After mapping, the ancient origin of the resulting reads can be verified based on the damage patterns typical of aDNA. The calculations were performed using DamageProfiler ([133], Chapter 3) and include the length distribution, hamming distance, and the frequency of the C to T base substitution. To further verify that the bacteria comes from the same time period as the mummified individual, i.e. is ancient, their damage patterns were compared with those obtained from the mapped human reads of the same sample. Assuming that both are coming from the same time, the patterns should be coherent.

VARIANT CALLING AND SNP EFFECT ANALYSIS The next and last step is variant calling using GATK UnifiedGenotyper [103, 207], which entails identifying SNPs and small InDels from next-generation sequencing data. The result is presented as a VCF file containing each specified position, the reference, and the allele call. Based on the created VCF file, MUSIAL [208] can be used to investigate further and process the variant calls, such as calculating SNP alignments for a given set of genomes and additional statistics. The applications described in this work were set as follows: the reference base was called if the position was covered at least three times, and the quality score was at least 30. The base was called an SNP if the quality score was at least 30, and 90% of the mapped reads contained this variant.

Moreover, the functional effect prediction software SnpEff [209] can be called directly within MUSIAL but also independently, which categorizes the effects of variants and predicts coding effects based on their genomic locations.

Phylogenetic analysis and estimation of divergence times

For *M. leprae*, all published and newly reconstructed genomes with a coverage of at least 5X of 80% of the genome were included in this analysis. MEGAX [210] was used to calculate a maximum parsimony tree using partial deletions (cutoff of 80%) and 1000 bootstraps. Besides, a maximum likelihood tree was generated using PhyML [211] with 100 bootstraps and optimizing tree topology, branch length, and rate parameters. Based on the temporal signal that could be detected using TempEst [212] and a date randomization test, the analysis of the divergence time and substitution rate was performed using the Bayesian framework BEAST version 2.5.0 [213] based on 161 modern and ancient strains, which is represented by the data set after excluding all hypermutated strains. In these strains, there is an unusually long branch but with roughly the same number of InDels and deleterious mutations in gene *nth*, an endonuclease III gene [214]. In addition, all SNPs belonging to known repeat regions and rRNA, as well as the positions covered by the negative control sample SK12 [27] were excluded.

For HBV, two different approaches were used for the phylogenetic analysis: a network structure using SplitTree version 4.15 [215] and a maximum likelihood tree, calculated using PhyML version 3.1 [175] and 100 bootstraps. The tree-like network is based on an alignment of 511 modern and ancient HBV genomes, as used and established in a previous study [29], using the parameter NeighborNet [216] with uncorrected P distances. The maximum likelihood tree is based on a selected subset of the alignment described above, representing all HBV genotypes. It consists of all ancient and 111 modern human and non-human primate HBV genomes [29–31, 217], as used in a previous study [29], and our new genome Abusir1543. The sequences were aligned with MAFFT version 7.407 [173] using the linsi algorithm.

Recombination analysis

As recombination is known to occur in HBV, a subset of 52 representative genomes, including all ancient strains, Abusir1543, and one strain per sub-genotype, was used for recombination analysis. Recombination is defined as the exchange of genetic material between organisms (Section 2.2). There are five commonly accepted methods for detecting recombination in DNA sequences: similarity methods, distance methods, phylogenetic methods, compatibility methods, and substitution distribution. These concepts are implemented in various software. However, the software RDP4 [218] was used for this specific analysis as it is specialized in virus DNA. In particular, the methods RDP [219], GENECONV [220], Chimaera [221], MaxChi [222], BootScan (secondary scan) [223], SiScan (secondary scan) [224], and 3Seq [225] were run. The window size of 100 nucleotides and the circular genome without reference parameter were set.

5.2.3 Results

Metagenomic analysis

The metagenomic screening of all 133 samples from different mummy tissue resulted in a majority of reads mapping to the bacterial genus Clostridium, independent of the investigated tissue or time (Appendix, Figure A.2). This genus is known to be involved in biological tissue decomposition and can survive well due to the endospores they build [226, 227]. A closer investigation of these reads, especially the damage patterns, showed various frequencies of base substitutions within the mapped reads, indicating different ages of the DNA. Therefore, reads mapping to Clostridia have been excluded from further analysis, as they most likely originate

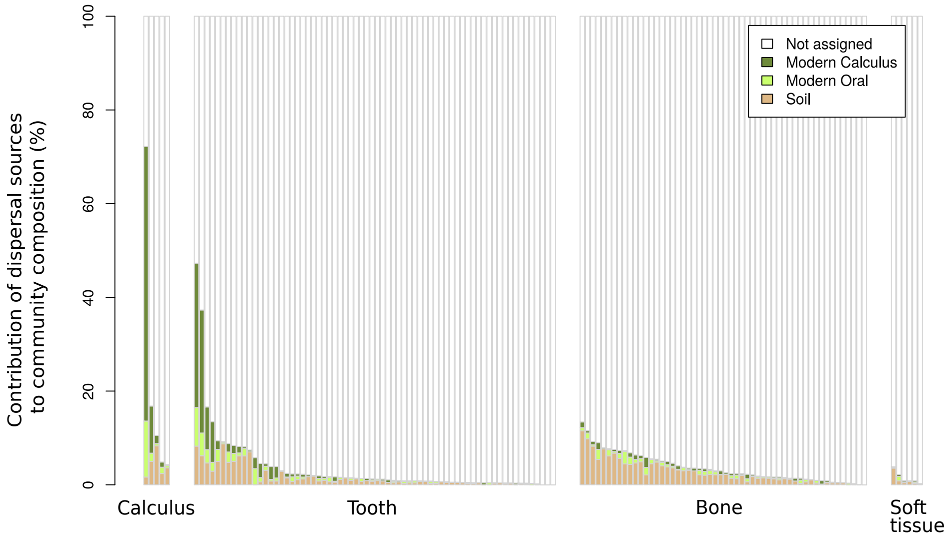


Figure 5.1: Community profiles of all microbes in bone, soft tissue, and oral samples. This diagram shows the proportions of soil (brown), modern oral tissue (light green), and modern calculus (dark green).

from environmental contamination or postmortem habitat of the analyzed tissue. Evaluation of the metagenomic sources [102] shows that most oral samples contain microbes expected in oral samples (Figure 5.1). Hardly any such signal could be detected in the other tissues (i.e. bone, soft tissue). In addition, there were many bacteria that had not been assigned. This is possibly due to the absence of comparative metagenomic profiles from different mummified tissues or decomposition materials. In the following sections, the different tissue is described separately. Within each tissue, samples were grouped according to the time period.

The investigation of the bone and soft tissue samples yields a general composition of the phyla Firmicutes, Actinobacteria, Proteobacteria, and Bacteroidetes (Appendix, Figure A.3a). Metagenomic analysis of these samples also revealed the presence of various pathogens. The first pathogen, *Proteus mirabilis*, was detected in four samples (Abusir1608b, Abusir1566b, Abusir1609b, and Abusir1645b) and is a possible agent of symptomatic infections of the urinary tract [228], but also known to be present in wound infections [229]. Furthermore, *Enterococcus faecalis* and *Enterococcus faecium* were detected in eight samples. These bacteria are typical inhabitants of the intestinal tract of healthy individuals. Still, they are also known to cause infections in humans, such as endocarditis and septicemia, and infections in the urinary tract [230]. Finally, *M. leprae* and HBV were detected and the analyses performed are described in detail in the following sections.

The ancient origin of all identified pathogens could be verified by showing the increasing frequency of base misincorporations at the ends of the fragment and the short fragment length using DamageProfiler [133]. In addition, the comparison of the damage patterns of the human DNA recovered from the respective samples yield similar patterns, also supporting the ancient origin. The ancient human DNA of these samples was already proven to be authentic in a previous study [119].

Mycobacterium leprae in Abusir1630

The metagenomic screening yielded the identification of a high number of reads in sample Abusir1630b that mapped against the genome of *Mycobacterium leprae*. Individual Abusir1630 was dated to 342-117 cal BCE, and the respective sample was taken from a bone. The excellent preservation of the sample allowed us to reconstruct the genome based on the shotgun sequencing without the need for additional enrichment for this specific pathogen. The genome was reconstructed based on the mapping against the reference genome of *M. leprae* (RefSeq ID: NC_002677.1) using EAGER [23]. About 97% of the genome could be reconstructed with a mean coverage of 35.35X and authentic damage patterns with a damage profile of 10.8 to 11.3 % and average fragment length of 44bp (Figure 5.2). The investigation of this sample mapping to the human genome resulted in comparable damage patterns, supporting the ancient origin. The analysis of the single nucleotide polymorphisms (SNPs) of the newly reconstructed *M. leprae* strain Abusir1630 yielded 3,342 informative SNPs. Among them, 47 missense, 41 synonymous, one stop-gain, and one stop-loss variant could be detected. Nine SNPs were exclusive to the leprosy strain Abusir1630, five belonging to non-coding regions, three intragenic variants, two synonymous, and one missense variant. The missense variant occurs on position 1,824,962 and affects gene 'dapA', which is involved in the synthesis of (S)-tetrahydrodipicolinate [231].

The phylogenetic analysis using a maximum likelihood and maximum parsimony approach (Appendix, Figure A.4 and A.5) resulted in a consistent placement of Abusir1630. The phylogeny of *M. leprae* is defined by six branches (0 to 5) and four SNP types (1 to 4), including 16 SNP sub-types (A to P). The two nomenclatures originated from two different research fields, including modern diagnostics (SNP types) [214, 232] and whole-genome investigations (branches) [27]. Although the former is constructed using a limited number of SNPs and provides less resolution, it is widely used in the leprae research field. The newly sequenced strain Abusir1630 falls within branch 4, clustering together with the modern strain S15, and was typed to be of SNP type 3L. This branch mainly contains modern strains from Africa, and one additional ancient genome from Czech Republic (Body188), dated to 800-1200 CE [28]. The short branch length indicates the genetic proximity of Abusir1630 to the most recent common ancestor (MRCA) of genotype 3L.

Due to the detectable temporal signal, a time-aware phylogeny was performed for the given data set using BEAST [213]. This result lead to an estimate of 5,844 years (y) before the most recent sample for the time to the most recent common ancestor (tMRCA). This points to a divergence time for the leprosy variation about 1300 years older than previously published [28], but overlapping with the confidence intervals once estimated [27, 28]. The divergence time of branch 4, estimated to 3,428 y, suggests an introduction of the strain in Egypt between 1410 and 117 BCE. However, additional strains of genotype 3L are needed to accurately determine the chronology of branch 4.

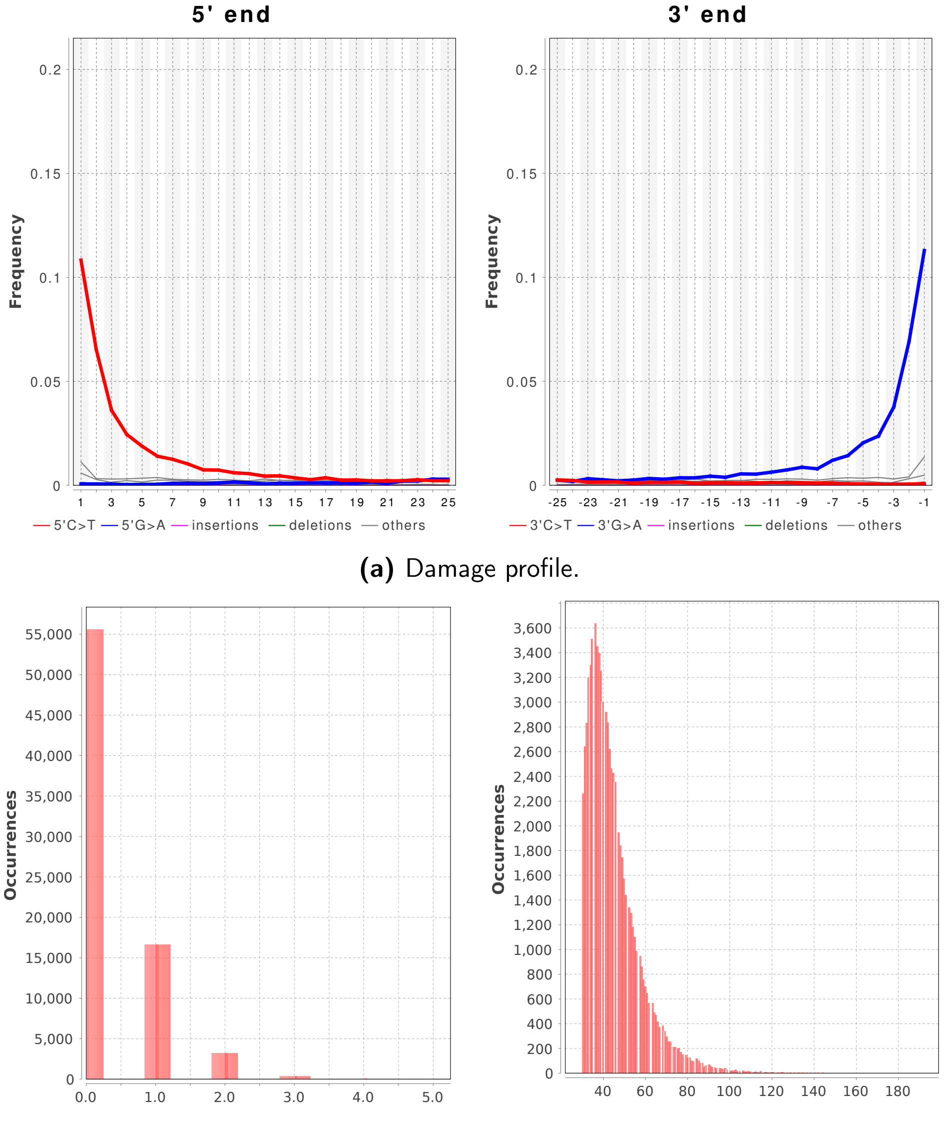
HBV in individual Abusir1543

Further investigations of the samples showed reads mapping to a hepatitis B virus reference genome in the bone and soft tissue of individual Abusir1543. As before, the excellent preservation of the sample allowed a genome reconstruction based on the shotgun sequencing without further enrichment. The mapping against the HBV genotype A reference genome (RefSeq ID: AY738142) yields 1594 mapped reads, 20.56X mean coverage covering 96.31% of the genome, and a damage profile of 10.4%, suggesting the ancient origin of the sample.

For the phylogenetic placement of strain Abusir1543, a whole-genome alignment of 511 modern and ancient HBV genomes was used to calculate a network (Figure 5.4) and maximum likelihood tree (Appendix, Figure A.6). Both the network and ML tree of HBV resulted in a consistent placement of Abusir1543 within genotype A. Besides several modern strains from Africa, this genotype contains six ancient strains from Eastern Europe, dating from around 1500 to 4200 years before the present. In detail, all ancient strains in this genotype fall basal to Abusir1543, which falls ancestral to genotypes A1 and A3, consisting of only modern African strains.

Additionally, the newly sequenced strain was tested for recombination due to known recombination events within the HBV genome [34]. Recombination causes mosaic sequences, where different patterns are inherited from one or multiple ancestral sequences called parents. The analysis of Abusir1543 resulted in two parents detected, namely a 4000-year-old strain RISE254 from Hungary as a minor parent and the genotype D strain extracted from a 1,500-year-old Italian mummified individual as a major parent. It is improbable that the strain Abusir1543 resulted from direct recombination of these strains; therefore, it was likely the ancestral stains of the two parents which were involved in the recombination.

A weak temporal signal could be detected for the HBV data set. Therefore, the divergence time estimation was investigated using BEAST [213]. The data set for this analysis consists of 128 modern and ancient strains as already used in previous analysis [30], as well as strain Abusir1543. The analysis estimated the tMRCA to 8923 BCE., thus matching previous results [30]. The tMRCA of Abusir1543 and genotypes A1 and A3 are estimated to 448 BCE. (2,467 y (2076-3021 y 95% HPD)). Confirming previous results [30], the mean clock rate is 1.30x10^{-5} substitutions per site per year (95% HPD interval: 9.99x10^{-6} - 1.61x10^{-5}) under this model. Moreover, the addition of the oldest genomes reconstructed so far did



(a) Damage profile.

(b) Edit distance.

(c) Length distribution.

Figure 5.2: Damage patterns of sample Abusir1630. All reads are from shotgun sequencing and mapped against the *M. leprae* reference genome.

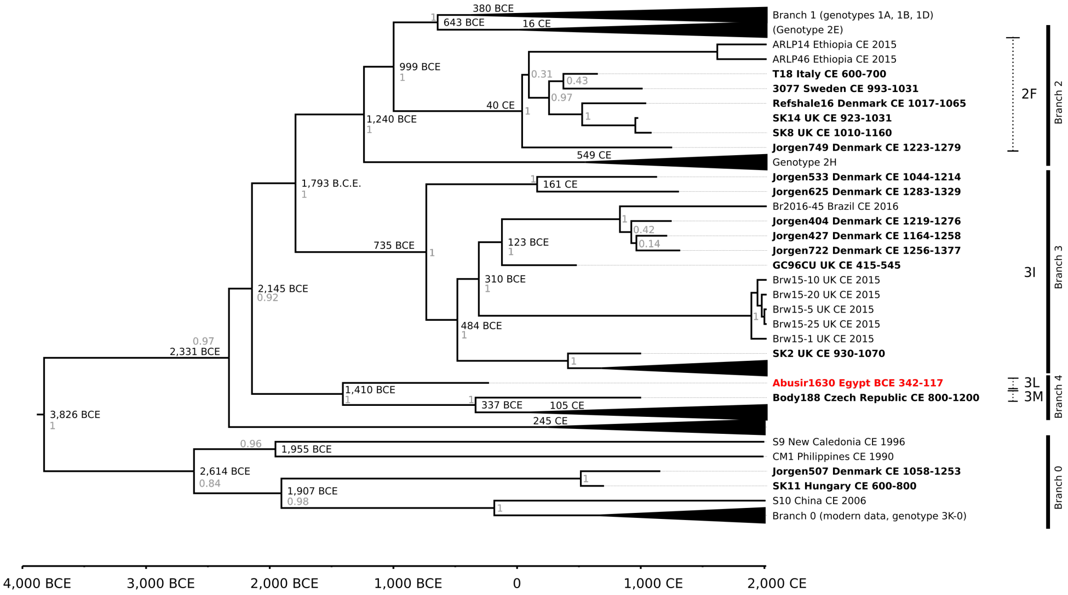


Figure 5.3: Based on 2641 informative SNP positions from 161 M. leprae samples, a Bayesian tree with maximum clade reliability was reconstructed. A strict molecular clock and the Bayesian skyline model were used. Ancient samples are in bold, and the newly added Abusir1630 genome is in red. Node labels are median divergence times in years BCE and CE. Posterior values are in gray. Genotypes are written in parentheses or marked with dotted lines. Branches are shown on the right side with black bars.

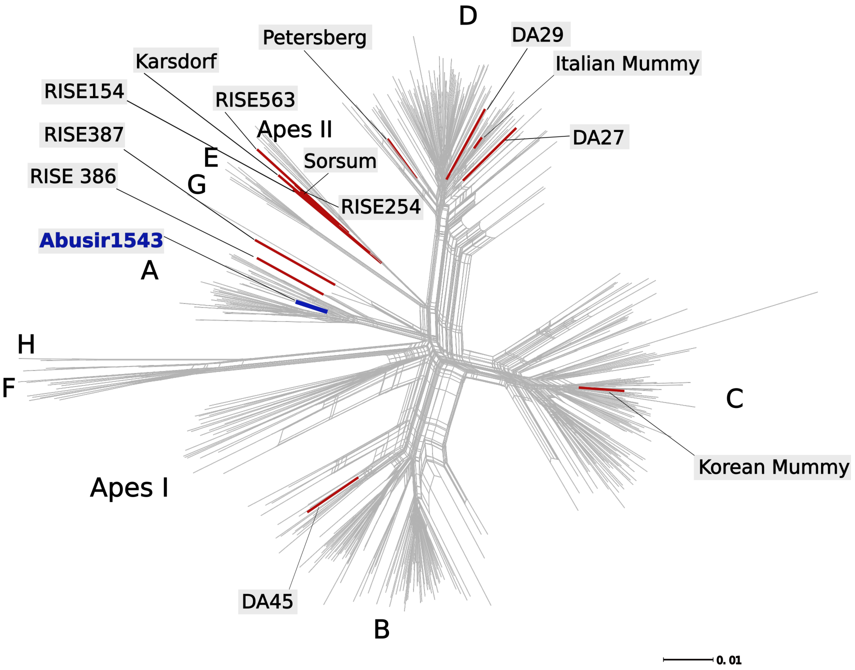


Figure 5.4: The phylogenetic network is based on 511 HBV genomes (Appendix, Table A.1). The newly sequenced genome is highlighted in blue, while previously published genomes are highlighted in red and labeled in black [29–31, 217]. The capital letters represent the different clades. The designation "Apes I" includes all strains of gibbons and orangutans, and "Apes II" includes the strains of gorillas and chimpanzees.

not influence the estimations. However, genetic dating is not expected to provide conclusive results because of the recombinations and mutations that occur. It can also be affected by crossing the barrier between humans and apes. [34–36].

Oral microbiome assessment

Since the data set used in this study was dominated by well-preserved oral samples, a more detailed picture of the microbial composition of the oral cavity was obtained. All identified ancient oral microbiome reads showed a reliable damage pattern (Appendix, Figure A.7) matching the profile of the mitochondrial reads of the corresponding individuals, further supporting the ancient origin. Contrary to the bone and soft-tissue samples, reliable modern comparative data sets were available for the oral microbiome, allowing a more in-depth analysis. First, for all oral samples, the composition of different microbial sources was assessed, and second, the general metagenomic composition on the phyla level. Lastly, a screening on a species level compared with the paleopathological investigations [233] provides an insight into the oral health status of specific individuals.

The investigation of the microbial sources contained in the five calculus samples yielded in a majority of unassigned species, which is not unexpected for ancient samples [25, 191]. Nevertheless, a high signal for modern calculus and modern oral communities in samples Abusir1519c (70.55%) and Abusir1594c (11.80%) provides a basis for further investigations. Considering the phyla level, the calculus samples were dominated by Firmicutes, Actinobacteria, Proteobacteria, Bacteroidetes, Chloroflexi, Fusobacteria, and Spirochetes (Appendix, Figure A.3). These results are consistent with already identified dominating bacteria in ancient calculus [25, 234]. Considerations at species level revealed the presence of fragments mapping to the bacteria of the Red Complex, a widespread bacterial complex consisting of *Tannerella forsythia*, *Porphyromonas gingivalis*, and *Treponema denticola* associated with periodontal disease [235]. Moreover, *Filifactor alocis* and *Olsenella uli*, which are associated with periodontitis and endodontic infections [236, 237], were also identified in two calculus samples.

As with the calculus samples, the majority of microbial species identified in the dental samples could not be classified to any of the sources (Figure 5.1). However, four samples (Abusir1580t, Abusir1650t, Abusir1614t, and Abusir1573t) showed a reasonable amount ($> 10\%$) of species that were associated with oral communities. In the remaining samples, the oral community represents between 0 and 5.8%. The samples are mainly composed of the phyla Firmicutes, Actinobacteria, Proteobacteria, Bacteroidetes, Chloroflexi, Fusobacteria, and Spirochetes (Appendix, Figure A.3), and on the species level, multiple bacteria connected to the oral microbiome have been detected. Fourteen samples were identified as harboring bacteria belonging to the Red Complex among others. Furthermore, nine tooth samples showed the presence of *Filifactor alocis* and *Olsenella uli*. *Streptococcus mutans*, which is known to be a significant contributor to tooth decay [238], could be identified in three samples.

Together with the paleopathological examinations already published for 18 of the tooth samples [233], a clearer picture of the general oral health of these individuals could be obtained. The comparison showed that for 13 samples, the genetic identification of oral disorders could be confirmed paleopathologically (Appendix, Table A.2). Two samples showed clear genetic evidence, while visual inspection showed no signs of lesions. In contrast, three samples showed clear paleopathological signs of oral lesions, while no oral pathogen could be detected genetically.

5.2.4 Discussion

This study analyzed a data set of 133 samples from mummified Egyptian individuals, spanning around 2000 years from the first intermediate to the Roman Period. The samples were taken from different tissues and the analysis was focused on the metagenomic composition of the various tissues, especially the oral microbiome, and the identification of pathogens associated with ancient Egyptians.

All samples showed a high amount of reads mapping to Clostridia with varying damage profiles. Their presence is not unexpected as Clostridia are known to be involved in the decomposition of biological material [226, 227]. Also, the varying damage profiles are a result of their presence in different stages of the decay and their ability to survive longterm, even in extreme climate conditions as in Egypt [239, 240]. The detection of microbial communities in all samples additionally showed a high amount of unassigned species, which is expected in ancient samples [191], and can be caused by various reasons. One reason is that these species belong to different communities, for example, the necrobiome,which describes the species composition harvesting a decaying corpse [241–243]. However, no comparative sample was available for this study as all studies on the necrobiome are based on 16s sequencing, making direct comparison difficult. Besides, the samples are metagenomic in nature, meaning that they also contain a variety of environmental DNA. Their study was challenging because no soil samples were available from this site at the time of the excavation. In addition, unassigned species may also be due to differences in the ancient and modern bacterial composition of different body parts and lacking knowledge of the actual composition of ancient communities, but also the different sample origin.

The investigations of time and tissue-specific metagenomic compositions of all samples were restricted by the lack of comparative data sets, especially for bone and soft-tissue samples. Also, the analysis of the general metagenomic composition on phyla level did not yield any detectable patterns specific to the time period or tissue, probably caused by the high environmental contamination of the samples. This is also reflected by the SourceTracker analysis that resulted in a high number of unassigned species. Although there were modern comparative samples for the oral microbiome, it also resulted in many unassigned species. This can be explained by the differences in the compositions between modern and ancient samples [25], or due to different populations, origins, health, or social background of the individuals, resulting in different microbial profiles. Aside from these challenges, specific oral pathogens could be detected in around 25% of the oral samples, showing authentic damage patterns, while 13 cases were confirmed through paleopathological analyses [233]. The differences between the genetic and the paleopathological analysis may come from an early stage of the infection, where lesions are not yet present, or vice versa. Moreover, multiple teeth were incomplete, so the bacterial could still be present in other teeth, while the infected ones have already dropped out. Moreover, there are also disease-associated species that are also contained in the healthy oral microbiome, such as Red Complex and *S. mutans* [186]. Nevertheless, the analysis showed that bacteria well-known in the oral microbiome of modern and ancient individuals with different backgrounds are also present in the studied individuals from ancient Egypt.

Additionally, the analysis of bone and soft tissue resulted in identifying five pathogens. However, only a minimal amount of reads could be detected, complicating any further investigation. Moreover, as both *P. mirabilis* and Enterococcus species usually are predominantly present in the urinary tract; miss-classification can not be excluded. In contrast, five of 12 individuals show clear damage patterns. These discrepancies can be explained by the relaxed mapping parameters chosen to also take the chemical modifications of ancient DNA into account or by environmental contamination.

The investigation of the bone and soft-tissue samples also resulted in the reconstruction of two complete pathogen genomes: a *M. leprae* and an HBV genome. The former was detected in a 2,200-year-old individual without showing any physical signs of leprosy infection, which is not untypical for an early stage of leprosy. This represents the first complete genome reconstructed from ancient Egypt. The only other record is of SNP type 3K/L/M and dated to the fourth to fifth-century [244, 245]. A detailed assessment of the SNP type was not possible due to poor DNA preservation. However, Abusir1630 (SNP type 3L) confirms the presence of this SNP type in ancient Egypt. As the first and so far oldest reconstructed genome from Egypt, Abusir1630 could help to understand the past and origin of leprosy. Currently, two hypotheses exist [28]: The first describes an origin in Western Eurasia, from where it has spread. The second assumes that *M. leprae* was present in a different region in the world and was then introduced into Europe before and during the Middle Ages. However, these models are only based on ancient data from Europe and modern strains. In addition to the introduction of strain Abusir1630 and its location in branch 4, modern strains from West Africa and Brazil as well as an ancient strain from the Czech Republic, suggest an origin in Eurasia, which is also supported by ancient European strains found in branch 0. However, neither of the hypotheses as of today can be favored. The dating analysis yielded a mean tMRCA around 1300 years older than previously published [28], dating to 3800 BCE, as well as a general shift of several hundred years back in the estimation of the divergence time in all branch splits, presumably due to the addition of the so far oldest strain Abusir1630. However, the 95% HPD intervals still show an overlap with the previous dates. Due to missing genetic investigations, the 4000 and 2600-year-old cases from India [246, 247], the 1500-year-old case from Italy [248], and the nearly 6000-year-old case from Hungary [249] could not be taken into considerations for the analysis of the divergence time. These cases were only confirmed osteologically without any molecular support. Nevertheless, these widespread records, especially the sample from Hungary that matches our tMRCA, suggest that leprosy's origin may go back further into the past.

The second fully reconstructed pathogen genome is a 2000-year-old HBV virus genotype A strain. HBV is the causative agent of human hepatitis, and the oldest genomes are dated to the Neolithic era [29]. The availability of bone and soft tissue of individual Abusir1543 allowed the comparison of the preservation of this specific pathogen in different tissues, resulting in better preservation in soft tissue. The virus has already been detected in two other mummified individuals [31, 217] and in different skeletal material [29, 30]. Accordingly, it cannot be concluded that a particular tissue is preferable for the recovery of HBV genetic material. The phylogenetic analysis showed that Abusir1543 falls within the genotype A clade, ancestral to genotypes A1 and A3, consisting of mainly modern African strains.

Basal to Abusir1543, two several thousand-year-old strains from Russia, a 2500-year-old strain from Hungary, and a 1500-year-old strain from Slovakia can be found. This confirms the hypothesis by Mühlemann *et al.* [30] that the ancestors of genotypes A1 and A3 originated in Eurasia and then migrated through Eastern Europe into Africa. While the phylogenetic support of the placement of Abusir1543 is excellent and consistent for all methods used, the overall support varies a lot and has to be considered with care. This instability of the phylogeny may be caused by the recombination that is known to occur in HBV. Therefore, an assessment of the divergence times most likely result in unreliable results. Although this analysis is done in several studies [250–252], they only considered selected genotypes or strains, where no recombination events can be detected. As the strains considered within this study show a positive temporal signal, a time-aware phylogenetic analysis was performed, yielding similar dates to previously published results [30]. However, due to the recombination events detected in strain Abusir1543 and the ongoing discussions about recombination events, and the lack of adapted methods, the dating results must be considered with great care.

5.3 Application 2: Variola virus genome isolated from an 18th century museum specimen

Text and figures in this chapter were adapted with modifications from our work published in Philosophical Transactions of the Royal Society B [253].

5.3.1 Introduction

Smallpox is known to be a highly contagious and lethal disease [254, 255], which caused several large scale epidemics [254, 255] before it was eradicated in 1980 CE [254]. The causative agent of smallpox is the Variola virus (VARV), a member of the genus Orthopoxvirus [254, 256, 257]. According various sources, VARV emerged around 3,000-4,000 years ago [258–261]. The first evidence of smallpox disease has been recorded in 1122 BCE China and 1500 BCE India and ancient Egyptian mummies dating to 1580-1100 BCE [254, 255, 262]. However, no genetic confirmation of these cases is available. The earliest reliable cases are ascertained in the 4th century CE China, 7th century CE India, and the Mediterranean, and 10th century CE southwestern Asia [254, 255, 263]. Nevertheless, with these few cases, the origin of VARV is still unclear.

Thanks to advances in ancient DNA (aDNA) research, a new understanding of disease origins and evolution can be gained by reconstructing ancient pathogen genomes [264, 265]. While the origin of various bacteria has been investigated and many genomes are already available (e.g., *Mycobacterium leprae* [27, 28] and

Yersinia pestis [127–129]), the research of ancient viruses is very limited. The attempts to sequence VARV genomes were of only limited success. In 2012, PCR fragments could be isolated from 17th-18th century CE Siberian mummies, which suggests a VARV origin approximately 2,000 years ago [130]. The first complete ancient genome has been reconstructed from a 17th century CE Lithuanian child mummy [32], followed by two specimens from the Czech National Museum in Prague dated to the 19th and early 20th century CE [33].

The divergence time estimate for a common ancestor of all VARV strains occurring in the 20th century CE yielded different dates covering a period between 1350 and 1645 CE [32, 33, 131, 266]. Consensus exists, however, that the point of divergence between the historical and 20th century AD occurring strains pre-dates the initiation of widespread inoculation in 1796 AD [254].

Furthermore, the time of common ancestry of clades PI and PII also appears to predate modern inoculation. The diversity within the modern clades does not appear to have arisen until the late 19th or early 20th century AD. These results suggest that the predominant VARV strains at that time were going through a severe bottleneck. At that time (the turn of the 20th century AD), worldwide smallpox vaccination programs began [254], causing the extinction of several older lineages [32].

By the 18th century AD, smallpox was endemic in Europe [267], with both epidemic frequency and mortality increasing [268–270], particularly among children [271]. To contribute to this period, this study reconstructed an 18th century CE VARV genome derived from a museum specimen of a child's leg made by the surgeon and anatomist John Hunter (1728-1793 CE) between 1760 and 1793 CE [272]. The phylogenetic analysis of the newly reconstructed historical VARV genome and all available historical and modern genomes results in a recent common ancestor dated to the 17th century CE, and suggests greater VARV diversity in the past than previously thought. This study also demonstrates how increasing the number of available historical or ancient genomes can lead to better resolution of phylogenetic inferences. In addition, the metagenomic composition of the conserved tissue was investigated.

5.3.2 Methods

The methods are described in more detail in [253] and the Section 5.2.2, and are only briefly depicted here.

Sample information, extraction and library preparation

The sample RCSHC/P 328 was taken from an ethanol-fixed infant leg embedded in liquid paraffin, stored at the Hunterian Museum at the Royal College of Surgeons of England. The DNA extraction followed a modified protocol from Devault *et al.* [273] and was conducted in the cleanroom facilities at the University of Zurich.

Six double-indexed sequencing libraries were generated from 10 μl of extract and extraction blank following a protocol optimized for aDNA [198, 199]. The sequencing was performed on the HiSeq2500 and HiSeq4000 Illumina platforms with 2x125+7+7 or 2x75+8+8 cycles, respectively, by the Functional Genomics Center Zurich (Switzerland).

Metagenomic screening & Genome reconstruction

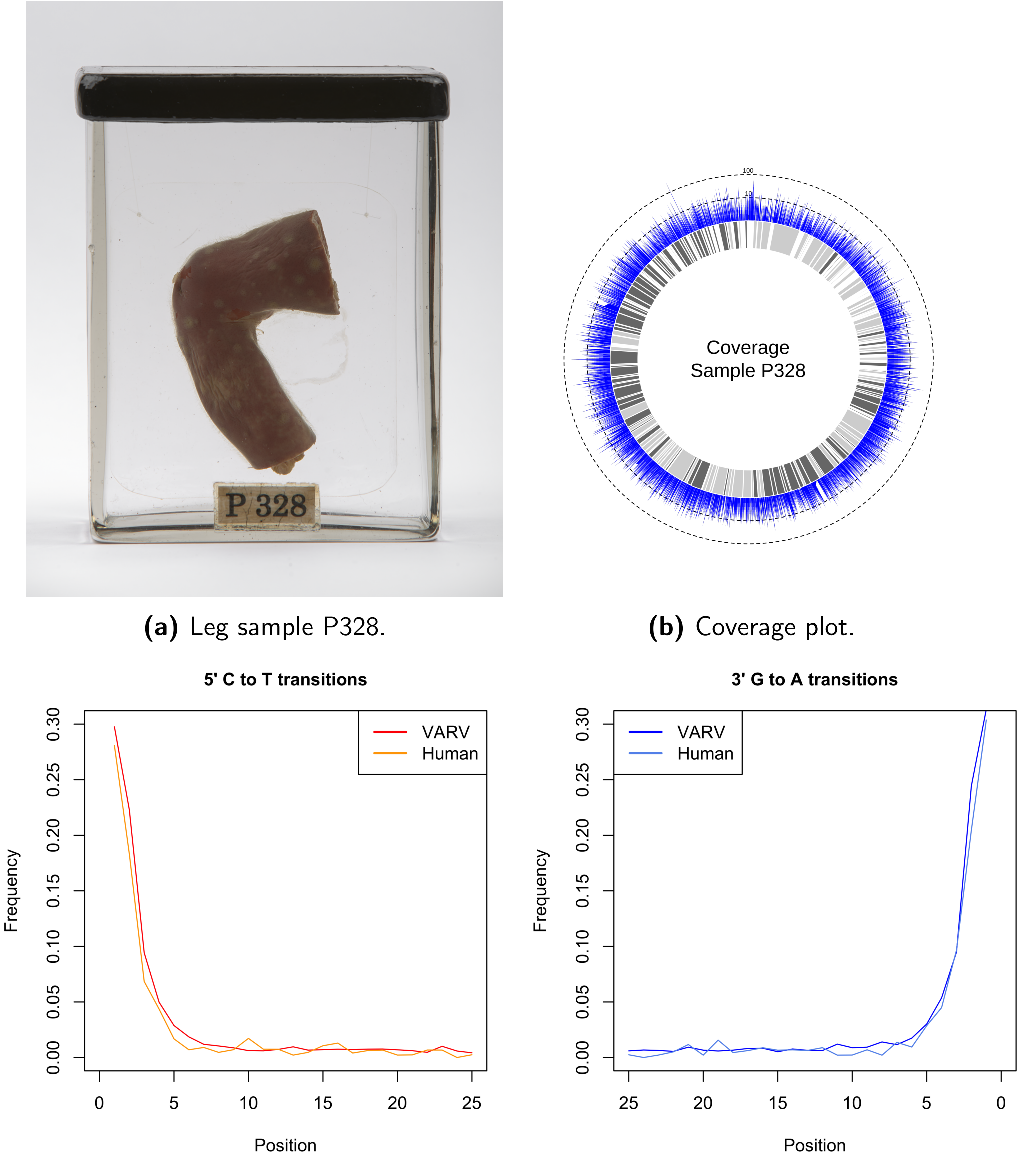
The metagenomic screening of all six libraries was conducted using MALT [101] with the parameters described in Section 5.2.2 and processed using MEGAN6 [202].

The VARV genome was reconstructed using the EAGER pipeline [23] using the parameters and methods as described in Section 5.2.2 with a variation in the mapping parameters. The processed sequencing data were mapped against the reference genome of VARV (RefSeq ID: NC_001611.1) using CircularMapper [23] with a minimum quality score of $q = 37$ and a maximum hamming distance of $n = 0.01$.

Phylogenetic analysis and estimation of divergence times

The whole-genome alignment, which serves as the basis for the phylogenetic analyses, consists of 57 genomes: 44 modern and three historic publicly available VARV genomes [33, 37, 266, 274, 275], as well as eight additional Orthopoxvirus genomes (camelpox virus, taterapox virus, cowpox virus, horsepox virus, monkeypox virus, raccoonpox virus, skunkpox virus and volepox virus) [37, 276–280] and a horsepox virus genome reconstructed from a vaccine manufactured in 1902 CE [281]. The sequences were aligned with MAFFT version 7.407 [173] using the FFT-NS-2 algorithm. Based on the resulting alignment and using all sites, a maximum-likelihood tree was calculated using RAxML version 8 [174] with 100 bootstraps.

To investigate the divergence times, BEAST2 [168] was applied to the data set described above, excluding the additional Orthopoxvirus genomes. Beforehand, due to controversial dating information, the tip dates of two historical strains, V563, and V1588, were estimated based on the time interval indicated by the original publication [33] and the corresponding comment [266]. A uniform distribution was used for strains P328 (1760-1793 CE) and VD21 (1643-1665 CE), and a normal distribution was used for V563 (mean: 1925, standard deviation: 20) and V1588 (mean: 1929, standard deviation: 60). Based on the resulting tip dates and the isolation dates of all modern strains, the divergence times and substitution rate have been estimated by using a strict molecular clock, the K81 substitution model (bModelTest [282]) and assuming constant population size [283] as tested best previously [32].



(a) Leg sample P328.

(b) Coverage plot.

(c) Damage profile.

Figure 5.5: Overview VARV sample P328. (a) Vessel containing specimen RC-SHC/P 328. (b) Coverage plot of the newly reconstructed VARV genome from sample P328. Blue indicates the coverage at a particular position (outer ring), coding areas of the genome are in grey (inner ring, the shades of grey indicate forward and reverse strand). The dotted circles indicate the coverage. (c) Damage profile of all reads mapping to VARV and human MT genome.

5.3.3 Results

Metagenomic analysis

The sample investigated in this study was collected from a museum specimen with a known diagnosis of smallpox, specimen RCSHC/P 328 (later referred to as P328, Figure 5.5) from the Hunterian collection at the Royal College of Surgeons of England. The sample, an infant's leg, is covered with smallpox lesions, and it is believed that the infant contracted the disease *in utero*. Following DNA extraction, six DNA libraries were generated for shotgun sequencing, obtaining approximately 150 million raw reads.

All reads mapping to the human genome were excluded to assess the metagenomic content of all six P328 and non-template control libraries. The remaining reads were mapped using MALT as described in Section 5.2.2. On average, 5.81% of the sequenced DNA could be assigned to the genomes of the reference database. The high number of unassigned reads may come from thus far unsequenced environmental bacteria. Overall, the P328 libraries are composed of viruses (53.77%), followed by bacteria (45.47% on average) and archaea (0.76% on average). While a background-specific metagenomic pattern could not be detected, the metagenomic analysis also yields a high amount of reads mapping to Poxviridae (6.85% - 24.38%), especially to the VARV genome (NC_001611.1) in all libraries except the non-template controls (Appendix, Figure A.8).

Ancient genome reconstruction

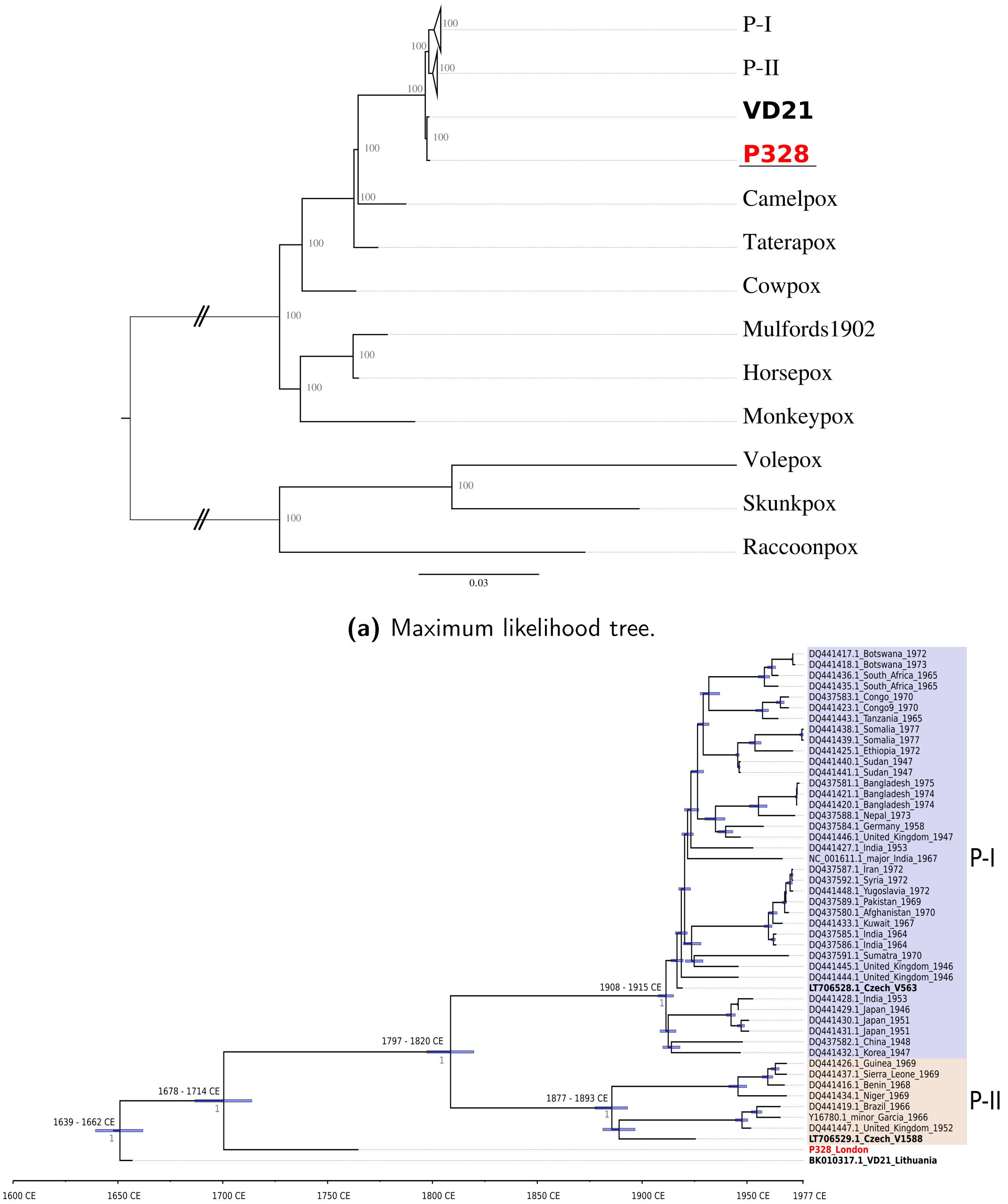
After combining all P328 libraries into a single file, all DNA fragments were mapped against the VARV (NC_001611.1) and the human (GRCh37.p13) reference genome using EAGER [23], respectively. The mapping against the human genome (resulting in 62,250,656 reads) yield the reconstruction of 95.42% of the human mitochondrial genome with a mean coverage of 19.62X and a damage profile of 28.07% to 30.68% (Figure 5.5b), verifying the ancient origin of the sample. This is supported by schmutzi [134], resulting in 1% of modern-day human contamination. The haplogroup was determined using HaploGrep2 v2.1.19 [164] and results in haplogroup H3b1b1, which is common in Europe. The mapping against the VARV reference genome resulted in 56,549 unique reads, and 85.18% of the VARV genome was covered with a mean coverage of 14X. (Figure 5.5c). The analysis of reads mapped to the VARV reference revealed a damage profile of 29.76% to 31.32% (Figure 5.5b). As the damage profile of the reads matches that of the human mitochondrial genome, this further confirms their ancient origin.

Phylogenetic analysis and estimation of divergence times

Based on a maximum likelihood tree, the newly reconstructed strain could be placed basal to all modern VARV strains, along with the historical strain VD21

[33, 131], as already shown previously [32, 33, 131] (Figure 5.6A, the complete tree is shown in Appendix Figure A.9). Furthermore, the root subdivides the smallpox strains into new- and old-world orthopoxviruses [277, 284].

A closer investigation of the dating was done for the Czech samples (V1588 and V563) due to discrepancies [33, 131], which could be solved with BEAST 2.5.2 [213]. The dates were estimated to 1925 CE for V1588 and 1920 CE for V563. Subsequently, the divergence times of all viral genomes were calculated (Figure 5.6B, Table 5.1). Also, the analysis estimated a mean evolutionary rate of VARV to be 10.67x10^6 nucleotide substitutions per site per year. The newly sequenced strain P328, dated to 1766 CE, falls basal to all modern human strains [37, 274, 275], as well as to the historic Lithuanian strain VD21 [32, 131] dated to 1656. The tMRCA of all used genomes is estimated at 1651 CE (1639-1662, 95% HPD). The median divergence time of P328 and the modern strains is dated to 1701 CE (1687-1714 CE, 95% HPD). In agreement with previously published results, V563 and V1588 fall into the modern P-I and P-II clades, respectively [33]. For testing the reliability of the estimate of divergence time, the BEAST analysis was reran excluding the strains V563 and V1588, producing near identical divergence times (Appendix, Figure A.10)



(a) Maximum likelihood tree.

(b) Dated Bayesian tree.

Figure 5.6: Phylogenetic analysis of VARV. (a) Maximum-likelihood tree derived from 57 Orthopoxvirus genomes. The historical genomes are in bold, the newly sequenced genome is in red and underlined. Bootstrap values are indicated as node labels in gray (100 BS). (b) Dated Bayesian maximum clade credibility tree calculated with BEAST 2.5.2 [168]. Nodes are labeled with the 95% HPD interval. Historical genomes are in bold; the newly added genome is in red and underlined. Posterior values are indicated as node labels in gray.

Table 5.1: Time to the most recent common ancestor (tMRCA) comparison of the dated VARV tree and individual branches from this study and previously published studies [32, 33, 131]. (Dates are given in calendar years (CE). HPD, highest posterior density.)

Branch splits, (CE)	This study		Duggan et al. [32]		Pajer et al. [33]	Smithson et al. [131]	
	mean tMRC	95% HPD	mean tMRCA	95% HPD	mean tMRCA	mean tMRCA	95% HPD
Split VD21/P328/ modern VARV	1651	1639–1662	1617	1588–1645	1350	1517	1470–1563
Split P328/ modern VARV	1701	1687–1714	na	na	na	na	na
Split P-I/P-II	1809	1797–1820	1764	1734–1793	1695	1623	1579–1667
Split P-I internal	1911	1908–1915	1910	1902–1917	1887	1881	1861–1897
Split P-II internal	1886	1877–1893	1870	1855–1885	1808	1794	1754–1828

5.3.4 Discussion

This study presents the successful reconstruction of an 18th century CE VARV genome at a mean coverage of about 14X using shotgun metagenomic data. Besides a capture-bias-free reconstruction of the genome, shotgun sequencing also has the advantage of analyzing the sample's metagenomic composition by comparing it against a reference database. For sample P328, this analysis yield predominantly plant-infecting viruses, such as the Dasheen mosaic virus, among the top ten hits. However, the damage profile of the reads mapping to these species does not show the expected damage profile. Also, a background-specific bacterial composition could not be identified. A high proportion of species for which sequences are unknown and preservation treatment of the samples in ethanol and paraffin could explain this, along with the absence of studies of the metagenome content of these preservation methods. Furthermore, to account for DNA damage that occurs with age, we applied a loose identity parameter during mapping. Nevertheless, the VARV genome was still mapped largely by reads whose ancient origin can be verified by a reliable damage profile (Figure 5.5C).

The phylogenetic analysis placed strain P328 in a sister group to all 20th-century VARV strains together with VD21, which is consistent with previously published results [32]. The common ancestor for all VARV strains, including P328, is dated to between 1639 and 1662 AD. This is consistent with the suggested common ancestor between modern strains and VD21, which is dated to AD 1588 to 1645 [32], both younger than the common ancestor suggested by Pajer *et al.*, which is dated to AD 1350 [33].

Similarly, our phylogenetic analysis calculates the divergence time of the PI and P-II clade between 1797 to 1820 CE, of P-I between 1908 to 1915 CE, and P-II from 1877 and 1893 (Figure 2B). Although these dates are slightly earlier than those reported by Duggan *et al.* [32], they are consistent with the hypothesis that the PI and P-II clades diverged before or during the smallpox vaccination in 1796 CE. This is also consistent with the study that the currently available VARV genome from the 20th century CE may represent only a tiny part of the genetic diversity of VARV in the past, as some strains are thought to have disappeared or are no longer detected due to loss of virulence [267]. Of the four historical strains included in our study, the younger V563 and V1588 [33] fall into the variety of modern VARV strains and group themselves in the P-I and P-II clades, respectively. In contrast, the 17th century VD21 [32, 131] and the 18th century P328 each form their sister group to the 20th-century strains. This observation is consistent with modern VARV strains not being representative of past viral diversity and is congruent with the scenario that selective pressure from increasing vaccinations led to the disappearance of several VARV lineages [32, 267]. However, obtaining more ancient and historical VARV genomes is crucial to capture the true diversity of VARV before the development of smallpox vaccination.

Finally, we addressed the discrepancies in the dating of the strains V563 and

V1588, which were dated by Pajer *et al.* to 60 and 160 years, respectively [33]. However, this was disputed, and an improved dating was published [266]. This dated both to 1920s CE. The dates of the study were re-examined. For this purpose, a tip calibration with normal distribution was taken out around the proposed younger dates. Furthermore, the possibility that the samples have the original proposed age was considered. These analyses resulted in a mean date of 1920 AD for V563 and 1925 AD for V1588. This is consistent with the dates proposed by Porter *et al.* [266]. Furthermore, our proposed 17th century CE dating of the common ancestor between modern and historic strains is much closer to the 16th-17th century CE dating presented by Duggan *et al.* [32], and Porter *et al.* [266] than to the 14th century CE dating obtained by Pajer *et al.* [33] when assuming an older date for strain V1588. Excluding V563 and V1588 from the analysis also demonstrated the stability of the results.

Another study by Smithson *et al.* [131] proposed an improved genome assembly for VD21 that, when used in phylogenetic analysis, dates the common ancestor between VD21 and modern VARV strains to the late 15th or early 16th century and the divergence between P-I and P-II to the late 16th or early 17th century. In this study, the newly improved genome reconstruction for VD21 was used, and dates closer to those of Duggan *et al.* [32], and older than those of Smithson *et al.* [131] were obtained.

The different and at times controversial dating of the VARV phylogeny presented here and in previous studies (e.g., [32, 33, 131]) demonstrates the importance of improving existing genome assemblies and additionally increasing the resolving power of such analyses. This can be achieved by adding additional historical genomes from an expanded spatial and temporal range. This is crucial to understand better the evolution of pathogens over a more extensive temporal transect and geographical distribution and provide additional calibration points for phylogenetic analyses. To this end, approaches that do not use hybridization, such as metagenomic shotgun sequencing, can make a valuable contribution to studying a wide range of hosts and their associated viruses.

CHAPTER 6

Conclusion and outlook

Due to NGS technologies and the constantly growing amount of data, bioinformatics is facing new challenges in various research fields. Examples are the identification and evolution of pathogens [27, 29, 101, 127] and humans [17, 44, 98, 285], clinical analyses to detect genetic defects [286], or the reconstruction of genomes from previously extinct species [98, 287]. In addition, aDNA and evolution studies, for example, are increasingly incorporating samples into individual projects. Also, NGS technology has led to the ability to sequence the entire genomic content rather than targeting only one species. The inherent challenges of bioinformatics are the efficient management and quick processing of large amounts of data and the development of new software to help biologists or archaeologists derive knowledge from the acquired data. In addition, researchers need reproducible methods with which projects can be re-evaluated at any time. Moreover, bioinformatics is demanded to analyze metagenomic samples, especially the differentiation of truly ancient DNA and modern contamination.

The first part of this thesis introduces two user-friendly tools for the efficient processing of NGS data. The challenges for bioinformatics are developing and implementing user-friendly software that supports users, regardless of knowledge level, to analyze their data. They must be able to process NGS data in a standardized and sound manner. In addition, the complete analysis must be reproducible so that the data can be re-analyzed at any time, providing the same results. To be as flexible as possible for new application areas or data types, it is advantageous if the software has a modular structure. This way, the existing software can easily be extended with new functions, such as new analysis methods or visualizations. In this work, the software DamageProfiler (Chapter 3) and mitoBench (Chapter 4) are presented.

DamageProfiler can be used to authenticate the ancient origin of DNA reads based on damage patterns characteristic of ancient DNA (Chapter 2.4). For a given set of mapped DNA fragments, it calculates the frequency of base mismatches, the read length distribution, and the difference between the DNA fragment and the reference genome. The runtime has been significantly improved compared to existing software [42, 43]. In addition, a graphical user interface and comprehensive documentation increase the usability of this software. A possible extension of DamageProfiler could be the implementation of a scaling option similar to mapDamage2 [42], which has already been requested by users. Furthermore, a web-based version of DamageProfiler also could improve its accessibility and usability.

While DamageProfiler can be applied to data from all kinds of species, the software mitoBench is specialized in human data. MitoBench consists of a workbench and a database. The workbench provides a range of specific analysis methods and statistics to investigate human mtDNA and is directly connected with the database, which provides a well-curated, extensive collection of published, complete human mitochondrial genomes, including analysis-relevant meta information. Possible extensions and improvements to mitoBench include a wide variety of approaches. While the current database layout is sufficient and efficient for the actual application, extending and further breaking down this layout into individual tables could allow private user accounts to upload personal records. This data would be accessible only to the owner and anyone with whom it is shared. In addition, the workbench currently only provides a database search based on a few selected attributes, which also results in data being downloaded that must later be sorted out again. An improvement will increase the accessibility of the database contents and allow for more detailed searches, which would reduce the amount of data that needs to be downloaded from the database. This would improve the efficiency further. The visualizations in mitoBench are implemented with JavaFX, which provides many additional formatting and layout options. Unfortunately, the image export in publication-suitable resolution and file format (e.g., SVG) is not supported. Other libraries, such as JFreeChart [288], or Orson Charts [289], can be used. However, this also changes the style of the visualization. Alternatively, mitoBench could be extended to include the functionality of exporting R scripts, which can then be executed and customized in R. This is not only conceivable for visualizations but also to perform further customized analyses. Since in many population genetic analyses only the hypervariable regions of mitochondrial genomes are used, mitoBench could be extended to extract this region as an additional output. In order to make this feasible, all genomes must be aligned against ,i.e., the rCRS or RSRS reference genome to ensure extracting the correct region. For this to be possible, all genomes would need to be aligned to the rCRS or RSRS reference genome to ensure that the correct region is extracted. For example, this could be built into the pipeline of aDNA processing (Section 4.2.4) or directly included in the mitoBench FASTA import functionality.

The second part (Chapter 5) presents two studies in which aDNA was analyzed from human remains. The successful identification of several pathogens by screening the metagenomics data set was shown. In addition, the genome reconstruction of three pathogens - *M. leprae*, HBV, and smallpox - without hybridization capture was demonstrated. This is noteworthy, as pathogen DNA typically represents only a minuscule portion of the metagenomic profile of an ancient or historical sample, so in most cases, hybridization capture is required to reconstruct genomes (e.g., [32, 127, 193]). While this technique can reconstruct genomes more efficiently with lower costs and higher coverage, the drawback is that only existing knowledge can be recovered. Evolutionary events such as genomic rearrangements will potentially be missed because the genome used as a reference for the design of the hybridization probes does not have them. While instances of genome reconstructions without enrichment for pathogen DNA do exist [27, 290–292], they only represent a minority of studies using ancient DNA. The reconstruction of the ancient genomes has been performed using the well-established pipeline EAGER [23]. However, further processing of the consensus sequence is not provided. A workbench similar to mitoBench, but focusing on pathogen genomes, would be helpful to improve the field of ancient pathogen genomics. First attempts are made with nf-core/eager [148], which already integrates the metagenomic screening of the sample. However, subsequent analyses such as phylogenetic reconstructions or recombination analysis are not included. Genetic recombination is still a major issue with many pathogen genomes. While it can be well managed in some species, e.g., *Treponema pallidum* [38, 132], it is still unresolved with HBV, for example. This work provided a new HBV genome showing signs of recombination to be included in further research in this field.

To summarize, this thesis describes two novel and improved methods and programs for the analysis of NGS data (Chapter 3 and 4). In addition, two aDNA studies are described in Chapter 5, also demonstrating how the developed tools can be applied. While the focus of Chapter 3 and 5 is mainly on ancient DNA, Chapter 4 combines the analysis of ancient and modern DNA. Although several investigations have been undertaken, the fields of software development and ancient DNA analysis, this is still potential for improvements in efficiency and user-friendliness of the software, but also the need for additional ancient genomes from several species - especially pathogens - to create a more complete picture of their evolution and origins. Finally, this work extends the applicability of metagenomic analysis to previously unused material, such as tissues from mummified Egyptian individuals and material preserved in ethanol.

Bibliography

[1] R. Higuchi, B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson. "DNA sequences from the quagga, an extinct member of the horse family". In: *Nature* 312.5991 (1984), pages 282–284.

[2] S. Pääbo. "Preservation of DNA in ancient Egyptian mummies". In: *Journal of Archaeological Science* 12.6 (1985), pages 411–417.

[3] N. Weyand, M. Bunnell, et al. "DNA sequence from Cretaceous period bone fragments". In: *Science* 266.5188 (1994), pages 1229–1232.

[4] R. Cano, H Poinar, and G. Poinar Jr. "Isolation and partial characterisation of DNA from the bee Proplebeia dominicana (Apidae: Hymenoptera) in 25–40 million year old amber". In: *Med. Sci. Res* 20 (1992), pages 249–251.

[5] R. Cano and H. Poinar. "Rapid isolation of DNA from fossil and museum specimens suitable for PCR." In: *Biotechniques* 15.3 (1993), pages 432–4.

[6] R. J. Cano, H. N. Poinar, N. J. Pieniazek, A. Acra, and G. O. Poinar. "Amplification and sequencing of DNA from a 120–135-million-year-old weevil". In: *Nature* 363.6429 (1993), pages 536–538.

[7] G. O. Poinar. *Life in amber*. Stanford University Press, 1992.

[8] Z. Hawass, Y. Z. Gad, S. Ismail, R. Khairat, et al. "Ancestry and pathology in King Tutankhamun's family". In: *Jama* 303.7 (2010), pages 638–647.

[9] H. Zischler, M. Hoss, O. Handt, A Von Haeseler, et al. "Detecting dinosaur DNA". In: *Science* 268.5214 (1995), pages 1192–1193.

[10] G Gutiérrez and A Marín. "The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems". en. In: *Mol. Biol. Evol.* 15.7 (1998), pages 926–929.

[11] F. Broushaki, M. G. Thomas, V. Link, S. López, et al. "Early Neolithic genomes from the eastern Fertile Crescent". In: *Science* 353.6298 (2016), pages 499–503.

[12] Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, et al. "Early farmers from across Europe directly descended from Neolithic Aegeans". In: *Proceedings of the National Academy of Sciences* 113.25 (2016), pages 6886–6891.

[13] A. Keller, A. Graefen, M. Ball, M. Matzas, et al. "New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing". In: *Nature communications* 3.1 (2012), pages 1–9.

[14] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518 (2014), pages 409–413.

[15] J. Krause, A. W. Briggs, M. Kircher, T. Maricic, et al. "A complete mtDNA genome of an early modern human from Kostenki, Russia". In: *Current Biology* 20.3 (2010), pages 231–236.

[16] M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, et al. "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104 (2012), pages 222–226.

[17] K. Prüfer, F. Racimo, N. Patterson, F. Jay, et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pages 43–49.

[18] M. Meyer, Q. Fu, A. Aximu-Petri, I. Glocke, et al. "A mitochondrial genome sequence of a hominin from Sima de los Huesos". In: *Nature* 505.7483 (2014), pages 403–406.

[19] S. Sawyer, G. Renaud, B. Viola, J.-J. Hublin, et al. "Nuclear and mitochondrial DNA sequences from two Denisovan individuals". In: *Proceedings of the National Academy of Sciences* 112.51 (2015), pages 15696–15700.

[20] H. Rougier, I. Crevecoeur, C. Beauval, C. Posth, et al. "Neandertal cannibalism and Neandertal bones used as tools in Northern Europe". In: *Scientific reports* 6.1 (2016), pages 1–11.

[21] C. Posth, C. Wißing, K. Kitagawa, L. Pagani, et al. "Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals". In: *Nature communications* 8.1 (2017), pages 1–9.

[22] M. Kircher. "Analysis of high-throughput ancient DNA sequencing data". In: *Ancient DNA*. Springer, 2012, pages 197–228.

[23] A. Peltzer, G. Jäger, A. Herbig, A. Seitz, et al. "EAGER: efficient ancient genome reconstruction". In: *Genome biology* 17.1 (2016), page 60.

[24] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, et al. "The human microbiome project". In: *Nature* 449.7164 (2007), pages 804–810.

[25] C. Warinner, J. F. M. Rodrigues, R. Vyas, C. Trachsel, et al. "Pathogens and host immunity in the ancient human oral cavity". In: *Nature genetics* 46.4 (2014), pages 336–344.

[26] R. Y. Tito, D. Knights, J. Metcalf, A. J. Obregon-Tito, et al. "Insights from characterizing extinct human gut microbiomes". In: *PloS one* 7.12 (2012), e51146.

[27] V. J. Schuenemann, P. Singh, T. A. Mendum, B. Krause-Kyora, et al. "Genome-wide comparison of medieval and modern Mycobacterium leprae". In: *Science* 341.6142 (2013), pages 179–183.

[28] V. J. Schuenemann, C. Avanzi, B. Krause-Kyora, A. Seitz, et al. "Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe". In: *PLoS pathogens* 14.5 (2018), e1006997.

[29] B. Krause-Kyora, J. Susat, F. M. Key, D. Kühnert, et al. "Neolithic and medieval virus genomes reveal complex evolution of hepatitis B". In: *Elife* 7 (2018), e36666.

[30] B. Mühlemann, T. C. Jones, P. de Barros Damgaard, M. E. Allentoft, et al. "Ancient hepatitis B viruses from the Bronze Age to the Medieval period". In: *Nature* 557.7705 (2018), pages 418–423.

[31] Z. P. Ross, J. Klunk, G. Fornaciari, V. Giuffra, et al. "The paradox of HBV evolution as revealed from a 16th century mummy". In: *PLoS pathogens* 14.1 (2018), e1006750.

[32] A. T. Duggan, M. F. Perdomo, D. Piombino-Mascali, S. Marciniak, et al. "17th century variola virus reveals the recent history of smallpox". In: *Current Biology* 26.24 (2016), pages 3407–3412.

[33] P. Pajer, J. Dresler, H. Kabíckova, L. Písa, et al. "Characterization of two historic smallpox specimens from a Czech museum". In: *Viruses* 9.8 (2017), page 200.

[34] P. Simmonds and S. Midgley. "Recombination in the genesis and evolution of hepatitis B virus genotypes". In: *Journal of virology* 79.24 (2005), pages 15467–15476.

[35] B. F. d. C. D. Souza, J. F. Drexler, R. S. d. Lima, M. d. O. H. V. d. Rosário, and E. M. Netto. "Theories about evolutionary origins of human hepatitis B virus in primates and humans". In: *Brazilian Journal of Infectious Diseases* 18.5 (2014), pages 535–543.

[36] A. Rasche, B. F. d. C. D. Souza, and J. F. Drexler. "Bat hepadnaviruses and the origins of primate hepatitis B viruses". In: *Current opinion in virology* 16 (2016), pages 86–94.

[37] J. J. Esposito, S. A. Sammons, A. M. Frace, J. D. Osborne, et al. "Genome sequence diversity and clues to the evolution of variola (smallpox) virus". In: *Science* 313.5788 (2006), pages 807–812.

[38] K. Majander, S. Pfrengle, A. Kocher, J. Neukamm, et al. "Ancient bacterial genomes reveal a high diversity of Treponema pallidum Strains in early Modern Europe". In: *Current Biology* 30.19 (2020), pages 3788–3803.

[39] T. Lindahl. "Instability and decay of the primary structure of DNA". In: *nature* 362.6422 (1993), pages 709–715.

[40] M. Hofreiter, D. Serre, H. N. Poinar, M. Kuch, and S. Pääbo. "ancient DNA". In: *Nature Reviews Genetics* 2.5 (2001), pages 353–359.

[41] A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, et al. "Patterns of damage in genomic DNA sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences* 104.37 (2007), pages 14616–14621.

[42] H. Jónsson, A. Ginolhac, M. Schubert, P. L. Johnson, and L. Orlando. "mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters". In: *Bioinformatics* 29.13 (2013), pages 1682–1684.

[43] P. Skoglund, B. H. Northoff, M. V. Shunkov, A. P. Derevianko, et al. "Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal". In: *Proceedings of the National Academy of Sciences* 111.6 (2014), pages 2229–2234.

[44] C. Posth, G. Renaud, A. Mittnik, D. G. Drucker, et al. "Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe". In: *Current Biology* 26.6 (2016), pages 827–833.

[45] Q. Fu, C. Posth, M. Hajdinjak, M. Petr, et al. "The genetic history of ice age Europe". In: *Nature* 534.7606 (2016), pages 200–205.

[46] G. Brandt, W. Haak, C. J. Adler, C. Roth, et al. "Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity". In: *Science* 342.6155 (2013), pages 257–261.

[47] P. Brotherton, W. Haak, J. Templeton, G. Brandt, et al. "Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans". In: *Nature communications* 4.1 (2013), pages 1–11.

[48] I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, et al. "Genomic insights into the origin of farming in the ancient Near East". In: *Nature* 536.7617 (2016), pages 419–424.

[49] L. Excoffier and H. E. Lischer. "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows". In: *Molecular ecology resources* 10.3 (2010), pages 564–567.

[50] K. Ishiya and S. Ueda. "MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing". In: *PeerJ* 5 (2017), e3406.

[51] W. Parson and A. Dür. "EMPOP–a forensic mtDNA database". In: *Forensic science international. Genetics* 1.2 (2007), 88—92. ISSN: 1872-4973. DOI: 10.1016/j.fsigen.2007.01.018. URL: https://doi.org/10.1016/j.fsigen.2007.01.018.

[52] M. T. Lott, J. N. Leipzig, O. Derbeneva, H. M. Xie, et al. "mtDNA variation and analysis using mitomap and mitomaster". In: *Current protocols in bioinformatics* 44.1 (2013), pages 1–23.

[53] R. Clima, R. Preste, C. Calabrese, M. A. Diroma, et al. "HmtDB 2016: data update, a better performing query system and human mitochondrial DNA haplogroup predictor". In: *Nucleic acids research* 45.D1 (2017), pages D698–D706.

[54] M. Ingman and U. Gyllensten. "mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences". In: *Nucleic acids research* 34.suppl_1 (2006), pages D749–D751.

[55] E. Ehler, J. Novotnỳ, A. Juras, M. Chyleński, et al. "AmtDB: a database of ancient human mitochondrial genomes". In: *Nucleic acids research* 47.D1 (2019), pages D29–D32.

[56] M. T. P. Gilbert, I. Barnes, M. J. Collins, C. Smith, et al. "Long-term survival of ancient DNA in Egypt: Response to Zink and Nerlich (2003)". In: *American journal of physical anthropology* 128.1 (2005), pages 110–114.

[57] S. Pääbo. "Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification". In: *Proceedings of the National Academy of Sciences* 86.6 (1989), pages 1939–1943.

[58] J. D. Watson and F. H. Crick. "Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid". In: *Nature* 171.4356 (1953), pages 737–738.

[59] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, et al. "Structure of a ribonucleic acid". In: *Science* (1965), pages 1462–1465.

[60] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12 (1977), pages 5463–5467.

[61] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, et al. "The sequence of the human genome". In: *science* 291.5507 (2001), pages 1304–1351.

[62] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, et al. "Initial sequencing and analysis of the human genome". In: (2001).

[63] J. A. Schloss. "How to get genomes at one ten-thousandth the cost". In: *Nature biotechnology* 26.10 (2008), pages 1113–1115.

[64] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén. "Real-time DNA sequencing using detection of pyrophosphate release". In: *Analytical biochemistry* 242.1 (1996), pages 84–89.

[65] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, et al. "Accurate multiplex polony sequencing of an evolved bacterial genome". In: *Science* 309.5741 (2005), pages 1728–1732.

[66] R. Drmanac, A. B. Sparks, M. J. Callow, A. L. Halpern, et al. "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays". In: *Science* 327.5961 (2010), pages 78–81.

[67] S. Balasubramanian, D. Klenerman, and C. Barnes. *Arrayed polynucleotides and their use in genome analysis*. US Patent App. 10/153,240. 2003.

[68] I. Braslavsky, B. Hebert, E. Kartalov, and S. R. Quake. "Sequence information can be obtained from single DNA molecules". In: *Proceedings of the National Academy of Sciences* 100.7 (2003), pages 3960–3964.

[69] J. Eid, A. Fehr, J. Gray, K. Luong, et al. "Real-time DNA sequencing from single polymerase molecules". In: *Science* 323.5910 (2009), pages 133–138.

[70] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, et al. "The potential and challenges of nanopore sequencing". In: *Nanoscience and technology: A collection of reviews from Nature Journals*. World Scientific, 2010, pages 261–268.

[71] C. Bleidorn. "Third generation sequencing: technology and its potential impact on evolutionary biodiversity research". In: *Systematics and biodiversity* 14.1 (2016), pages 1–8.

[72] P. K. Gupta. "Single-molecule DNA sequencing technologies for future genomics research". In: *Trends in biotechnology* 26.11 (2008), pages 602–611.

[73] D. Laehnemann, A. Borkhardt, and A. C. McHardy. "Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction". In: *Briefings in Bioinformatics* 17.1 (2015), pages 154–179. ISSN: 1467-5463. DOI: 10.1093/bib/bbv029. URL: https://doi.org/10.1093/bib/bbv029.

[74] Illumina. "An introduction to Next-Generation Sequencing Technology". In: (2017).

[75] A. W. Briggs and P. Heyn. "Preparation of next-generation sequencing libraries from damaged DNA". In: *Ancient DNA*. Springer, 2012, pages 143–154.

[76] M.-T. Gansauge and M. Meyer. "Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA". In: *Nature protocols* 8.4 (2013), pages 737–748.

[77] M.-T. Gansauge, T. Gerber, I. Glocke, P. Korlević, et al. "Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase". In: *Nucleic acids research* 45.10 (2017), e79–e79.

[78] M.-T. Gansauge, A. Aximu-Petri, S. Nagel, and M. Meyer. "Manual and automated preparation of single-stranded DNA libraries for the sequencing of DNA from ancient biological remains and other sources of highly degraded DNA". In: *Nature Protocols* 15.8 (2020), pages 2279–2300.

[79] *NovaSeq6000 Sequencing System*. https://www.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-6000-spec-sheet-770-2016-025/novaseq-6000-spec-sheet-770-2016-025.pdf. 2020.

[80] T. Nakazato, T. Ohta, and H. Bono. "Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive". In: *PLoS One* 8.10 (2013), e77910.

[81] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". In: *Briefings in bioinformatics* 14.2 (2013), pages 178–192.

[82] V. M. Ingram. "A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin". In: *Nature* 178.4537 (1956), pages 792–794.

[83] J. C. Chang and Y. W. Kan. "beta 0 thalassemia, a nonsense mutation in man". In: *Proceedings of the National Academy of Sciences* 76.6 (1979), pages 2886–2889.

[84] A. Hamosh, T. M. King, B. J. Rosenstein, M. Corey, et al. "Cystic fibrosis patients bearing both the common missense mutation, Gly$\rightarrow$ Asp at codon 551 and the ΔF508 mutation are clinically indistinguishable from ΔF508 homozygotes, except for decreased risk of meconium ileus". In: *American journal of human genetics* 51.2 (1992), page 245.

[85] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, et al. "The international HapMap project". In: (2003).

[86] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, et al. "An initial map of insertion and deletion (INDEL) variation in the human genome". In: *Genome research* 16.9 (2006), pages 1182–1190.

[87] J. M. Mullaney, R. E. Mills, W. S. Pittard, and S. E. Devine. "Small insertions and deletions (INDELs) in human genomes". In: *Human molecular genetics* 19.R2 (2010), R131–R136.

[88] E. Hazkani-Covo, R. M. Zeller, and W. Martin. "Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes". In: *PLoS Genet* 6.2 (2010), e1000834.

[89] J. Dabney, M. Meyer, and S. Pääbo. "Ancient DNA damage". In: *Cold Spring Harbor perspectives in biology* 5.7 (2013), a012567.

[90] D. Carroll. "Genetic Recombination". In: *Encyclopedia of Genetics*. Edited by S. Brenner and J. H. Miller. New York: Academic Press, 2001, pages 841 –845. ISBN: 978-0-12-227080-2. DOI: https://doi.org/10.1006/rwgn.2001.0543. URL: http://www.sciencedirect.com/science/article/pii/B0122270800005437.

[91] M. J. McDonald, D. P. Rice, and M. M. Desai. "Sex speeds adaptation by altering the dynamics of molecular evolution". In: *Nature* 531.7593 (2016), pages 233–236.

[92] C. M. Marra, S. K. Sahi, L. C. Tantalo, C. Godornes, et al. "Enhanced molecular typing of Treponema pallidum: geographical distribution of strain types and association with neurosyphilis". In: *The Journal of infectious diseases* 202.9 (2010), pages 1380–1388.

[93] H. Pětrošová, M. Zobaníková, D. Čejková, L. Mikalova, et al. "Whole genome sequence of Treponema pallidum ssp. pallidum, strain Mexico A, suggests recombination between yaws and syphilis strains". In: *PLoS Negl Trop Dis* 6.9 (2012), e1832.

[94] M. Stoneking. *An introduction to molecular anthropology*. John Wiley & Sons, 2016.

[95] E. Willerslev, E. Cappellini, W. Boomsma, R. Nielsen, et al. "Ancient biomolecules from deep ice cores reveal a forested southern Greenland". In: *Science* 317.5834 (2007), pages 111–114.

[96] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, and S. Pääbo. "Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA". In: *PloS one* 7.3 (2012), e34131.

[97] S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Després, et al. "Genetic analyses from ancient DNA". In: *Annu Rev Genet* 38.1 (2004), pages 645–679.

[98] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, et al. "A draft sequence of the Neandertal genome". In: *science* 328.5979 (2010), pages 710–722.

[99] M. L. Carpenter, J. D. Buenrostro, C. Valdiosera, H. Schroeder, et al. "Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries". In: *The American Journal of Human Genetics* 93.5 (2013), pages 852–864.

[100] J. B. Hays and B. H. Zimm. "Flexibility and stiffness in nicked DNA". In: *Journal of molecular biology* 48.2 (1970), pages 297–317.

[101] Å. J. Vågene, A. Herbig, M. G. Campana, N. M. R. García, et al. "Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico". In: *Nature ecology & evolution* 2.3 (2018), pages 520–528.

[102] D. Knights, J. Kuczynski, E. S. Charlson, J. Zaneveld, et al. "Bayesian community-wide culture-independent microbial source tracking". In: *Nature methods* 8.9 (2011), pages 761–763.

[103] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, et al. "From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline". In: *Current protocols in bioinformatics* 43.1 (2013), pages 11–10.

[104] M. Schubert, A. Ginolhac, S. Lindgreen, J. F. Thompson, et al. "Improving ancient DNA read mapping against modern reference genomes". In: *BMC genomics* 13.1 (2012), page 178.

[105] D. R. Wolstenholme. "Animal Mitochondrial DNA: Structure and Evolution". In: edited by D. R. Wolstenholme and K. W. Jeon. Volume 141. International Review of Cytology. Academic Press, 1992, pages 173 –216. DOI: https://doi.org/10.1016/S0074-7696(08)62066-5. URL: http://www.sciencedirect.com/science/article/pii/S0074769608620665.

[106] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, et al. "Sequence and organization of the human mitochondrial genome". In: *Nature* 290.5806 (1981), pages 457–465.

[107] C. A. Hutchison, J. E. Newbold, S. S. Potter, and M. H. Edgell. "Maternal inheritance of mammalian mitochondrial DNA". In: *Nature* 251.5475 (1974), pages 536–538.

[108] G. S. Michaels, W. W. Hauswirth, and P. J. Laipis. "Mitochondrial DNA copy number in bovine oocytes and somatic cells". In: *Developmental biology* 94.1 (1982), pages 246–251.

[109] L. Pikó and L. Matsumoto. "Number of mitochondria and some properties of mitochondrial DNA in the mouse egg". In: *Developmental biology* 49.1 (1976), pages 1–10.

[110] W. M. Brown, M. George, and A. C. Wilson. "Rapid evolution of animal mitochondrial DNA". In: *Proceedings of the National Academy of Sciences* 76.4 (1979), pages 1967–1971.

[111] E. Hagström, C. Freyer, B. J. Battersby, J. B. Stewart, and N.-G. Larsson. "No recombination of mtDNA after heteroplasmy for 50 generations in the mouse maternal germline". In: *Nucleic acids research* 42.2 (2013), pages 1111–1116.

[112] D. A. Merriwether, A. G. Clark, S. W. Ballinger, T. G. Schurr, et al. "The structure of human mitochondrial DNA variation". In: *Journal of molecular evolution* 33.6 (1991), pages 543–555.

[113] M. Cieslak, M. Pruvost, N. Benecke, M. Hofreiter, et al. "Origin and history of mitochondrial DNA lineages in domestic horses". In: *PloS one* 5.12 (2010), e15311.

[114] J. Gretzinger, M. Molak, E. Reiter, S. Pfrengle, et al. "Large-scale mitogenomic analysis of the phylogeography of the Late Pleistocene cave bear". In: *Scientific reports* 9.1 (2019), pages 1–11.

[115] O. Thalmann, B. Shapiro, P. Cui, V. J. Schuenemann, et al. "Complete mitochondrial genomes of ancient canids suggest a European origin of domestic dogs". In: *Science* 342.6160 (2013), pages 871–874.

[116] C. F. Aquadro and B. D. Greenberg. "Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals". In: *Genetics* 103.2 (1983), pages 287–312.

[117] M. J. Johnson, D. C. Wallace, S. D. Ferris, M. C. Rattazzi, and L. L. Cavalli-Sforza. "Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns". In: *Journal of Molecular Evolution* 19.3-4 (1983), pages 255–271.

[118] Q. Fu, A. Mittnik, P. L. Johnson, K. Bos, et al. "A revised timescale for human evolution based on ancient mitochondrial genomes". In: *Current biology* 23.7 (2013), pages 553–559.

[119] V. J. Schuenemann, A. Peltzer, B. Welte, W. P. Van Pelt, et al. "Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods". In: *Nature communications* 8.1 (2017), pages 1–11.

[120] J. Jarczak, Ł. Grochowalski, B. Marciniak, J. Lach, et al. "Mitochondrial DNA variability of the Polish population". In: *European Journal of Human Genetics* 27.8 (2019), pages 1304–1314.

[121] G. U. Rehman. "Mitochondrial DNA analysis of Chitrali population of Pakistan from ancient human bones". In: *Meta Gene* (2020), page 100821.

[122] M. Van Oven. "PhyloTree Build 17: Growing the human mitochondrial DNA tree". In: *Forensic Science International: Genetics Supplement Series* 5 (2015), e392–e394.

[123] R. L. Cann, M. Stoneking, and A. C. Wilson. "Mitochondrial DNA and human evolution". In: *Nature* 325.6099 (1987), pages 31–36.

[124] M. Stoneking, L. B. Jorde, K. Bhatia, and A. C. Wilson. "Geographic variation in human mitochondrial DNA from Papua New Guinea." In: *Genetics* 124.3 (1990), pages 717–733.

[125] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. "African populations and the evolution of human mitochondrial DNA". In: *Science* 253.5027 (1991), pages 1503–1507.

[126] C. P. Andam, C. J. Worby, Q. Chang, and M. G. Campana. "Microbial genomics of ancient plagues and outbreaks". In: *Trends in microbiology* 24.12 (2016), pages 978–990.

[127] K. I. Bos, V. J. Schuenemann, G. B. Golding, H. A. Burbano, et al. "A draft genome of Yersinia pestis from victims of the Black Death". In: *Nature* 478.7370 (2011), pages 506–510.

[128] M. A. Spyrou, R. I. Tukhbatova, C.-C. Wang, A. A. Valtueña, et al. "Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague". In: *Nature communications* 9.1 (2018), pages 1–10.

[129] N. Rascovan, K.-G. Sjögren, K. Kristiansen, R. Nielsen, et al. "Emergence and spread of basal lineages of Yersinia pestis during the Neolithic decline". In: *Cell* 176.1-2 (2019), pages 295–305.

[130] P. Biagini, C. Thèves, P. Balaresque, A. Geraut, et al. "Variola virus in a 300-year-old Siberian mummy." In: (2012).

[131] C. Smithson, J. Imbery, and C. Upton. "Re-assembly and analysis of an ancient variola virus genome". In: *Viruses* 9.9 (2017), page 253.

[132] N. Arora, V. J. Schuenemann, G. Jäger, A. Peltzer, et al. "Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster". In: *Nature microbiology* 2.1 (2016), pages 1–6.

[133] J. Neukamm, A. Peltzer, and K. Nieselt. "DamageProfiler: Fast damage pattern calculation for ancient DNA". In: *Bioinformatics* (2021). btab190. ISSN: 1367-4803.

[134] G. Renaud, V. Slon, A. T. Duggan, and J. Kelso. "Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA". In: *Genome biology* 16.1 (2015), page 224.

[135] F. Racimo, G. Renaud, and M. Slatkin. "Joint estimation of contamination, error and demography for nuclear DNA from ancient humans". In: *PLoS genetics* 12.4 (2016), e1005972.

[136] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. "ANGSD: analysis of next generation sequencing data". In: *BMC bioinformatics* 15.1 (2014), page 356.

[137] R. J. Cano, F. Tiefenbrunner, M. Ubaldi, C. Del Cueto, et al. "Sequence analysis of bacterial DNA in the colon and stomach of the Tyrolean Iceman". In: *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 112.3 (2000), pages 297–309.

[138] R. Y. Tito, S. Macmil, G. Wiley, F. Najar, et al. "Phylotyping and functional analysis of two ancient human microbiomes". In: *PLoS One* 3.11 (2008), e3703.

[139] R. Hübler, F. M. Key, C. Warinner, K. I. Bos, et al. "HOPS: Automated detection and authentication of pathogen DNA in archaeological remains". In: *Genome biology* 20.1 (2019), pages 1–13.

[140] B. Grüning, R. Dale, A. Sjödin, B. A. Chapman, et al. "Bioconda: sustainable and comprehensive software distribution for the life sciences". In: *Nature methods* 15.7 (2018), pages 475–476.

[141] H. Li, B. Handsaker, A. Wysoker, T. Fennell, et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pages 2078–2079.

[142] P. Ewels, M. Magnusson, S. Lundin, and M. Käller. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (2016), pages 3047–3048.

[143] G. Renaud, K. Hanghøj, E. Willerslev, and L. Orlando. "gargammel: a sequence simulator for ancient DNA". In: *Bioinformatics* 33.4 (2017), pages 577–579.

[144] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". In: *Nature methods* 9.4 (2012), page 357.

[145] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *bioinformatics* 25.14 (2009), pages 1754–1760.

[146] H. Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: *arXiv preprint arXiv:1303.3997* (2013).

[147] H. Li. "Minimap2: pairwise alignment for nucleotide sequences". In: *Bioinformatics* 34.18 (2018), pages 3094–3100.

[148] J. A. F. Yates, T. C. Lamnidis, M. Borry, A. A. Valtueña, et al. "Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager". In: *PeerJ* 9 (2021), e10947.

[149] M. M. Nass and S. Nass. "Intramitochondrial fibers with DNA characteristics I. Fixation and electron staining reactions". In: *Journal of Cell Biology* 19.3 (1963), pages 593–611.

[150] P. Pamilo and M. Nei. "Relationships between gene trees and species trees." In: *Molecular biology and evolution* 5.5 (1988), pages 568–583.

[151] C. W. Birky Jr. "Evolution and population genetics of organelle genes: mechanisms and models". In: *Evolution at the molecular level/edited by Robert K. Selander, Andrew G. Clark, and Thomas S. Whittman* (1991).

[152] J. W. O. Ballard and M. C. Whitlock. "The incomplete natural history of mitochondria". In: *Molecular ecology* 13.4 (2004), pages 729–744.

[153] A. Gaziev and G. Shaikhaev. "Nuclear mitochondrial pseudogenes". In: *Molecular Biology* 44.3 (2010), pages 358–368.

[154] T. Hlaing, W. Tun-Lin, P. Somboon, D. Socheat, et al. "Mitochondrial pseudogenes in the nuclear genome of Aedes aegypti mosquitoes: implications for past and future population genetic studies". In: *Bmc Genetics* 10.1 (2009), pages 1–12.

[155] J. V. Lopez, N. Yuhki, R. Masuda, W. Modi, and S. J. O'Brien. "Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat". In: *Journal of molecular evolution* 39.2 (1994), pages 174–190.

[156] O. Thalmann, D. Serre, M. Hofreiter, D. Lukas, et al. "Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA". In: *Molecular Ecology* 14.1 (2005), pages 179–188.

[157] K. Thangaraj, M. B. Joshi, A. G. Reddy, A. A. Rasalkar, and L. Singh. "Sperm mitochondrial mutations as a cause of low sperm motility". In: *Journal of andrology* 24.3 (2003), pages 388–392.

[158] Y.-G. Yao, Q.-P. Kong, A. Salas, and H.-J. Bandelt. "Pseudomitochondrial genome haunts disease studies". In: *Journal of medical genetics* 45.12 (2008), pages 769–772.

[159] C. Der Sarkissian, O. Balanovsky, G. Brandt, V. Khartanovich, et al. "Ancient DNA reveals prehistoric gene-flow from Siberia in the complex human population history of North East Europe". In: *PLoS Genet* 9.2 (2013), e1003296.

[160] I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, et al. "Genetic origins of the Minoans and Mycenaeans". In: *Nature* 548.7666 (2017), pages 214–218.

[161] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, et al. "Genome-wide patterns of selection in 230 ancient Eurasians". In: *Nature* 528.7583 (2015), pages 499–503.

[162] A. Furtwängler, A. B. Rohrlach, T. C. Lamnidis, L. Papac, et al. "Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland". In: *Nature communications* 11.1 (2020), pages 1–11.

[163] E. Skourtanioti, Y. S. Erdal, M. Frangipane, F. B. Restelli, et al. "Genomic History of Neolithic to Bronze Age Anatolia, Northern Levant, and Southern Caucasus". In: *Cell* 181.5 (2020), pages 1158–1175.

[164] H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, et al. "HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing". In: *Nucleic acids research* 44.W1 (2016), W58–W63.

[165] B. A. Mahesh, E Kannan, G. D. J. Davis, P Venkatesan, and P. Ragunath. "GenPop-An Online Tool to Analyze Human Population Genetic Data". In: *Bioinformation* 16.2 (2020), page 149.

[166] M. Scheibye-Knudsen, K. Scheibye-Alsing, C. Canugovi, D. L. Croteau, and V. A. Bohr. "A novel diagnostic tool reveals mitochondrial pathology in human diseases and aging". In: *Aging (Albany NY)* 5.3 (2013), page 192.

[167] M. Ziemann, Y. Eren, and A. El-Osta. "Gene name errors are widespread in the scientific literature". In: *Genome biology* 17.1 (2016), pages 1–3.

[168] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, et al. "BEAST 2: a software platform for Bayesian evolutionary analysis". In: *PLoS Comput Biol* 10.4 (2014), e1003537.

[169] A. J. Drummond and A. Rambaut. "BEAST: Bayesian evolutionary analysis by sampling trees". In: *BMC evolutionary biology* 7.1 (2007), pages 1–8.

[170] A. Peltzer. "Computational methods for ancient genome reconstruction". PhD thesis. Eberhard Karls Universität Tübingen, 2019.

[171] M. G. Kendall et al. "course in multivariate analysis". In: (1965).

[172] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. "Bayesian phylogenetics with BEAUti and the BEAST 1.7". In: *Molecular biology and evolution* 29.8 (2012), pages 1969–1973.

[173] K. Katoh and D. M. Standley. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". In: *Molecular biology and evolution* 30.4 (2013), pages 772–780.

[174] A. Stamatakis. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies". In: *Bioinformatics* 30.9 (2014), pages 1312–1313.

[175] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, et al. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0". In: *Systematic biology* 59.3 (2010), pages 307–321.

[176] T. T. Community. *TestFX: Testing GUI applications in Java.* `https://github.com/TestFX/TestFX`. 2020.

[177] *UNSD — Methodology.* `https://unstats.un.org/unsd/methodology/m49/`. Accessed on 01.01.2021.

[178] *Geonames.* `http://download.geonames.org/export/dump`. Accessed on 01.01.2021.

[179] *OpenStreetMap.* `https://www.openstreetmap.org`. Accessed on 01.01.2021.

[180] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. *Glottolog 4.3*. Jena, 2020. DOI: `10.5281/zenodo.4061162`. URL: `https://glottolog.org/accessed2021-01-01`.

[181] J. Köster and S. Rahmann. "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19 (2012), pages 2520–2522.

[182] A. Modi, H. Lancioni, I. Cardinali, M. R. Capodiferro, et al. "The mitogenome portrait of Umbria in Central Italy as depicted by contemporary inhabitants and pre-Roman remains". In: *Scientific reports* 10.1 (2020), pages 1–12.

[183] M. Cerezo, A. Achilli, A. Olivieri, U. A. Perego, et al. "Reconstructing ancient mitochondrial DNA links between Africa and Europe". In: *Genome research* 22.5 (2012), pages 821–826.

[184] A. Olivieri, M. Pala, F. Gandini, B. H. Kashani, et al. "Mitogenomes from two uncommon haplogroups mark late glacial/postglacial expansions from the near east and neolithic dispersals within Europe". In: *PloS one* 8.7 (2013), e70492.

[185] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, et al. "Metagenomic analysis of the human distal gut microbiome". In: *science* 312.5778 (2006), pages 1355–1359.

[186] E. M. Bik, C. D. Long, G. C. Armitage, P. Loomer, et al. "Bacterial diversity in the oral cavity of 10 healthy individuals". In: *The ISME journal* 4.8 (2010), pages 962–974.

[187] E. A. Grice and J. A. Segre. "The skin microbiome". In: *Nature reviews microbiology* 9.4 (2011), pages 244–253.

[188] I. Nasidze, J. Li, D. Quinque, K. Tang, and M. Stoneking. "Global diversity in the human salivary microbiome". In: *Genome research* 19.4 (2009), pages 636–643.

[189] H. R. Preus, O. J. Marvik, K. A. Selvig, and P. Bennike. "Ancient bacterial DNA (aDNA) in dental calculus from archaeological human remains". In: *Journal of Archaeological Science* 38.8 (2011), pages 1827–1831.

[190] A. T. Ozga, M. A. Nieves-Colón, T. P. Honap, K. Sankaranarayanan, et al. "Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus". In: *American journal of physical anthropology* 160.2 (2016), pages 220–228.

[191] C. Warinner, A. Herbig, A. Mann, J. A. Fellows Yates, et al. "A robust framework for microbial archaeology". In: *Annual review of genomics and human genetics* 18 (2017), pages 321–356.

[192] J. Neukamm, S. Pfrengle, M. Molak, A. Seitz, et al. "2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals". In: *BMC biology* 18.1 (2020), pages 1–18.

[193] F. Maixner, B. Krause-Kyora, D. Turaev, A. Herbig, et al. "The 5300-year-old Helicobacter pylori genome of the Iceman". In: *Science* 351.6269 (2016), pages 162–165.

[194] L. Pagani, S. Schiffels, D. Gurdasani, P. Danecek, et al. "Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians". In: *The American Journal of Human Genetics* 96.6 (2015), pages 986–991.

[195] O. Loreille, S. Ratnayake, A. L. Bazinet, T. B. Stockwell, et al. "Biological sexing of a 4000-year-old Egyptian mummy head to assess the potential of nuclear DNA recovery from the most damaged and limited forensic specimens". In: *Genes* 9.3 (2018), page 135.

[196] A. Cooper and H. N. Poinar. "Ancient DNA: do it right or not at all". In: *Science* 289.5482 (2000), pages 1139–1139.

[197] J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, et al. "Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments". In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pages 15758–15763.

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

[199] M. Kircher, S. Sawyer, and M. Meyer. "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform". In: *Nucleic acids research* 40.1 (2012), e3–e3.

[200] S. Andrews et al. *FastQC: a quality control tool for high throughput sequence data*. `https://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. 2010.

[201] M. Schubert, S. Lindgreen, and L. Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". In: *BMC research notes* 9.1 (2016), pages 1–7.

[202] S. Beier, R. Tappu, and D. H. Huson. "Functional analysis in metagenomics using MEGAN 6". In: *Functional Metagenomics: Tools and Applications*. Springer, 2017, pages 65–74.

[203] M. Barbara A, N. Karen E, P. Mihai, C. Heather H, et al. "A framework for human microbiome research". In: *Nature* 486.7402 (2012).

[204] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, et al. "Structure, function and diversity of the healthy human microbiome". In: *nature* 486.7402 (2012), page 207.

[205] I. M. Velsko, J. A. F. Yates, F. Aron, R. W. Hagan, et al. "Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage". In: *Microbiome* 7.1 (2019), page 102.

[206] M. Köberl, H. Müller, E. M. Ramadan, and G. Berg. "Desert farming benefits from microbial potential in arid soils and promotes diversity and plant health". In: *PLoS One* 6.9 (2011), e24452.

[207] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data". In: *Nature genetics* 43.5 (2011), page 491.

[208] A. Seitz. *MUSIAL - MUlti Sample varIant AnaLysis*. `https://github.com/Integrative-Transcriptomics/MUSIAL`. 2017.

[209] P. Cingolani, A. Platts, L. L. Wang, M. Coon, et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3". In: *Fly* 6.2 (2012), pages 80–92.

[210] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura. "MEGA X: molecular evolutionary genetics analysis across computing platforms". In: *Molecular biology and evolution* 35.6 (2018), pages 1547–1549.

[211] J. C. Gervin, M. Behrenfeld, C. R. McClain, J. Spinhirne, et al. "PhyLM: A Mission Design Concept for an Optical/Lidar Instrument to Measure Ocean Productivity and Aerosols from Space". In: (2004).

[212] A. Rambaut, T. T. Lam, L. Max Carvalho, and O. G. Pybus. "Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen)". In: *Virus evolution* 2.1 (2016), vew007.

[213] R. Bouckaert, T. G. Vaughan, J. Barido-Sottani, S. Duchêne, et al. "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis". In: *PLoS computational biology* 15.4 (2019), e1006650.

[214] A. Benjak, C. Avanzi, P. Singh, C. Loiseau, et al. "Phylogenomics and antimicrobial resistance of the leprosy bacillus Mycobacterium leprae". In: *Nature communications* 9.1 (2018), pages 1–11.

[215] D. H. Huson. "SplitsTree: analyzing and visualizing evolutionary data." In: *Bioinformatics (Oxford, England)* 14.1 (1998), pages 68–73.

[216] D. Bryant and V. Moulton. "Neighbor-net: an agglomerative method for the construction of phylogenetic networks". In: *Molecular biology and evolution* 21.2 (2004), pages 255–265.

[217] G. Kahila Bar-Gal, M. J. Kim, A. Klein, D. H. Shin, et al. "Tracing hepatitis B virus to the 16th century in a Korean mummy". In: *Hepatology* 56.5 (2012), pages 1671–1680.

[218] D. P. Martin, B. Murrell, M. Golden, A. Khoosal, and B. Muhire. "RDP4: Detection and analysis of recombination patterns in virus genomes". In: *Virus evolution* 1.1 (2015).

[219] D. Martin and E. Rybicki. "RDP: detection of recombination amongst aligned sequences". In: *Bioinformatics* 16.6 (2000), pages 562–563.

[220] M. Padidam, S. Sawyer, and C. M. Fauquet. "Possible emergence of new geminiviruses by frequent recombination". In: *Virology* 265.2 (1999), pages 218–225.

[221] D. Posada and K. A. Crandall. "Evaluation of methods for detecting recombination from DNA sequences: computer simulations". In: *Proceedings of the National Academy of Sciences* 98.24 (2001), pages 13757–13762.

[222] J. M. Smith. "Analyzing the mosaic structure of genes". In: *Journal of molecular evolution* 34.2 (1992), pages 126–129.

[223] M. O. Salminen, J. K. CARR, D. S. BURKE, and F. E. McCUTCHAN. "Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning". In: *AIDS research and human retroviruses* 11.11 (1995), pages 1423–1425.

[224] M. J. Gibbs, J. S. Armstrong, and A. J. Gibbs. "Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences". In: *Bioinformatics* 16.7 (2000), pages 573–582.

[225] M. F. Boni, D. Posada, and M. W. Feldman. "An exact nonparametric method for inferring mosaic structure in sequence triplets". In: *Genetics* 176.2 (2007), pages 1035–1047.

[226] C. L. Wells and T. D. Wilkins. "Clostridia: sporeforming anaerobic bacilli". In: *Medical Microbiology. 4th edition*. University of Texas Medical Branch at Galveston, 1996.

[227] G. T. Javan, S. J. Finley, T. Smith, J. Miller, and J. E. Wilkinson. "Cadaver thanatomicrobiome signatures: the ubiquitous nature of Clostridium species in human decomposition". In: *Frontiers in microbiology* 8 (2017), page 2096.

[228] J. N. Schaffer and M. M. Pearson. "Proteus mirabilis and urinary tract infections". In: *Urinary Tract Infections: Molecular Pathogenesis and Clinical Management* (2017), pages 383–433.

[229] R. Mordi and M. Momoh. "Incidence of Proteus species in wound infections and their sensitivity pattern in the University of Benin Teaching Hospital". In: *African Journal of Biotechnology* 8.5 (2009).

[230] B. E. Murray. "The life and times of the Enterococcus." In: *Clinical microbiology reviews* 3.1 (1990), pages 46–65.

[231] "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic Acids Research* 49.D1 (2021), pages D480–D489.

[232] R. Sharma, P. Singh, W. Loughry, J. M. Lockhart, et al. "Zoonotic leprosy in the southeastern United States". In: *Emerging Infectious Diseases* 21.12 (2015), page 2127.

[233] B. Welte. *Zeitzeugen Aus Dem Wüstensand. Die Altägyptischen Mumienschädel Aus Abusir El-Meleq*. VML Verlag Marie Leidorf, 2016.

[234] L. S. Weyrich, S. Duchene, J. Soubrier, L. Arriola, et al. "Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus". In: *Nature* 544.7650 (2017), pages 357–361.

[235] I. N. Rôças, J. F. Siqueira Jr, K. R. Santos, A. M. Coelho, and R. de Janeiro. "'Red complex' (Bacteroides forsythus, Porphyromonas gingivalis, and Treponema denticola) in endodontic infections: a molecular approach". In: *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology* 91.4 (2001), pages 468–471.

[236] M. Göker, B. Held, S. Lucas, M. Nolan, et al. "Complete genome sequence of Olsenella uli type strain (VPI D76D-27C T)". In: *Standards in Genomic Sciences* 3.1 (2010), pages 76–84.

[237] R. J. Palmer Jr. "Composition and development of oral bacterial communities". In: *Periodontology 2000* 64.1 (2014), pages 20–39.

[238] A Sharma, R Somani, et al. "Dermatoglyphic interpretation of dental caries and its correlation to salivary bacteria interactions: An in vivo study". In: *Journal of Indian Society of Pedodontics and Preventive Dentistry* 27.1 (2009), page 17.

[239] W. L. Nicholson, N. Munakata, G. Horneck, H. J. Melosh, and P. Setlow. "Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments". In: *Microbiology and molecular biology reviews* 64.3 (2000), pages 548–572.

[240] S. Filippidou, T. Junier, T. Wunderlin, C.-C. Lo, et al. "Under-detection of endospore-forming Firmicutes in metagenomic data". In: *Computational and structural biotechnology journal* 13 (2015), pages 299–306.

[241] I. Can, G. T. Javan, A. E. Pozhitkov, and P. A. Noble. "Distinctive thanatomicrobiome signatures found in the blood and internal organs of humans". In: *Journal of microbiological methods* 106 (2014), pages 1–7.

[242] J. L. Pechal, T. L. Crippen, M. E. Benbow, A. M. Tarone, et al. "The potential use of bacterial community succession in forensics as described by high throughput metagenomic sequencing". In: *International Journal of Legal Medicine* 128.1 (2014), pages 193–205.

[243] J. L. Metcalf. "Estimating the postmortem interval using microbes: Knowledge gaps and a path to technology adoption". In: *Forensic Science International: Genetics* 38 (2019), pages 211–218.

[244] H. D. Donoghue, A. Marcsik, C. Matheson, K. Vernon, et al. "Co–infection of Mycobacterium tuberculosis and Mycobacterium leprae in human archaeological samples: a possible explanation for the historical decline of leprosy". In: *Proceedings of the Royal Society B: Biological Sciences* 272.1561 (2005), pages 389–394.

[245] C. A. Roberts, M. E. Lewis, and K. Manchester. *The past and present of leprosy: archaeological, historical, palaeopathological and clinical approaches: 3rd International Congress on the Evolution and palaeoepidemiology of the infectious diseases, ICEPID, 26-31 July 1999, University of Bradford; proceedings.* Archaeopress, 2002.

[246] G. Robbins, V. M. Tripathy, V. N. Misra, R. K. Mohanty, et al. "Ancient skeletal evidence for leprosy in India (2000 BC)". In: *PloS one* 4.5 (2009), e5669.

[247] R Dharmendra. "Leprosy in ancient Indian medicine". In: *Int J Lepr* 15.4 (1947), pages 424–430.

[248] V. Mariotti, O Dutour, M. Belcastro, F Facchini, and P Brasili. "Probable early presence of leprosy in Europe in a Celtic skeleton of the 4th–3rd century BC (Casalecchio di Reno, Bologna, Italy)". In: *International Journal of Osteoarchaeology* 15.5 (2005), pages 311–325.

[249] K. Köhler, A. Marcsik, P. Zádori, G. Biro, et al. "Possible cases of leprosy from the Late Copper Age (3780-3650 cal BC) in Hungary". In: *PloS one* 12.10 (2017), e0185966.

[250] D. Paraskevis, G. Magiorkinis, E. Magiorkinis, S. Y. Ho, et al. "Dating the origin and dispersal of hepatitis B virus infection in humans and primates". In: *Hepatology* 57.3 (2013), pages 908–916.

[251] I. E. Andernach, O. E. Hunewald, and C. P. Muller. "Bayesian inference of the evolution of HBV/E". In: *PLoS One* 8.11 (2013), e81690.

[252] G. Zehender, E. Ebranati, E. Gabanelli, R. Shkjezi, et al. "Spatial and temporal dynamics of hepatitis B virus D genotype in Europe and the Mediterranean Basin". In: *PloS one* 7.5 (2012), e37198.

[253] G. Ferrari, J. Neukamm, H. T. Baalsrud, A. M. Breidenstein, et al. "Variola virus genome sequenced from an eighteenth-century museum specimen supports the recent origin of smallpox". In: *Philosophical Transactions of the Royal Society B* 375.1812 (2020), page 20190572.

[254] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, I. D. Ladnyi, et al. *Smallpox and its eradication*. Volume 6. World Health Organization Geneva, 1988.

[255] D. R. Hopkins. *The greatest killer: smallpox in history*. Volume 793. University of Chicago Press, 2002.

[256] B Moss. "Fields virology". In: *Poxviridae. Philadelphia, PA: Lippincott, Williams, and Wilkins* (2007), pages 2905–45.

[257] Z. S. Moore, J. F. Seward, and J. M. Lane. "Smallpox". In: *The Lancet* 367.9508 (2006), pages 425–435.

[258] I. V. Babkin and I. N. Babkina. "The origin of the variola virus". In: *Viruses* 7.3 (2015), pages 1100–1112.

[259] A. L. Hughes, S. Irausquin, and R. Friedman. "The evolutionary biology of poxviruses". In: *Infection, Genetics and Evolution* 10.1 (2010), pages 50–59.

[260] S. N. Shchelkunov. "How long ago did smallpox virus emerge?" In: *Archives of virology* 154.12 (2009), pages 1865–1871.

[261] I. V. Babkin and I. N. Babkina. "A retrospective study of the orthopoxvirus molecular evolution". In: *Infection, Genetics and Evolution* 12.8 (2012), pages 1597–1604.

[262] C. Dixon. "Smallpox in Tripolitania, 1946: an epidemiological and clinical study of 500 cases, including trials of penicillin treatment". In: *Epidemiology & Infection* 46.4 (1948), pages 351–377.

[263] Y. C. Li, D. Gardner, S. Walsh, and M. Vitalis. "EA & Damon, IK 2007.'On The Origin of Smallpox: Correlating Variola Phylogenics with Historical Smallpox Record'". In: *PNAS* 104.40 (), pages 15–787.

[264] S. Marciniak and H. N. Poinar. "Ancient pathogens through human history: A paleogenomic perspective". In: *Paleogenomics*. Springer, 2018, pages 115–138.

[265] K. I. Bos, D. Kühnert, A. Herbig, L. R. Esquivel-Gomez, et al. "Paleomicrobiology: Diagnosis and evolution of ancient pathogens". In: *Annual review of microbiology* 73 (2019), pages 639–666.

[266] A. F. Porter, A. T. Duggan, H. N. Poinar, and E. C. Holmes. "Comment: characterization of two historic smallpox specimens from a Czech museum". In: *Viruses* 9.10 (2017), page 276.

[267] C. Thèves, P. Biagini, and E. Crubézy. "The rediscovery of smallpox". In: *Clinical Microbiology and Infection* 20.3 (2014), pages 210–218.

[268] R. J. Davenport, J. Boulton, and L. Schwarz. "Urban inoculation and the decline of smallpox mortality in eighteenth-century cities—a reply to R azzell". In: *The Economic history review* 69.1 (2016), pages 188–214.

[269] C. Duncan, S. Duncan, and S. Scott. "Oscillatory dynamics of smallpox and the impact of vaccination". In: *Journal of theoretical biology* 183.4 (1996), pages 447–454.

[270] R. J. Davenport, M. Satchell, and L. M. W. Shaw-Taylor. "The geography of smallpox in England before vaccination: A conundrum resolved". In: *Social Science & Medicine* 206 (2018), pages 75–85.

[271] J. Landers, L. John, et al. *Death and the metropolis: studies in the demographic history of London, 1670-1830*. 20. Cambridge University Press, 1993.

[272] *SurgiCat*. `http://surgicat.rcseng.ac.uk/`. 2019.

[273] A. M. Devault, G. B. Golding, N. Waglechner, J. M. Enk, et al. "Second-pandemic strain of Vibrio cholerae from the Philadelphia cholera outbreak of 1849". In: *New England Journal of Medicine* 370.4 (2014), pages 334–340.

[274] S. N. Shchelkunov, A. V. Totmenin, V. N. Loparev, P. F. Safronov, et al. "Alastrim smallpox variola minor virus genome DNA sequences". In: *Virology* 266.2 (2000), pages 361–386.

[275] S. N. Shchelkunov, A. V. Totmenin, and L. S. Sandakhchiev. "Analysis of the nucleotide sequence of 23.8 kbp from the left terminus of the genome of variola major virus strain India-1967". In: *Virus research* 40.2 (1996), pages 169–183.

[276] C. Afonso, E. Tulman, Z Lu, L Zsak, et al. "The genome of camelpox virus". In: *Virology* 295.1 (2002), pages 1–9.

[277] C. Smithson, N. Tang, S. Sammons, M. Frace, et al. "The genomes of three North American orthopoxviruses". In: *Virus genes* 53.1 (2017), pages 21–34.

[278] S. N. Shchelkunov, A. V. Totmenin, I. V. Babkin, P. F. Safronov, et al. "Human monkeypox and smallpox viruses: genomic comparison". In: *Febs letters* 509.1 (2001), pages 66–70.

[279] E. Tulman, G. Delhon, C. Afonso, Z. Lu, et al. "Genome of horsepox virus". In: *Journal of virology* 80.18 (2006), pages 9244–9258.

[280] P. W. Dabrowski, A. Radonić, A. Kurth, and A. Nitsche. "Genome-wide comparison of cowpox viruses reveals a new clade related to Variola virus". In: *PloS one* 8.12 (2013), e79953.

[281] L. Schrick, S. H. Tausch, P. W. Dabrowski, C. R. Damaso, et al. "An early American smallpox vaccine based on horsepox". In: *New England Journal of Medicine* 377.15 (2017), pages 1491–1492.

[282] R. R. Bouckaert and A. J. Drummond. "bModelTest: Bayesian phylogenetic site model averaging and model comparison". In: *BMC evolutionary biology* 17.1 (2017), page 42.

[283] J. F. C. Kingman. "The coalescent". In: *Stochastic processes and their applications* 13.3 (1982), pages 235–248.

[284] D. S. Carroll, G. L. Emerson, Y. Li, S. Sammons, et al. "Chasing Jenner's vaccine: revisiting cowpox virus classification". In: *PLoS One* 6.8 (2011), e23086.

[285] J. H. Marcus, C. Posth, H. Ringbauer, L. Lai, et al. "Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia". In: *Nature communications* 11.1 (2020), pages 1–14.

[286] D. C. Collins, R. Sundar, J. S. Lim, and T. A. Yap. "Towards precision medicine in the clinic: from biomarker discovery to novel therapeutics". In: *Trends in pharmacological sciences* 38.1 (2017), pages 25–40.

[287] J. Krause, P. H. Dear, J. L. Pollack, M. Slatkin, et al. "Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae". In: *Nature* 439.7077 (2006), pages 724–727.

[288] D. Gilbert. "The jfreechart class library". In: *Developer Guide. Object Refinery* 7 (2002).

[289] O. R. Limited. *Orson Charts*. `https://github.com/jfree/orson-charts`. 2020.

[290] G. L. Kay, M. J. Sergeant, Z. Zhou, J. Z.-M. Chan, et al. "Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe". In: *Nature communications* 6.1 (2015), pages 1–9.

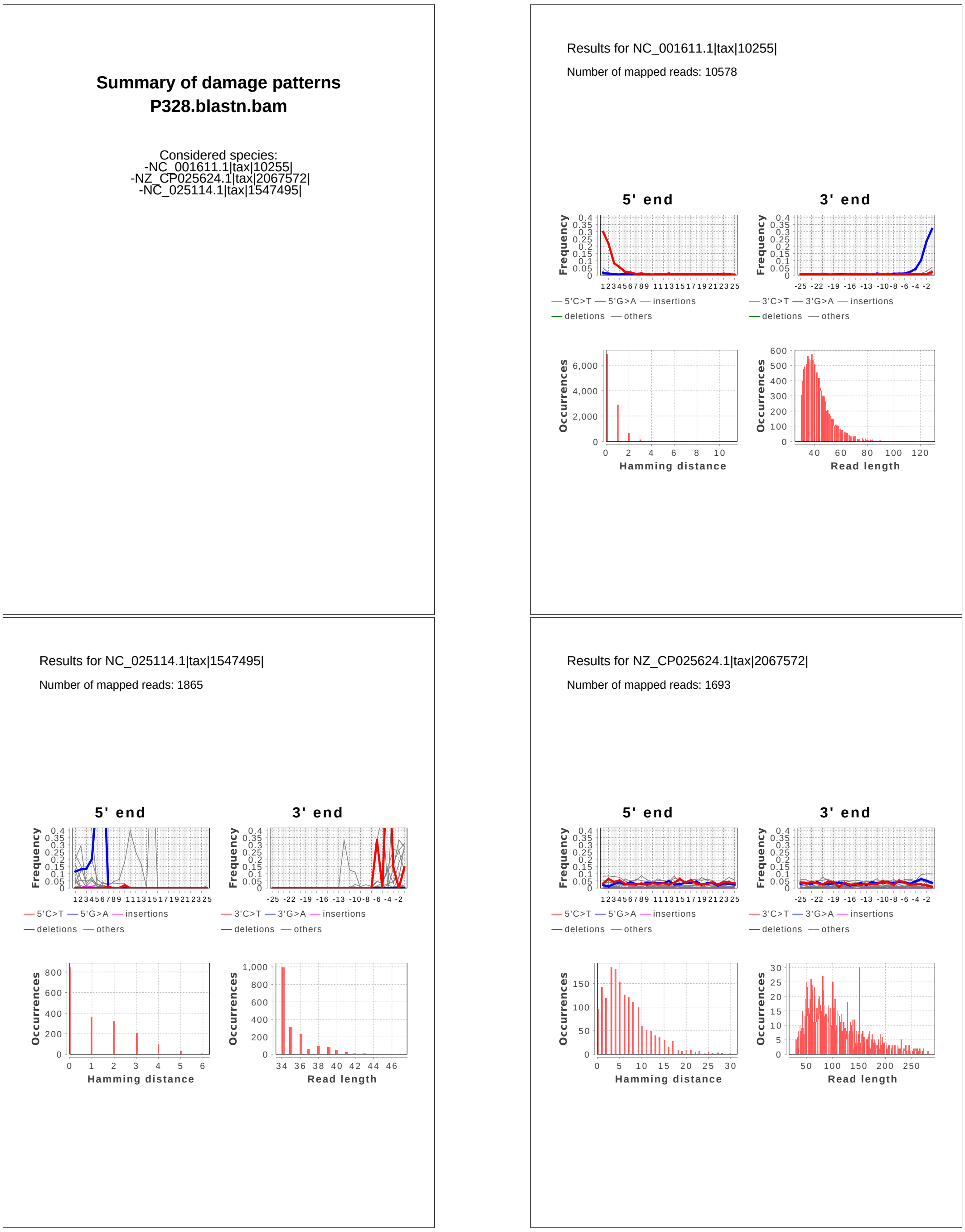
[291] Y.-L. Xiao, J. C. Kash, S. B. Beres, Z.-M. Sheng, et al. "High-throughput RNA sequencing of a formalin-fixed, paraffin-embedded autopsy lung tissue sample from the 1918 influenza pandemic". In: *The Journal of pathology* 229.4 (2013), pages 535–545.

[292] O. Smith, A. Clapham, P. Rose, Y. Liu, et al. "A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus". In: *Scientific reports* 4 (2014), page 4003.

APPENDIX A

Supplementary Information

A.1 Supplementary Figures



Summary of damage patterns
P328.blastn.bam

Considered species:
-NC_001611.1|tax|10255|
-NZ_CP025624.1|tax|2067572|
-NC_025114.1|tax|1547495|

Results for NC_001611.1|tax|10255|

Number of mapped reads: 10578

Results for NC_025114.1|tax|1547495|

Number of mapped reads: 1865

Results for NZ_CP025624.1|tax|2067572|

Number of mapped reads: 1693

Figure A.1: PDF summarizing the result of the metagenomic analysis with DamageProfiler. The damage patterns of each species are visualized on a separate page.

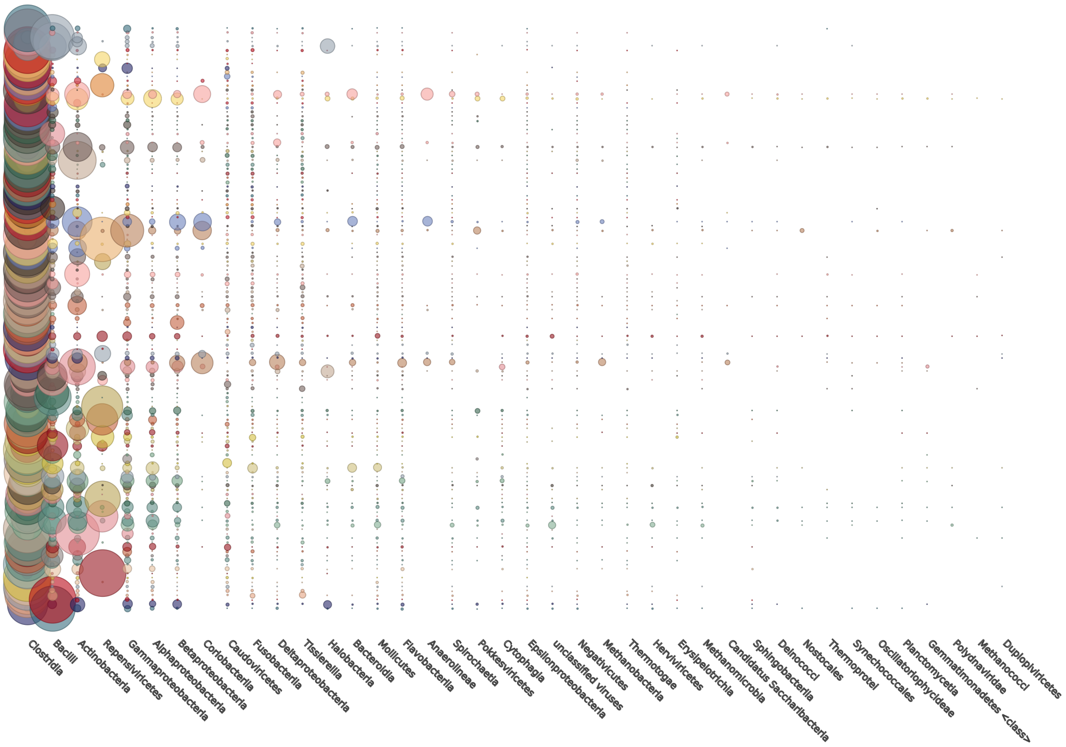


Figure A.2: Overview of the metagenomic results per sample on genus level. Each row/color represents a sample and the size of the circle symbolizes the number of reads assigned to the corresponding genus.

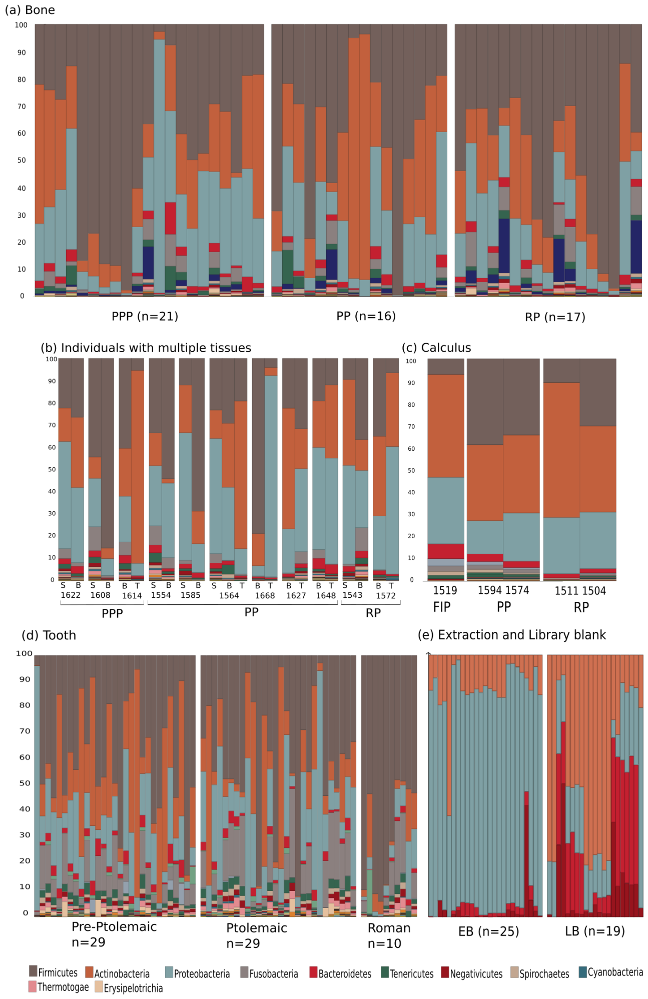


Figure A.3: Metagenomic composition of all samples. Comparison of bacterial and metagenomic composition of all samples, tissues, and time periods: first intermediate period (FIP), pre-Ptolemaic period (PPP), Ptolemaic period (PP), and Roman period (RP). Individual IDs Abusir<ID> in Figure B and C are abbreviated to <ID> for ease of reading. (a) Bacterial composition of all bone samples. Numbers indicate the number of samples per time period. (b) Comparison of metagenomic composition between different tissues of an individual. The letter 's' represents soft tissue, 'b' for bone, and 't' for teeth are used to distinguish between tissue types. (c) Comparison of the metagenomic composition of all calculus samples. (d) Comparison of the metagenomic composition of all dental samples over all time periods (calculus not included). (e) Metagenomic composition of all library and extraction blanks.

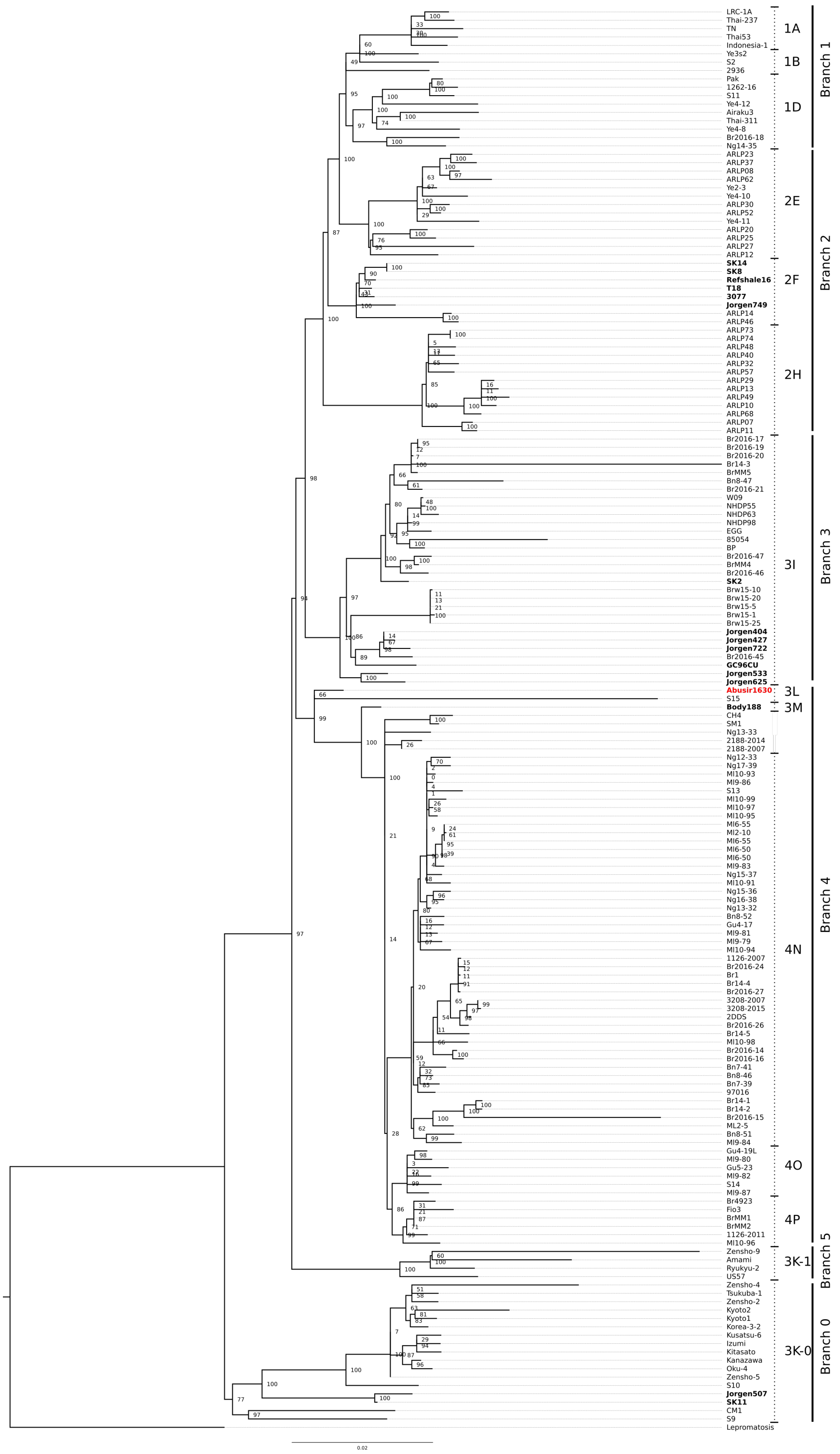


Figure A.4: Unfolded maximum likelihood tree of all published modern and ancient leprosy genomes, including the newly sequenced strain Abusir1630. The ancient genomes are in bold, the new strain Abusir1630 is highlighted in red. Bootstrap values are given as node labels.

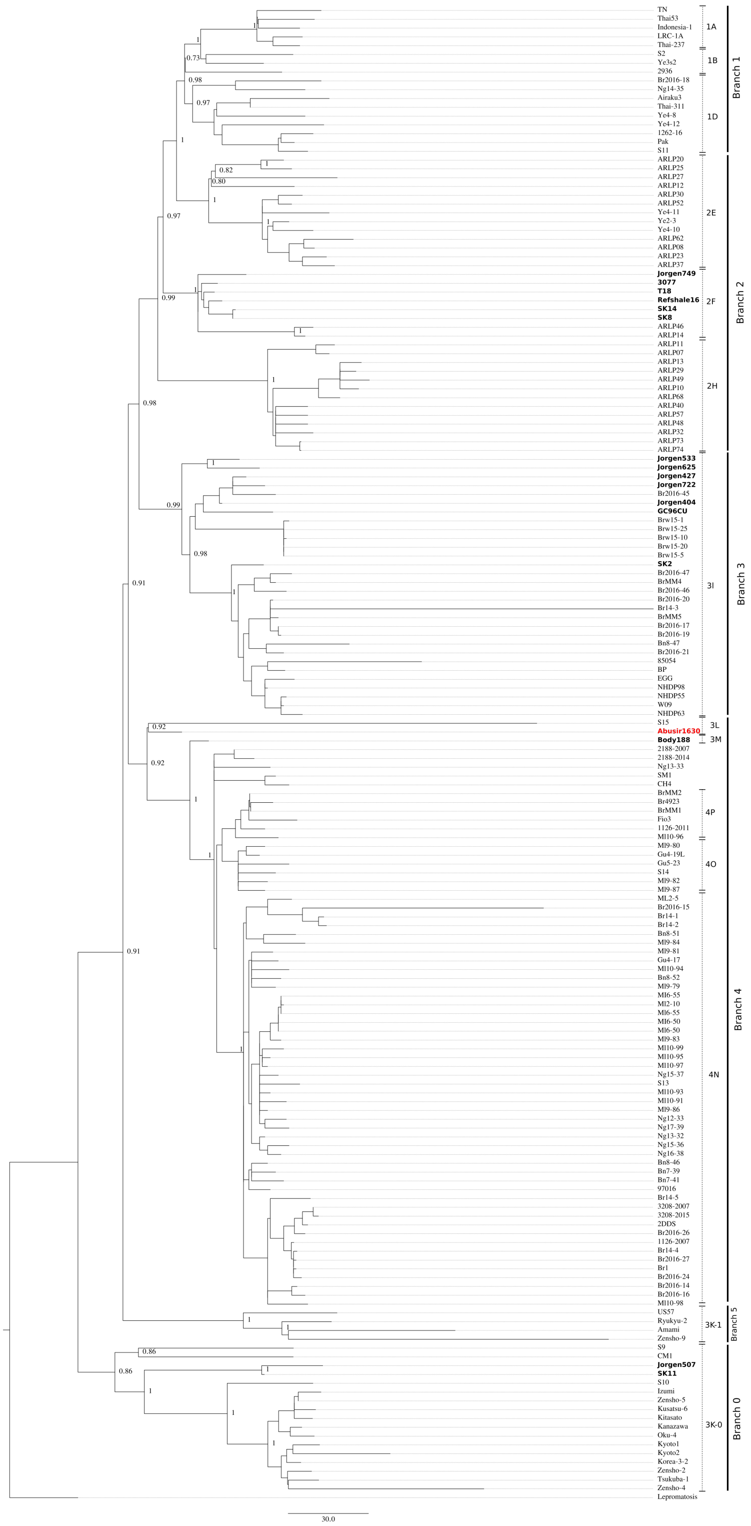


Figure A.5: Unfolded maximum parsimony tree of all published modern and ancient leprosy genomes, including the newly sequenced strain Abusir1630. The ancient genomes are in bold, the new strain Abusir1630 is highlighted in red. Bootstrap values are given as node labels.

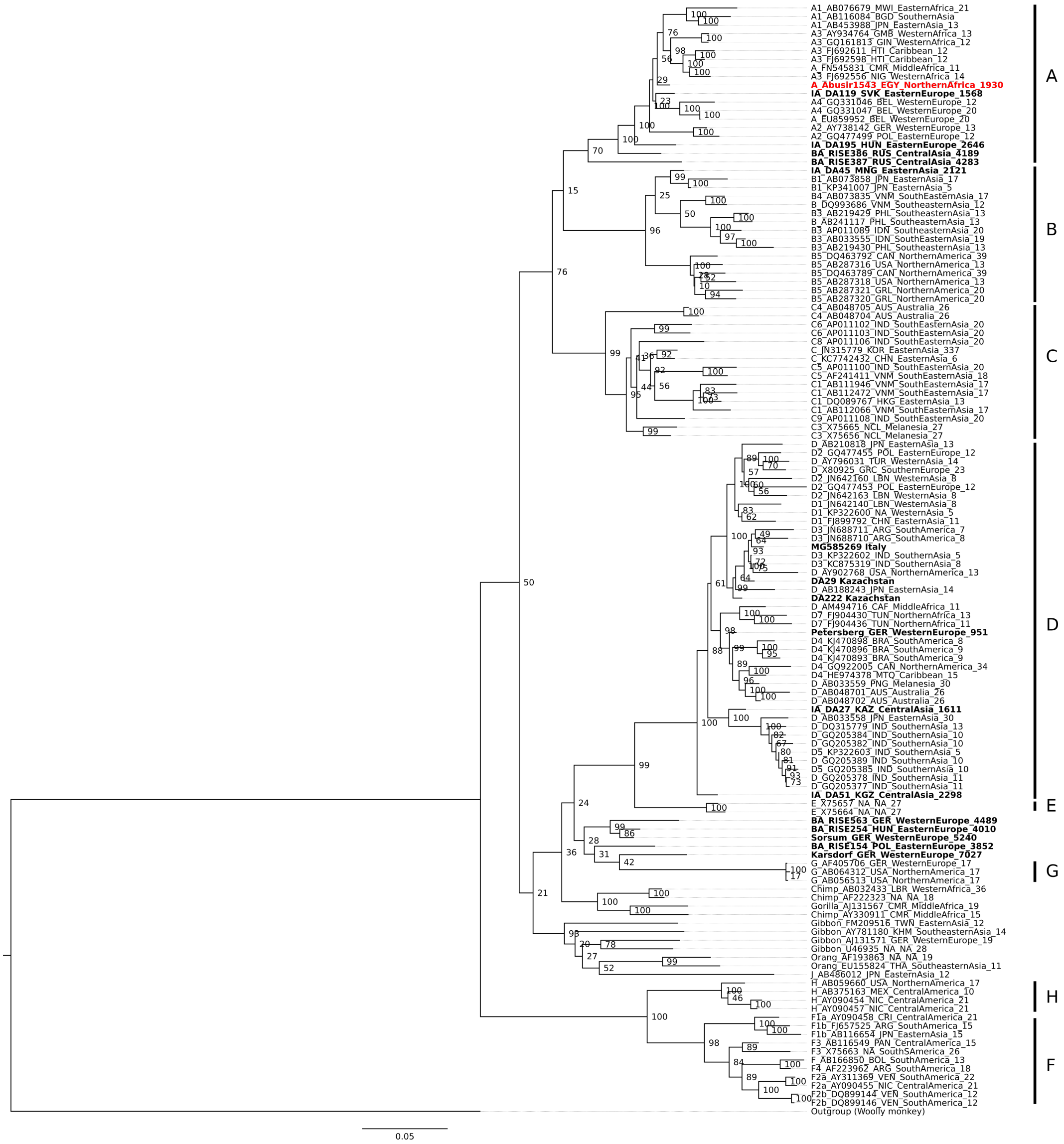
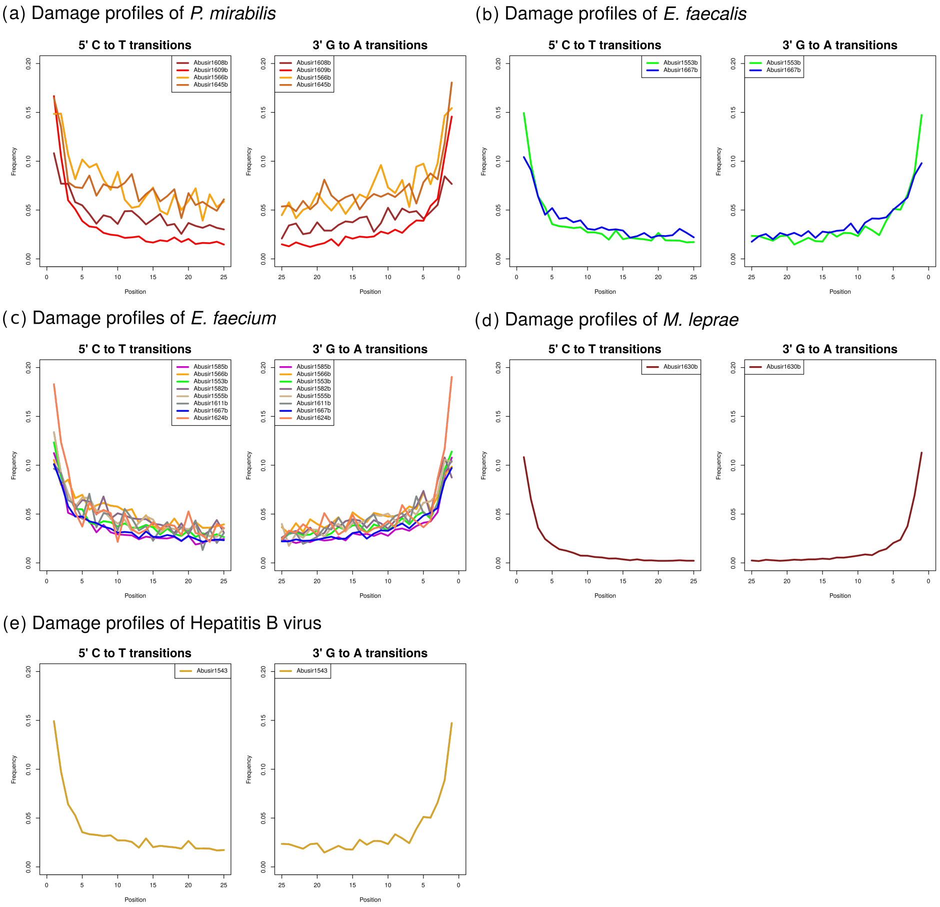

A1_AB076679_MWI_EasternAfrica_21
A1_AB116084_BGD_SouthernAsia
A1_AB453988_JPN_EasternAsia_13
A3_AY934764_GMB_WesternAfrica_13
A3_GQ161813_GIN_WesternAfrica_12
A3_FJ692611_HTI_Caribbean_12
A3_FJ692598_HTI_Caribbean_12
A_FN545831_CMR_MiddleAfrica_11
A3_FJ692556_NIG_WesternAfrica_14
A_Abusir1543_EGY_NorthernAfrica_1930
IA_DA119_SVK_EasternEurope_1568
A4_GQ331046_BEL_WesternEurope_12
A4_GQ331047_BEL_WesternEurope_20
A_EU859952_BEL_WesternEurope_20
A2_AY738142_GER_WesternEurope_13
A2_GQ477499_POL_EasternEurope_12
IA_DA195_HUN_EasternEurope_2646
BA_RISE386_RUS_CentralAsia_4189
BA_RISE387_RUS_CentralAsia_4283
IA_DA45_MNG_EasternAsia_2121
B1_AB073858_JPN_EasternAsia_17
B1_KP341007_JPN_EasternAsia_5
B4_AB073835_VNM_SouthEasternAsia_17
B_DQ993686_VNM_SoutheasternAsia_12
B3_AB219429_PHL_SoutheasternAsia_13
B_AB241117_PHL_SoutheasternAsia_13
B3_AP011089_IDN_SoutheasternAsia_20
B3_AB033555_IDN_SouthEasternAsia_19
B3_AB219430_PHL_SoutheasternAsia_13
B5_DQ463792_CAN_NorthernAmerica_39
B5_AB287316_USA_NorthernAmerica_13
B5_DQ463789_CAN_NorthernAmerica_39
B5_AB287318_USA_NorthernAmerica_13
B5_AB287321_GRL_NorthernAmerica_20
B5_AB287320_GRL_NorthernAmerica_20
C4_AB048705_AUS_Australia_26
C4_AB048704_AUS_Australia_26
C6_AP011102_IND_SouthEasternAsia_20
C6_AP011103_IND_SouthEasternAsia_20
C8_AP011106_IND_SouthEasternAsia_20
C_JN315779_KOR_EasternAsia_337
C_KC7742432_CHN_EasternAsia_6
C5_AP011100_IND_SouthEasternAsia_20
C5_AF241411_VNM_SouthEasternAsia_18
C1_AB111946_VNM_SouthEasternAsia_17
C1_AB112472_VNM_SouthEasternAsia_17
C1_DQ089767_HKG_EasternAsia_13
C1_AB112066_VNM_SouthEasternAsia_17
C9_AP011108_IND_SouthEasternAsia_20
C3_X75665_NCL_Melanesia_27
C3_X75656_NCL_Melanesia_27
D_AB210818_JPN_EasternAsia_13
D2_GQ477455_POL_EasternEurope_12
D_AY796031_TUR_WesternAsia_14
D_X80925_GRC_SouthernEurope_23
D2_JN642160_LBN_WesternAsia_8
D2_GQ477453_POL_EasternEurope_12
D2_JN642163_LBN_WesternAsia_8
D1_JN642140_LBN_WesternAsia_8
D1_KP322600_NA_WesternAsia_5
D1_FJ899792_CHN_EasternAsia_11
D3_JN688711_ARG_SouthAmerica_7
D3_JN688710_ARG_SouthAmerica_8
MG585269 Italy
D3_KP322602_IND_SouthernAsia_5
D3_KC875319_IND_SouthernAsia_8
D_AY902768_USA_NorthernAmerica_13
DA29 Kazachstan
D_AB188243_JPN_EasternAsia_14
DA222 Kazachstan
D_AM494716_CAF_MiddleAfrica_11
D7_FJ904430_TUN_NorthernAfrica_13
D7_FJ904436_TUN_NorthernAfrica_11
Petersberg_GER_WesternEurope_951
D4_KJ470898_BRA_SouthAmerica_8
D4_KJ470896_BRA_SouthAmerica_9
D4_KJ470893_BRA_SouthAmerica_9
D4_GQ922005_CAN_NorthernAmerica_34
D4_HE974378_MTQ_Caribbean_15
D_AB033559_PNG_Melanesia_30
D_AB048701_AUS_Australia_26
D_AB048702_AUS_Australia_26
IA_DA27_KAZ_CentralAsia_1611
D_AB033558_JPN_EasternAsia_30
D_DQ315779_IND_SouthernAsia_13
D_GQ205384_IND_SouthernAsia_10
D_GQ205382_IND_SouthernAsia_10
D5_KP322603_IND_SouthernAsia_5
D_GQ205389_IND_SouthernAsia_10
D5_GQ205385_IND_SouthernAsia_10
D_GQ205378_IND_SouthernAsia_11
D_GQ205377_IND_SouthernAsia_11
IA_DA51_KGZ_CentralAsia_2298
E_X75657_NA_NA_27
E_X75664_NA_NA_27
BA_RISE563_GER_WesternEurope_4489
BA_RISE254_HUN_EasternEurope_4010
Sorsum_GER_WesternEurope_5240
BA_RISE154_POL_EasternEurope_3852
Karsdorf_GER_WesternEurope_7027
G_AF405706_GER_WesternEurope_17
G_AB064312_USA_NorthernAmerica_17
G_AB056513_USA_NorthernAmerica_17
Chimp_AB032433_LBR_WesternAfrica_36
Chimp_AF222323_NA_NA_18
Gorilla_AJ131567_CMR_MiddleAfrica_19
Chimp_AY330911_CMR_MiddleAfrica_15
Gibbon_FM209516_TWN_EasternAsia_12
Gibbon_AY781180_KHM_SoutheasternAsia_14
Gibbon_AJ131571_GER_WesternEurope_19
Gibbon_U46935_NA_NA_28
Orang_AF193863_NA_NA_19
Orang_EU155824_THA_SoutheasternAsia_11
J_AB486012_JPN_EasternAsia_12
H_AB059660_USA_NorthernAmerica_17
H_AB375163_MEX_CentralAmerica_10
H_AY090454_NIC_CentralAmerica_21
H_AY090457_NIC_CentralAmerica_21
F1a_AY090458_CRI_CentralAmerica_21
F1b_FJ657525_ARG_SouthAmerica_15
F1b_AB116654_JPN_EasternAsia_15
F3_AB116549_PAN_CentralAmerica_15
F3_X75663_NA_SouthSAmerica_26
F_AB166850_BOL_SouthAmerica_13
F4_AF223962_ARG_SouthAmerica_18
F2a_AY311369_VEN_SouthAmerica_22
F2a_AY090455_NIC_CentralAmerica_21
F2b_DQ899144_VEN_SouthAmerica_12
F2b_DQ899146_VEN_SouthAmerica_12
Outgroup (Woolly monkey)
A
B
C
D
E
G
H
F
0.05

Figure A.6: Unfolded maximum likelihood tree based on 129 HBV genomes (Appendix, Table A.1). The ancient genomes are in bold, the newly sequenced genome in red and bold. The bootstrap values are given for the main branches as node labels.



(a) Damage profiles of *P. mirabilis*

(b) Damage profiles of *E. faecalis*

(c) Damage profiles of *E. faecium*

(d) Damage profiles of *M. leprae*

(e) Damage profiles of Hepatitis B virus

Figure A.7: Combined damage profiles of pathogens identified in bone samples. Damage profile of reads mapping to (a) *Proteus mirabilis*, (b) *Enterococcus faecalis*, (c) *Enterococcus faecium*, (d) *Mycobacterium leprae* (sample Abusir1630b), and (e) hepatitis B virus (combined libraries of individual Abusir1543).

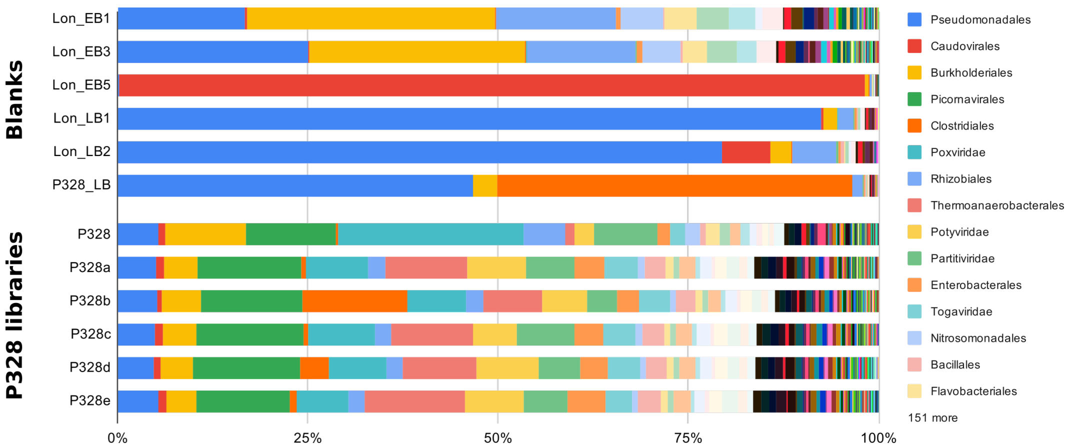


Figure A.8: Metagenomic composition of all libraries for sample P328, library and extraction blanks at the order level. All P328 libraries show a high amount of Poxviridae (turquoise), which is completely absent in the blank.

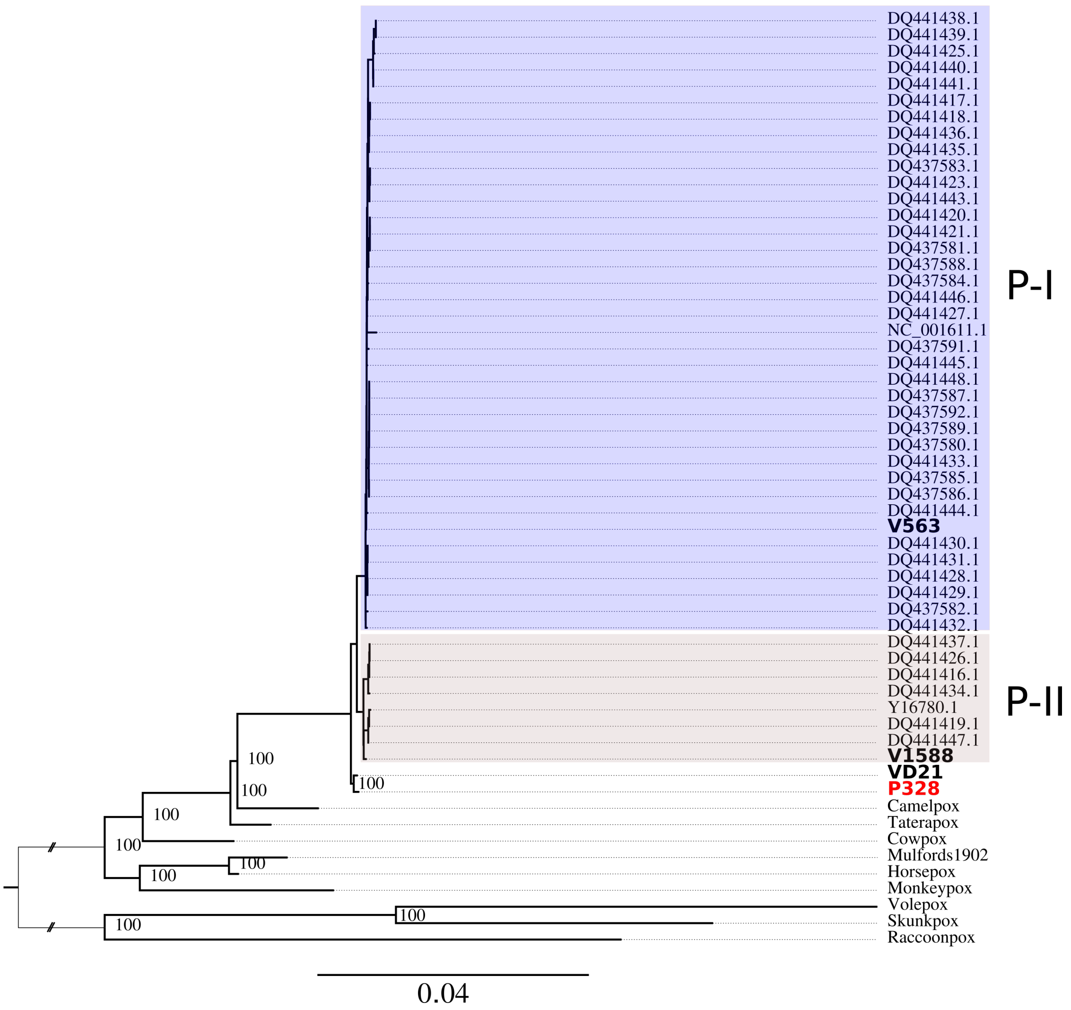


Figure A.9: Unfolded Maximum Likelihood tree including 57 Orthopoxvirus genomes. Bootstrap values are given as node labels. The historic genomes are in bold, the newly added genome in red. Bootstrap values are given for the main branches as node labels.

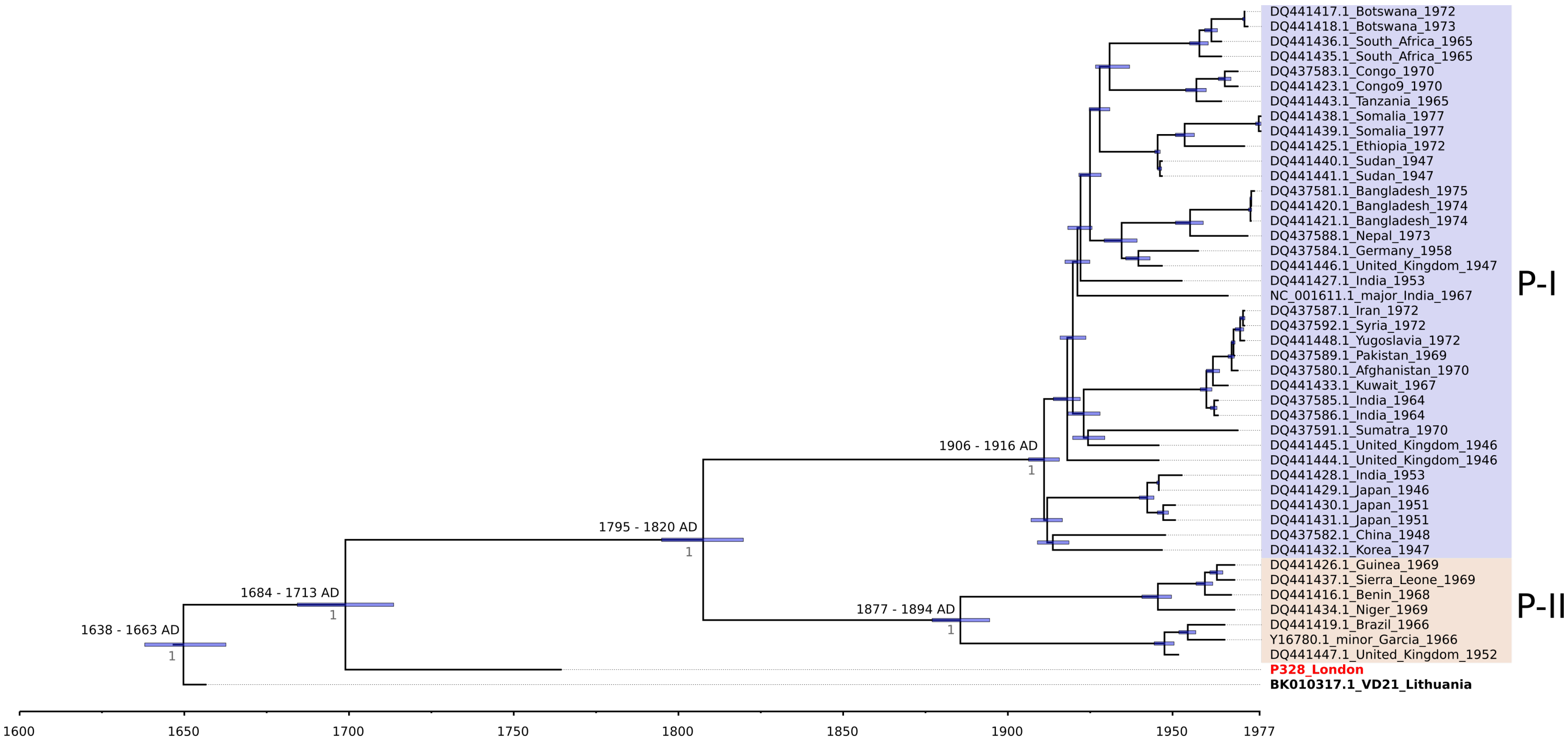


Figure A.10: Dated Bayesian Maximum Clade Credibility tree reconstructed with BEAST 2.5.5 (using a strict clock and constant population size) excluding strains V1588 and V563. The nodes are labeled with the 95% HPD interval. Historic genomes are in bold, the newly added genome in red. Posterior values are given as node labels in grey.

A.2 Supplementary Tables

Table A.1: Listing of all publish available HBV genomes that were used for the phylogenetic reconstruction. The accession IDs in italic and bold indicate the strains used for the maximum likelihood tree reconstruction.

FN545831
AB116092
AB194951
AF297621
AM184125
AY161138
AY233275
AY233280
AY233290
EU366129
EU859930
FJ692558
FJ692592
FJ692596
FJ692608
FM199978
FN545826
FN545828
FN545833
JN182323
JN182327
JQ023661
JQ707397
KF214660
KF922415
KF922430
KF922434
KJ854685
KJ854693
KM606737
KP168428
KP168435
KT151612
KU605533
KU605537
KX357650
EU859952
ERS3636018
ERS3636025
ERS3636093
ERS3636094
ERS3636095
ERS3636096
ERS3636097
ERS3636098
ERS3636099
ERS3636100
ERS3636101
AB116084
AB453988
AB076679
AY738142
GQ477499
AY934764
FJ692556
FJ692598
FJ692611
GQ161813
GQ331046
GQ331047
DQ993686
KP341007
AB010289
AB014366
AB031267
AB073821
AB073836
AB073849
AB073853
AB100695
AB106884
AB212625
AB231909
AB246340
AB287315
AB287323
AB300371

FJ562260
GQ358137
GQ358146
GQ358148
GQ358151
GQ924621
GQ924624
GQ924626
GQ924630
GQ924635
GQ924637
GQ924641
GQ924645
GQ924656
HM011466
HM011467
HM011471
HM011475
HM011476
HM011478
HM011482
HM011483
HM011487
HM011490
HM011496
HM011499
HM011503
HQ700546
JF436921
JN827419
JQ027311
JQ027312
JQ027313
JQ027330
JQ027334
JQ429081
JQ707737
JX026879
JX661471
KC774370
KJ173297
KJ173342
KJ173379
KJ173401
KJ410502
KJ790200
KJ803795
KJ803796
KJ803805
KJ803808
KJ803817
KJ803820
KM875420
KP148414
KP148452
KP148582
KP406278
KP659249
KP659250
KU964274
KU964383
KX276770
KX276772
KX276774
KX276783
KX276785
KX276786
KX276787
KX276791
KX276792
KX276794
KX276795
KX276796
KX276797
KX276798
KX276800
KX276806

AB014371
AB049610
AB074047
AB111120
AB112063
AB113878
AB115417
AB176642
AB195930
AB195931
AB198079
AB300361
AB367392
AB367420
AB670258
AB670259
AB670263
AB670285
AB670295
AB670298
AB697502
AB697510
AB900109
AB900113
AB931170
AP011099
AY167091
AY206386
AY641559
AY641563
D23682
DQ089768
DQ089777
DQ089785
DQ089788
DQ089790
DQ089795
DQ089802
DQ089804
DQ890381
EU306725
EU498227
EU522071
EU670263
EU796069
EU939539
EU939568
EU939597
FJ386601
FJ386626
FJ386644
FJ787452
FJ899767
FJ899783
GQ358157
GQ358158
GQ377586
GQ377632
GQ475321
GQ924642
GQ924643
HM011479
HM011488
HM011495
HQ700456
HQ700506
HQ700516
HQ700517
HQ700522
HQ700564
HQ700575
HQ700576
JF828925
JN827415
JN827421
JQ027317
JQ027324

KJ410494
KJ410496
KJ410515
KJ803766
KJ803777
KJ803779
KJ803790
KJ803809
KJ803818
KJ803823
KJ803826
KR013837
KR013859
KR013871
KT364751
KU679937
KU679947
KU679951
KU964045
KU964236
KU964358
KX276836
KX276841
KX276844
KX276846
KX276848
KX276850
KX276853
KX276855
AB112472
AB111946
AB112066
DQ089767
X75656
X75665
AB048705
AB048704
AP011100
AF241411
AP011103
AP011102
AP011106
AP011108
AB048702
AB188243
AB210818
AM494716
AY796031
AY902768
DQ315779
GQ205377
GQ205378
GQ205384
GQ205389
KC875319
KP322600
KP322602
KP322603
X80925
AB119255
AB188241
AB270541
AB330367
AB555496
AB555500
AB674416
AB674424
AB674425
AB674427
AB674428
AF043594
AJ344116
AJ627218
AJ627220
AJ627222
AY090452
DQ304548

GQ922002
GQ922003
GU456637
GU456643
GU456651
GU456654
GU456658
GU456665
GU456669
GU456674
GU456678
GU456679
GU456682
GU456684
HQ700449
HQ700458
HQ700510
HQ700513
JF754588
JF754592
JF754597
JF754611
JF754612
JF754617
JF754631
JN040762
JN040766
JN040768
JN040769
JN040779
JN040818
JN040822
JN257160
JN257162
JN257165
JN257172
JN257177
JN257190
JN257202
JN642133
JN642135
JN642136
JN642149
JN642159
JN664913
JN664919
JN664920
JN664921
JN664922
JN664931
JN664932
JN664936
JN688683
JN688695
JN688712
JN688713
JN792912
JQ707529
JQ707699
JX470760
KC774444
KC875342
KF192832
KF679990
KJ647353
KJ647355
KJ843187
KM524338
KM524358
KM577668
KP090181
KP168419
KU668435
KU736927
KX357622
L27106
X65258

AB219533
AB274977
EU239220
FN545823
FN594751
FN594760
GQ161775
HM363586
HM363592
HM363603
HM363611
KU736913
AB166850
AB214516
AB365453
DQ899142
DQ899145
DQ899148
JN688720
JN792921
JQ272888
KF199901
KJ638660
KJ638662
KJ676694
X75658
AY090458
FJ657525
AB116654
AY311369
AY090455
DQ899146
DQ899144
AB116549
X75663
AF223962
AB056513
AF405706
AB064312
HE981175
AB375163
AY090454
AY090457
AB059659
AB059660
AB486012
AF222323
FM209516
AJ131567
AY781180
EU155824
U46935
AB032432
AB823658
AB823659
AB823660
AB823661
AB823662
AF193864
AF222322
AF242586
AF498266
AJ131569
AJ131574
AM117396
AY077735
AY077736
AY330912
AY330913
AY330914
AY330915
AY330916
AY330917
AY781182
AY781186
AY781187
EU155821

AB032433
AB555498
AB642093
AB642101
AB713528
AB828708
AB900098
AB900107
AB931169
AY206383
AY206391
AY800392
DQ463791
DQ993684
DQ995802
DQ995804
EF494381
EU158262
EU522072
EU939634
EU939677
EU939678
FJ023631
FJ032342
FJ386582
FJ386648
KX276807
KX276812
KX276813
KX276815
KX276817
KX276819
KX276821
KX276825
KX276827
KX276830
KX276858
AB241117
AB073858
AB219430
AP011089
AB219429
AB033555
AB073835
AB287316
AB287318
AB287320
DQ463789
DQ463792
AB287321
KC774243
JQ040132
JQ040167
JQ429078
JQ801498
JX507211
JX870000
KC774180
KC774196
KC774214
KC774226
KC774238
KC774240
KC774244
KC774297
KC774302
KC774351
KC774352
KC774357
KF214670
KF214673
KF873514
KF873526
KF873536
KF873541
KF873544
JQ664503

MG585269
DQ315778
DQ399006
DQ464173
DQ464174
DQ486025
EU155893
EU414139
EU414140
EU414141
EU594432
EU594434
FJ904395
FJ904402
FJ904422
FJ904426
FJ904427
FJ904433
FJ904438
FJ904439
FJ904445
GQ167302
GQ184322
GQ205380
GQ477452
GQ477456
X80926
GQ205382
GQ205385
JN315779
AB048701
AB033558
AB033559
FJ899792
JN642140
GQ477455
JN642160
GQ477453
JN642163
JN688710
JN688711
HE974378
KJ470898
KJ470896
GQ922005
KJ470893
FJ904436
FJ904430
X75664
X75657
AB106564
EU155829
FJ798098
AF193863
AJ131571
AY330911
ERR2299806
ERR2299807
ERR2299808
LT992438
LT992439
LT992440
LT992441
LT992442
LT992443
LT992444
LT992447
LT992448
LT992454
LT992455
LT992459

Table A.2: List of pathogens that were detected in the oral samples. The last column indicates if a lesion could be detected by visual inspection of the samples (neg=negative, n.e.=not evaluable, n.i.=no information, pos=positive).

Sample	Genetically detected pathogen	Visual Indication			
		Dental Caries	Parodontose	Parodontitis	Calculus
Abusir1433t	F. alocis	neg	NA	NA	neg
Abusir1504c	-	n.e.	neg	neg	n.i.
Abusir1518t	*T. forsythia*	pos	neg	neg	n.i.
Abusir1519c	*T. forsythia, P. Gingivalis, T. denticola, F. alocis, O. uli*	neg	pos	pos	n.i.
Abusir1521t	*O. uli*	pos	n.i.	n.i.	neg
Abusir1563t	*F. alocis*	neg	pos	neg	n.i.
Abusir1564t	*O. uli*	neg	n.i.	n.i.	pos
Abusir1580t	*T. forsythia, P. Gingivalis*	n.e.	pos	pos	n.i.
Abusir1584t	*T. forsythia*	neg	pos	pos	n.i.
Abusir1594c	*T. forsythia, P. Gingivalis, T. denticola, O. uli*	neg	pos	pos	n.i.
Abusir1614t	*F. alocis, O. uli*	n.e.	pos	n.e.	n.i.
Abusir1618t	*T. forsythia, P. Gingivalis*	pos	pos	pos	n.i.
Abusir1627t	-	n.e.	pos	n.e.	n.i.
Abusir1650t	*T. forsythia, P. Gingivalis, T. denticola, F. alocis*	pos	pos	neg	n.i.
Abusir1655t	*O. uli*	neg	n.i.	n.i.	pos
Abusir1660t	*F. alocis*	pos	pos	neg	pos
Abusir1668t	-	n.e.	n.e.	n.e.	n.i.
Abusir3533t	-	neg	n.i.	n.i	neg

Table A.3: Metagenomic composition of all six P328 and six non-template libraries on different levels (domain, order, species).

	Samples						Blanks					
	P328	P328a	P328b	P328c	P328d	P328e	Lon_EB1	Lon_EB3	Lon_EB5	Lon_LB1	Lon_LB2	P328_LB
Level: Domain												
Bacteria	42.34%	43.58%	51.22%	43.61%	43.90%	48.17%	98.67%	99.68%	0.68%	99.42%	93.11%	99.84%
Archaea	0.29%	0.85%	0.69%	0.93%	0.79%	0.98%	0.04%	0.06%	0.13%	0.00%	0.00%	0.00%
Viruses	57.37%	55.57%	48.09%	55.46%	55.31%	50.84%	1.28%	0.26%	99.19%	0.58%	6.89%	0.15%
Level: Order												
Pseudomonadales	5.25%	5.08%	5.23%	4.86%	4.77%	5.25%	16.74%	25.00%	0.12%	92.43%	79.40%	46.61%
Caudovirales	0.93%	0.93%	0.61%	1.01%	0.78%	1.05%	0.26%	0.18%	98.07%	0.30%	6.35%	0.04%
Burkholderiales	10.61%	4.49%	5.06%	4.44%	4.24%	4.05%	32.50%	28.34%	0.50%	1.79%	2.67%	3.13%
Picornavirales	11.93%	13.63%	13.38%	14.04%	14.15%	12.26%	0.03%	0.01%	0.00%	0.04%	0.00%	0.09%
Clostridiales	0.26%	0.61%	13.71%	0.68%	3.84%	0.84%	0.13%	0.14%	0.04%	0.03%	0.17%	46.61%
Poxviridae	24.38%	8.14%	7.81%	8.71%	7.57%	6.85%	0.06%	0.04%	0.00%	0.05%	0.06%	0.00%
Rhizobiales	5.48%	2.30%	2.24%	2.25%	2.15%	2.06%	15.67%	14.32%	0.29%	2.08%	5.80%	1.33%
Thermoanaerobacterales	1.17%	10.73%	7.75%	10.67%	9.67%	13.19%	0.00%	0.04%	0.01%	0.00%	0.00%	0.07%
Potyviridae	2.56%	7.79%	5.85%	5.73%	8.13%	7.76%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%
Partitiviridae	8.27%	6.25%	3.97%	7.58%	5.45%	5.86%	0.05%	0.04%	0.01%	0.04%	0.18%	0.12%
Enterobacterales	1.80%	4.05%	2.86%	3.82%	3.66%	4.86%	0.63%	0.81%	0.01%	0.29%	0.32%	0.16%
Togaviridae	1.89%	4.38%	4.02%	4.23%	4.26%	3.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Nitrosomonadales	2.03%	0.85%	0.86%	0.87%	0.78%	0.82%	5.56%	5.02%	0.09%	0.10%	0.09%	0.03%
Bacillales	0.67%	2.74%	2.57%	2.89%	2.72%	3.01%	0.22%	0.26%	0.01%	0.13%	0.40%	0.06%
Flavobacteriales	1.81%	1.07%	0.92%	0.84%	0.91%	0.95%	4.29%	3.21%	0.12%	0.05%	0.19%	0.04%
Sphingomonadales	1.42%	0.74%	0.71%	0.71%	0.71%	0.68%	4.20%	3.88%	0.07%	0.26%	0.41%	0.22%
Flaviviridae	1.42%	2.08%	1.71%	2.19%	2.15%	2.25%	0.00%	0.00%	0.00%	0.00%	0.00%	0.05%
Cytophagales	1.20%	0.68%	0.57%	0.61%	0.57%	0.51%	3.45%	2.64%	0.04%	0.02%	0.00%	0.02%
ssRNA viruses	0.82%	1.41%	1.59%	1.68%	1.52%	1.43%	0.95%	0.04%	0.05%	0.13%	0.32%	0.03%
Micrococcales	0.81%	0.53%	0.46%	0.50%	0.50%	0.54%	2.62%	2.56%	0.08%	0.37%	0.52%	0.11%
Nidovirales	0.74%	1.53%	1.48%	1.83%	1.67%	1.65%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Spirochaetales	0.21%	1.61%	1.11%	1.70%	1.42%	2.10%	0.01%	0.02%	0.00%	0.00%	0.00%	0.02%
dsRNA viruses	0.54%	1.05%	0.98%	1.10%	1.15%	1.21%	0.01%	0.01%	0.01%	0.01%	0.03%	0.02%
Astroviridae	1.39%	0.95%	0.86%	1.02%	1.00%	0.84%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Lactobacillales	0.63%	0.92%	0.66%	0.81%	0.82%	1.01%	0.16%	0.23%	0.00%	0.13%	0.33%	0.05%
Mycoplasmatales	0.69%	0.98%	0.74%	1.03%	0.90%	1.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
unclassified viruses	0.92%	0.93%	0.73%	1.04%	0.88%	0.83%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Bacteroidetes Order II. Incertae sedis	0.02%	1.04%	0.73%	0.97%	1.00%	1.38%	0.01%	0.00%	0.00%	0.00%	0.00%	0.01%
Corynebacteriales	0.63%	0.34%	0.34%	0.35%	0.37%	0.28%	0.91%	0.93%	0.02%	0.33%	0.57%	0.13%
Totiviridae	0.02%	1.17%	0.79%	0.84%	1.16%	0.98%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Caulobacterales	0.44%	0.22%	0.22%	0.18%	0.21%	0.20%	1.55%	1.27%	0.03%	0.20%	0.37%	0.05%
Campylobacterales	0.18%	0.65%	0.67%	0.85%	0.65%	0.73%	0.09%	0.09%	0.00%	0.00%	0.00%	0.02%
Candidatus Nanopelagicales	0.44%	0.20%	0.20%	0.18%	0.20%	0.17%	1.36%	1.10%	0.03%	0.00%	0.00%	0.00%
Propionibacteriales	0.45%	0.27%	0.28%	0.26%	0.28%	0.27%	0.46%	0.48%	0.01%	0.34%	0.63%	0.05%
Oceanospirillales	1.09%	0.47%	0.46%	0.49%	0.42%	0.39%	0.11%	0.27%	0.00%	0.00%	0.00%	0.02%
Xanthomonadales	0.25%	0.24%	0.23%	0.18%	0.21%	0.22%	0.69%	0.66%	0.02%	0.32%	0.37%	0.21%
Thermococcales	0.03%	0.60%	0.51%	0.70%	0.57%	0.75%	0.00%	0.00%	0.13%	0.00%	0.00%	0.01%
Tymovirales	0.13%	0.48%	0.60%	0.80%	0.63%	0.49%	0.00%	0.00%	0.00%	0.00%	0.00%	0.03%
Virgaviridae	0.03%	0.61%	0.48%	0.62%	0.59%	0.67%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Rhodobacterales	0.27%	0.20%	0.14%	0.12%	0.16%	0.17%	0.68%	0.80%	0.01%	0.09%	0.11%	0.03%
Bromoviridae	0.03%	0.53%	0.40%	0.54%	0.54%	0.68%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%
Nostocales	0.09%	0.56%	0.33%	0.55%	0.51%	0.63%	0.03%	0.04%	0.00%	0.00%	0.00%	0.00%
Hypoviridae	0.39%	0.46%	0.40%	0.46%	0.48%	0.41%	0.00%	0.01%	0.00%	0.00%	0.00%	0.02%
Rhodocyclales	0.27%	0.19%	0.15%	0.16%	0.15%	0.14%	0.58%	0.65%	0.03%	0.03%	0.00%	0.03%
Herpesvirales	0.26%	0.41%	0.39%	0.49%	0.42%	0.28%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Rhodospirillales	0.20%	0.19%	0.15%	0.16%	0.18%	0.19%	0.36%	0.45%	0.02%	0.03%	0.10%	0.04%
Neisseriales	0.29%	0.17%	0.15%	0.16%	0.20%	0.16%	0.34%	0.34%	0.02%	0.03%	0.00%	0.03%
Phycodnaviridae	0.03%	0.29%	0.28%	0.54%	0.27%	0.31%	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%
Cellvibrionales	0.20%	0.12%	0.09%	0.10%	0.09%	0.09%	0.45%	0.60%	0.01%	0.00%	0.00%	0.00%
unclassified DNA viruses	0.55%	0.22%	0.20%	0.23%	0.20%	0.20%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%
Level: Species (top 50)												
Dasheen mosaic virus	3.51%	9.05%	9.70%	6.59%	6.95%	9.39%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gentian mosaic virus	9.71%	5.90%	7.45%	6.37%	7.47%	7.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Laceyella sp. FBKL4.010	0.36%	6.67%	4.23%	6.20%	4.34%	4.77%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Variola virus	17.42%	2.48%	2.87%	3.20%	2.88%	3.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Clostridium botulinum	0.13%	0.73%	4.43%	0.61%	16.36%	0.53%	0.02%	0.02%	0.00%	57.64%	0.00%	0.00%
Rhizoctonia solani dsRNA virus 3	13.81%	2.70%	2.56%	5.11%	1.93%	2.98%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Thermoanaerobacter wiegelii	1.49%	4.44%	3.52%	3.42%	2.67%	4.24%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%
Myrmica scabrinodis virus 1	2.89%	1.99%	2.43%	2.34%	2.87%	2.52%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Whataroa virus	1.13%	1.98%	2.67%	2.35%	2.52%	2.77%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Heterobasidion partitivirus 7	2.89%	2.40%	2.28%	2.02%	1.63%	2.80%	0.00%	0.00%	0.00%	0.01%	0.02%	0.00%
Salivirus FHB	9.04%	1.18%	1.62%	1.52%	1.46%	1.59%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Salmonella enterica	0.15%	2.96%	2.12%	2.31%	1.48%	2.51%	0.02%	0.02%	0.00%	0.00%	0.00%	0.00%
Pseudomonas sp. NC02	3.53%	1.58%	1.47%	1.56%	1.65%	1.61%	6.41%	10.68%	0.02%	16.88%	41.77%	34.37%
Pseudomonas fluorescens	3.17%	1.69%	1.50%	1.47%	1.63%	1.59%	6.39%	10.30%	0.03%	16.80%	39.43%	33.93%
Thermoanaerobacterium thermosaccharolyticum	0.05%	2.60%	1.65%	1.68%	1.26%	1.84%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Borrelia duttonii	0.23%	2.34%	1.58%	1.85%	1.24%	1.82%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Bacillus licheniformis	0.05%	1.92%	1.40%	1.47%	1.12%	1.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Hepacivirus C	1.70%	1.36%	1.29%	1.49%	1.05%	1.26%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%
Caldicellulosiruptor bescii	0.05%	1.61%	1.43%	1.42%	1.18%	1.40%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Lasius neglectus virus 1	1.88%	0.87%	1.16%	1.70%	0.91%	1.22%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Acidovorax sp. KKS102	4.82%	0.75%	0.80%	0.89%	0.83%	0.89%	9.17%	8.05%	0.03%	0.03%	0.00%	0.00%
Bacillus megaterium	0.49%	1.02%	1.28%	1.47%	1.36%	1.14%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%
Escherichia coli	0.39%	1.69%	1.14%	1.15%	0.95%	1.33%	0.05%	0.05%	0.00%	0.03%	0.09%	0.21%
Aspergillus foetidus dsRNA mycovirus	0.82%	1.23%	1.21%	1.04%	1.03%	1.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Jingmen tick virus	1.49%	1.18%	1.17%	0.90%	0.87%	1.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Rhodothermus marinus	0.05%	1.62%	1.18%	1.11%	0.86%	1.25%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Piscine myocarditis-like virus	0.05%	1.15%	1.40%	0.97%	0.97%	1.43%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%
Diabrotica virgifera virgifera virus 2	2.01%	0.73%	0.88%	0.95%	0.71%	0.89%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Methylotenera versatilis	3.30%	0.61%	0.60%	0.70%	0.69%	0.68%	5.47%	4.97%	0.03%	0.00%	0.09%	0.13%
Caldanaerobacter subterraneus	0.21%	1.16%	0.87%	1.04%	0.68%	0.98%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Marine RNA virus JP-B	0.05%	1.27%	0.94%	0.95%	0.64%	1.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Caldicellulosiruptor obsidiansis	0.13%	1.07%	0.86%	0.89%	0.63%	0.89%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Mycoplasma hyopneumoniae	0.72%	0.91%	0.76%	0.89%	0.61%	0.82%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Bacillus phage Stitch	0.05%	1.12%	0.83%	0.77%	0.58%	0.98%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
White bream virus	0.70%	0.75%	0.84%	0.86%	0.73%	0.75%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Cytophaga hutchinsonii	2.42%	0.44%	0.50%	0.57%	0.54%	0.54%	4.13%	3.14%	0.02%	0.00%	0.00%	0.00%
Thermococcus chitonophagus	0.08%	0.89%	0.68%	0.81%	0.61%	0.73%	0.00%	0.00%	0.13%	0.00%	0.00%	0.00%

Candidatus Portiera aleyrodidarum	2.34%	0.39%	0.44%	0.51%	0.49%	0.49%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Cadicivirus A	0.98%	0.62%	0.60%	0.63%	0.57%	0.61%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Lactobacillus sakei	0.03%	0.93%	0.67%	0.65%	0.52%	0.75%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Dill cryptic virus 1	1.00%	0.62%	0.54%	0.45%	0.34%	0.68%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%
Campylobacter hominis	0.10%	0.62%	0.58%	0.67%	0.53%	0.62%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Canine kobuvirus	0.21%	0.68%	0.58%	0.63%	0.44%	0.62%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Mamastrovirus 3	1.16%	0.46%	0.53%	0.53%	0.42%	0.49%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Broad bean mottle virus	0.00%	0.75%	0.58%	0.57%	0.44%	0.58%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Southern elephant seal virus	0.18%	0.48%	0.53%	0.77%	0.42%	0.61%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Mink coronavirus 1	0.10%	0.62%	0.52%	0.68%	0.46%	0.51%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Ndumu virus	0.72%	0.46%	0.56%	0.47%	0.41%	0.53%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Nostocales cyanobacterium HT-58-2	0.03%	0.68%	0.54%	0.57%	0.35%	0.60%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Hibiscus latent Singapore virus	0.03%	0.62%	0.49%	0.46%	0.37%	0.55%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table A.4: HaploGrep2 [164] results of all complete modern mitochondrial genomes used for the evaluation of mitoBench. The table shows the accession ID of the sample, and the comparison of the published and the mitoBench result.

Accession ID	HG mitoBench	HG published
JN214449	L2a1a2	L2a1a2
JN214450	3d1b1a	3d1b1a
JN214459	3d1b1a	3d1b1a
JN214460	1b1a16	1b1a16
KF146264	W1c	W1c
KF146293	W	W
MN687107	H	H
MN687108	U1a1a	U1a1a
MN687109	X2d1	X2d1
MN687110	U5a1	U5a1
MN687111	U5b2b4a	U5b2b4a
MN687112	H	H
MN687113	K2a9	K2a9
MN687114	H94	H94
MN687115	K1a19	K1a19
MN687116	X2b+226	X2b+226
MN687117	H1h1	H1h1
MN687118	H1	H1
MN687119	N1b1a	N1b1a
MN687120	H2b	H2b
MN687121	R0a2n	R0a2n
MN687122	H5m	H5m
MN687123	H15b1	H15b1
MN687124	H10a1	H10a1
MN687125	V1a1	V1a1
MN687126	H1+16189	H1+16189
MN687127	X2i+@225	X2i+@225
MN687128	T2b	T2b
MN687129	H61	H61
MN687130	H3	H3
MN687131	H5a1	H5a1
MN687132	H1a4	H1a4
MN687133	U3b2a1	U3b2a1
MN687134	H1e1b1	H1e1b1
MN687135	K1b1a1	K1b1a1
MN687136	H536	H5'36
MN687137	U5b1+16189+@16192	U5b1+16189+@16192
MN687138	U5b1+16189+@16192	U5b1+16189+@16192
MN687139	J2a1a1a	J2a1a1
MN687140	H81	H81
MN687141	J1b2	J1b2
MN687142	T2b4a1	T2b4a1
MN687143	K1b1+(16093)	K1b1+ (16093)
MN687144	HV4	HV4
MN687145	J1c3d	J1c3d
MN687146	H13a2b3	H13a2b3
MN687147	T2	T2
MN687148	H1	H1
MN687149	R0a1+152	R0a1+152
MN687150	J1c3d	J1c3d
MN687151	H1bz	H1bz
MN687152	H7b1	H7b1
MN687153	J1c3d	J1c3d
MN687154	U5a1f1	U5a1f1
MN687155	U5a1f1	U5a1f1
MN687156	J1c3d	J1c3d
MN687157	U5a1a1	U5a1a1
MN687158	H1e4	H1e4
MN687159	H+16291	H+16291
MN687160	T1b4	T1b4
MN687161	X2i+@225	X2i+@225
MN687162	U3b2b	U3b2b
MN687163	W6	W6
MN687164	K1b1	K1b1
MN687165	J2a1a1a	J2a1a1
MN687166	T2e	T2e
MN687167	J1c3d	J1c3d
MN687168	J1c2	J1c2
MN687169	U4a1b1	U4a1b1
MN687170	H1ak	H1ak
MN687171	J1c3+189	J1c3+189
MN687175	T2+16189	T2+16189
MN687176	H1	H1
MN687177	H1q3	H1q3
MN687178	H	H
MN687188	H1a	H1a
MN687190	H4a1c	H4a1c
MN687191	H1c	H1c
MN687192	H13a1a2	H13a1a2
MN687193	H3	H3
MN687194	T2+16189	T2+16189

MN687195	X2	X2
MN687196	H44b	H44b
MN687197	H+16129	H+16129
MN687198	H2a2a	H2a2a
MN687199	U3b	U3b
MN687200	U3b	U3b
MN687201	U7	U7
MN687202	T2e	T2e
MN687203	HV	HV
MN687204	U5b1f1a	U5b1f1a
MN687205	X2d1	X2d1
MN687206	H9a	H9a
MN687207	U1a1a	U1a1a
MN687208	H66	H66
MN687209	H2a3	H2a3
MN687210	H	H
MN687211	N1a3a	N1a3a
MN687212	N1b1a	N1b1a
MN687213	U5a1g	U5a1g
MN687214	H+195	H+195
MN687215	H	H
MN687216	H1e1	H1e1
MN687217	H13	H13
MN687218	J1c3	J1c3
MN687219	X2	X2
MN687220	X2	X2
MN687221	HV+16311	HV+16311
MN687222	T2c1a	T2c1a
MN687223	K1a	K1a
MN687224	J1c3	J1c3
MN687225	W6	W6
MN687226	U3b	U3b2
MN687227	W6	W6
MN687228	H1c	H1c
MN687229	H1b1g	H1b1g
MN687231	H3ap	H3ap
MN687232	T2a1b	T2a1b
MN687233	H1+16311	H1+16311
MN687234	J1c4	J1c4
MN687235	T1b	T1b
MN687236	U2e123	U2e1'2'3
MN687237	W+194	W+194
MN687238	U4b1a2a	U4b1a2a
MN687239	H1+16189	H1+16189
MN687240	U5b2a2a1	U5b2a2a1
MN687241	U5a1	U5a1
MN687242	H1au	H1au
MN687243	J1c	J1c
MN687244	J2b1c	J2b1c
MN687245	U4b1a1a1	U4b1a1a1
MN687246	V1a	V1a
MN687247	X2	X2
MN687248	H+152	H+152
MN687249	H	H
MN687250	U7b	U7b
MN687251	J2a1a1a	J2a1a1
MN687252	U5a1b1	U5a1b1
MN687253	J1c2q	J1c2q
MN687254	J1c15	J1c15
MN687255	J1c3d	J1c3d
MN687256	J1c5	J1c5
MN687257	H7a1	H7a1
MN687258	H61	H61
MN687260	H5a2	H5a2
MN687262	K1a4f	K1a4f
MN687263	W4	W4
MN687264	K1a	K1a
MN687265	T2c1a	T2c1a
MN687266	U4c1	U4c1
MN687267	H	H
MN687268	T2b	T2b
MN687269	H12a	H12a
MN687270	K1a4a	K1a4a
MN687271	W3a1	W3a1
MN687272	K1a4j1	K1a4j1
MN687273	J2b1a1	J2b1a1
MN687274	U8b1b1	U8b1b1
MN687275	H13	H13
MN687276	H1aj	H1aj
MN687277	J1c3	J1c3
MN687278	D4j12	D4j12
MN687279	H18	H18
MN687280	H4a1a	H4a1a
MN687281	J1d1b1	J1d1b1
MN687282	H1b1	H1b1
MN687284	H5b	H5b
MN687285	H11a2	H11a2
MN687286	W6	W6

MN687287	H4a1a	H4a1a
MN687288	K1b1a1	K1b1a1
MN687289	H7a1a	H7a1a
MN687290	J2a2a2	J2a2a2
MN687291	HV0+195	HV0+195
MN687292	K1a2a	K1a2a
MN687293	U5b1e1	U5b1e1
MN687294	H4a1a1a	H4a1a1a
MN687295	H1ak	H1ak
MN687296	2c	2c
MN687297	N1b1a2	N1b1a2

APPENDIX B

Publications

B.1 Articles

2021

- **Neukamm J**, Peltzer A, Achilli A, Balanovsky O, Andreson R, Bajić V, Balanovsky O, Barbieri B, Bodner M, Gandini F, Hübner A, Macholdt E, Metspalu M, Olivieri A, Pala M, Parson W, Prüfer K, Richards MB, Saag L, Schönherr S, Stoneking M, Torroni A, van Oven M, Weißensteiner H, Zaporozhchenko V, Nieselt K & Haak W. *mitoBench: Novel interactive methods and repository for population genetics of human mitochondrial DNA* **in preparation**.

- Pfrengle S, **Neukamm J**[1], Guellil M, Keller M, Molak M, Avanzi C, Kushniarevich A,Montes N, Neumann GU, Reiter E, Tukhbatova RI, Berezina NY, Buzhilova AP, Korobov DS, Hamre SS, Matos VMJ, Ferreira MT, González-Garrido L, Wasterlain SN, Lopes C, Santos AL, Antunes-Ferreira N, Duarte V, Silva AM, Melo L, Sarkic N, Saag L, Tambets K, Busso P, Cole ST, Avlasovich A, Roberts CA, Sheridan A, Cessford C, Robb J, Krause J, Scheib CL, Inskip SA & Schuenemann VJ. *Mycobacterium leprae diversity and population dynamics in medieval Europe from novel ancient genomes.* **BMC Biology** (Accepted).

- Yates JA, Lamnidis TC, Borry M, Valtueña AA, Fagnerås Z, Clayton S, Garcia MU, **Neukamm J** & Peltzer A. *Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager.* **PeerJ**, March 2021, 16;9:e10947.

[1] The first two authors contributed equally to this study.

2020

- **Neukamm J**, Peltzer A & Nieselt K. *DamageProfiler: Fast damage pattern calculation for ancient DNA* **Bioinformatics**, April 2020, btab190.
- Morozova I, Kasianov A, Bruskin S, **Neukamm J**, Molak M, Chekalin E, Batieva E, Pudło A, Rühli FJ & Schuenemann VS. *New ancient Eastern European Yersinia pestis genomes illuminate the dispersal of plague in Europe.* **Philosophical Transaction of the Royal Society B**, November 2020, 375(1812):20190569.
- Ferrari G & **Neukamm J**[2], Baalsruda HT, Breidenstein AM, Ravinet M, Phillips C, Rühli FJ, Bouwman A & Schuenemann VJ. *Variola virus genome isolated from an 18th century museum specimen supports the recent origin of smallpox.* **Philosophical Transaction of the Royal Society B**, November 2020, 375(1812):20190572.
- Majander K, Pfrengle S, Kocher A, **Neukamm J**, du Plessis L, Pla-Díaz M, Arora N, Akgül G, Salo K, Schats R, Inskip S, Oinonen M, Valk H, Nalve M, Kriiska A, Onkamo P, González-Candelas F, Kühnert D, Krause J & Schuenemann VJ. *Ancient bacterial genomes reveal a high diversity of Treponema pallidum strains in early modern Europe.* **Current Biology**, September 2020.
- **Neukamm J**, Pfrengle S, Molak M, Seitz A, Francken M, Eppenberger P, Avanzi C, Reiter E, Urban C, Welte B, Stockhammer PW, Teßmann B, Herbig A, Harvati K, Nieselt K, Krause J & Schuenemann VJ. *2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals.* **BMC Biology**, August 2020, 18(1):1-8.

2019

- Morozova I, Öhrström LM, Eppenberger P, Bode-Lesniewska B, Gascho D, Haas C, Akgül G, **Neukamm J**, Röthlin KA, Imhof A, Shved N, Papageorgopoulou C & Rühli FJ. *Ongoing tissue changes in an experimentally mummified human leg.* **The Anatomical Record**, December 2019.
- Gretzinger J, Molak M, Reiter E, Pfrengle S, Urban C, **Neukamm J**, Blant M, Conard NJ, Cupillard C, Dimitrijević V, Drucker DG, Hofman-Kamińska E, Kowalczyk R, Krajcarz MT, Krajcarz M, Münzel SC, Peresani M, Romandini M, Rufí I, Soler J, Terlato G, Krause J, Bocherens H & Schuenemann VJ. *Large-scale mitogenomic analysis of the phylogeography of the Late Pleistocene cave bear.* **Scientific Reports**, August 2019, 1-11.

[2] The first two authors contributed equally to this study.

2018

- Ferrari G, Lischer HE, **Neukamm J**, Rayo E, Borel N, Pospischil A, Rühli F, Bouwman AS & Campana MG. *Assessing Metagenomic Signals Recovered from Lyuba, a 42,000-Year-Old Permafrost-Preserved Woolly Mammoth Calf.* **Genes**, September 2018, 9:436.

B.2 Posters & presentations

- **Neukamm J**, Pfrengle S, Molak M, Seitz A, Francken M, Eppenberger P, Avanzi C, Reiter E, Urban C, Welte B, Tessmann B, Herbig A, Harvati K, Nieselt K, Krause J and Schuenemann VJ.
  2,000-year-old pathogen genomes reconstructed from mummies provide insights into the state of health of ancient Egyptians. **Talk** at the Congress of the European Society for Evolutionary Biology (ESEB) in Turku/Finland, August 23, 2019.

- **Neukamm J**, Pfrengle S, Molak M, Seitz A, Francken M, Eppenberger P, Avanzi C, Reiter E, Urban C, Welte B, Tessmann B, Herbig A, Harvati K, Nieselt K, Krause J and Schuenemann VJ.
  2,000-year-old pathogen genomes reconstructed from mummies provide insights into the state of health of ancient Egyptians. **Poster** at the 5th annual Meeting of the International Society for Evolution, Medicine, and Public Health (ISEMPH) in Zurich/Switzerland, August 15, 2019.

- **Neukamm J**, Peltzer A, Achilli A, Balanovsky O, Barbieri C, Bodner M, Gandini F, Macholdt E, Olovieri A, Pala M, Parson W, Richards MB, Schönherr S, Stoneking M, Torroni A, van Oven M, Weissensteiner H, Zaporozhchenko V, Nieselt K and Haak W. *MitoBench & MitoDB – novel interactive methods for population genetics of human mitochondrial DNA* **Talk** at the 8th International Symposium on Biomolecular Archaeology (ISBA) 2018 in Jena/Germany, September 21, 2018.

- **Neukamm J**, Pfrengle S, Molak M, Seitz A, Francken M, Welte B, Harvati K, Nieselt K, Krause J and Schuenemann VJ.
  A 2,200 year old Mycobacterium leprae genome from an Egyptian mummy. **Talk** at the SMBE 2018 Conference in Yokohama/Japan, June 09, 2018.

- **Neukamm J**, Pfrengle S, Molak M, Seitz A, Francken M, Welte B, Harvati K, Nieselt K, Krause J and Schuenemann VJ.
  Metagenomic analysis of Egyptian mummies. **Talk** at the 9th World Congress on Mummy Studies (WCOMS) in Tenerife/Spain, May 24, 2018.

- **Neukamm J**, Schuenemann VJ, Krause J and Nieselt K. *DamageProfiler: Calculation of damage patterns in next generation sequencing data from ancient DNA.* **Poster** at the meeting of students in evolution and ecology in Tübingen/Germany, November 04, 2016.
- **Neukamm J**, Schuenemann VJ, Krause J and Nieselt K. *DamageProfiler: Calculation of damage patterns in next generation sequencing data from ancient DNA.* **Poster** at the 7th International Symposium on Biomolecular Archaeology (ISBA) 2016 in Oxford/England, September 15, 2016.

APPENDIX C

Academic teaching experience

C.1 Supervised lectures and course

WS 2020/21

- Lecture: Evolutionary Medicine: Ancient pathogens and pathologies (Guest lecture) for Bachelor students

- Practical Course: *Evolutionary Medicine - Advanced Topic* for Bachelor students

SS 2019

- Lecture: *Evolutionary Medicine - Advanced Topics* (Guest lecture) for Bachelor students

- Practical Course: *Evolutionary Medicine - Advanced Topics* for Bachelor students

SS 2018

- Lecture: *Evolutionary Medicine - Advanced Topics* (Guest lecture) for Bachelor students

- Practical Course: *Evolutionary Medicine - Advanced Topics* for Bachelor students

WS 2017/18

- Seminar: *Biologie für Archäologen* for Bachelor students

SS 2017

- Practical course: *Analysis of NGS data from ancient DNA* for Master students

WS 2016/17

- Seminar: *Biologie für Archäologen* for Bachelor students

SS 2016

- Practical course: *Analysis of NGS data from ancient DNA* for Master students

C.2 Co-Supervised theses

2020

- Donike Sejdiu. *Investigation of evolutionary history of Mycobacterium tuberculosis during the 19th century* December 2020, Masterthesis

2017

- Susanne Jodoin. *Genetic analysis of the medieval population of Schleswig in northern Germany*. October 2017, Masterthesis

2016

- Joscha Gretzinger. *Der Höhlenbär aus der Prepoštská-Höhle, Slowakei. Zur phylogenetischen Stellung und Klassifikation innerhalb der Gattung Ursus anhand alter mitochondrialer DNS*. October 2016, Bachelorthesis