
Computational methods for ancient genome reconstruction

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. Alexander Peltzer
aus Ulm/Donau
Tübingen
2018

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

Tag der mündlichen Qualifikation: 18. Mai 2018

Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatterin:	Apl. Prof. Dr. Kay Nieselt
2. Berichterstatter:	Prof. Dr. Johannes Krause

*Dedicated to my parents
Doris † & Gerhard*

*"If we admit a first cause,
the mind still craves to know
whence it came and how it arose."*

- Charles Darwin, *"Life and Letters of Charles Darwin, 1873"*

Zusammenfassung

Anwendungen von Next Generation Sequencing (NGS)-Technologien sind zum De-facto-Standard in der systematischen Analyse der genetischen Komposition von Organismen geworden. Dies gilt nicht nur für die moderne DNA-Analyse, sondern auch für alte DNA (aDNA), bei der die NGS-Methoden die PCR-basierten Ansätze in den letzten Jahren fast vollständig ersetzen konnten.

Die Erforschung der DNA Variation in historischen Populationen bietet vielversprechende Möglichkeiten, fehlende Verbindungen in der Geschichte der Menschheit aufzudecken, die ansonsten nur schwer feststellbar wären. Die Methoden, die die Paläogenetik für die Erforschung der antiken Populationen bereitstellt, konnten unser Verständnis der Menschheitsgeschichte bereits in hohem Maße verändern. Diese Möglichkeiten sind zwar positiv, es existieren jedoch noch immer einige Probleme bei der Analyse von NGS-Daten von antiken Proben, die explizite Herausforderungen an die Bioinformatik stellen. Da aDNA Forschungsprojekte beispielsweise nur geringe Mengen an DNA generieren, müssen bioinformatische Methoden mit einem geringen DNA Gehalt umgehen können. Darüber hinaus stellen die inhärenten Herausforderungen der aDNA, wie z.B. DNA-Fehlpaarungen sowie die Kontamination durch moderne Quellen, weitere Herausforderungen für die erfolgreiche Analyse der aDNA dar. Der besondere Forschungsschwerpunkt der Bioinformatik in der aDNA Analyse liegt daher auf der Rekonstruktion alter Genome und der anschließenden Datenanalyse rekonstruierter Genomdaten. Ursprüngliche Analysemethoden waren bisher in ihrer Anwendbarkeit auf einige wenige Forschungsfragen beschränkt und erforderten daher die Anpassung der jeweiligen Werkzeuge, auch bei leicht unterschiedlichen Fragestellungen, im Kontext der aDNA Analyse.

Das Hauptthema dieser Dissertation konzentriert sich auf die Entwicklung von EAGER, einem Framework zur Analyse von aDNA Daten mit einer Vielzahl von Anwendungsfällen und Verbesserungen gegenüber bisher veröffentlichten Methoden. EAGER verfügt über mehrere neue Analysemethoden, die darauf abzielen, so viel aDNA wie möglich aus aDNA Sequenzierungsprojekten zu rekonstruieren. Darüber hinaus ist die Pipeline eine integrierte Lösung zur Analyse von aDNA Daten, bei der mehrere Analysemethoden kombiniert zur Rekonstruktion alter Genome eingesetzt werden, um eine möglichst einfache Nutzung der Methode zu erreichen. Die Anwendbarkeit von EAGER wurde in verschiedenen aDNA Analyseprojekten demonstriert und wird im Rahmen dieser Arbeit in Anwendungen zur Rekonstruktion des Genoms von George Bähr, dem Architekten der Dresdner Frauenkirche, und insgesamt 90 altägyptischen Individuen aus dem nordägyptischen Abusir El-Meleq veranschaulicht.

Neben der eigentlichen Analyse spielt in der heutigen Forschung die Handhabung genomischer Daten und Metadaten eine entscheidende Rolle. Während Sequen-

zierungsprojekte in den letzten Jahren florierten und immer mehr Daten produzierten, hatten effiziente bioinformatische Methoden, die den Anwendern bei der Organisation, Speicherung und Analyse ihrer jeweiligen Daten helfen sollten, Schwierigkeiten, mit der steigenden Datenmenge Schritt zu halten. Die gegenwärtige Situation in der Populationsgenetik sieht daher ein Defizit an bioinformatischen Anwendungen, die es Forschern ermöglichen, ihre Daten in größeren Kohorten zu organisieren und zu analysieren. Der zweite Teil dieser Arbeit konzentriert sich daher auf die konzeptionelle Einführung von MitoBench und MitoDB. Die Idee von MitoBench basiert auf dem Konzept einer modernen Analyseanwendung, die von Forschern genutzt werden kann, um ihre mitochondrialen Populationsgenetikdaten mit Metadaten aus einer Vielzahl von Ressourcen zu integrieren. Darüber hinaus bietet die Idee von MitoDB eine zentral zugängliche Datenbank für mitochondriale DNA mit Metadaten, die als Datenquelle für zukünftige Analyseprojekte im Kontext der mitochondrialen Populationsgenetik dienen soll.

Da die Sequenzierungskosten stetig sinken und damit die Stichprobengrößen für Forschungsprojekte mit derselben Geschwindigkeit ansteigen, werden neue Methoden und Frameworks für die Analyse von aDNA benötigt. Insgesamt hat das in dieser Dissertation erläuterte Framework EAGER sowie die konzeptionellen Ideen von MitoBench und MitoDB dazu beigetragen, die Menschheitsgeschichte in mehreren Projekten besser zu begreifen, und wird den Forschern hoffentlich auch weiterhin helfen, die Veränderungen aufzuklären, die vergangene Populationen erfahren haben.

Abstract

Applications of next-generation sequencing (NGS) technologies have become the de facto standard in the systematic analysis of the genetic composition of organisms. Aforementioned is not just valid for modern DNA analysis, but also for ancient DNA (aDNA) where NGS methods have almost entirely replaced PCR-based approaches.

Studying DNA variation in ancient humans provides promising opportunities to unravel missing links in human history that are otherwise hard to detect. Today, the methods that paleogenetics can provide to study ancient populations can change the understanding of human history to a large extent. While this is encouraging, there are still several issues in analyzing NGS data from ancient specimens, posing challenges to bioinformatics. As aDNA research projects typically produce low levels of DNA extract, bioinformatic methods have to cope with low DNA content. Furthermore, the inherent challenges of aDNA, such as DNA misincorporation patterns as well as human DNA contamination not only from modern sources, pose further challenges for the successful analysis of aDNA. Therefore, the primary research focus of bioinformatics in aDNA analysis lies on the reconstruction of ancient genomes and the subsequent data analysis of reconstructed genomes. Pipeline scripts that were used before, were limited in their applicability to few research questions and therefore required the adaption of the respective tools even for slightly different research scopes.

The main topic of this dissertation concentrates on the development of EAGER, a framework for the analysis of aDNA data with a variety of use cases and improvements in contrast to previously published methods. EAGER features several newly contributed analysis methods, aiming at recovering as much aDNA as possible from sequencing experiments. Additionally, the pipeline provides an integrated solution to analyze aDNA data in an advanced way, running several state of the art analysis methods to reconstruct ancient genomes. The applicability of EAGER has been demonstrated in various aDNA analysis projects and is further illustrated within this thesis, having been applied to the reconstruction of the genome of George Bähr, the architect of Dresden Frauenkirche, and a total of 90 ancient Egyptian individuals from Abusir El-Meleq in Northern Egypt.

Apart from the general processing as handled in EAGER, the handling of genomic data and metadata is of crucial importance in today's research. While sequencing projects prospered in the last couple of years, thereby producing more and more data, efficient bioinformatic tools to aid users in organizing, storing and analyzing their respective data have had difficulties in keeping up with the increasing amount of data produced. Therefore, the current situation in population genetics suffers struggles with a lack of bioinformatics methods capable of accommodating researchers with functionality to organize and analyze their data in larger sample

cohorts.

The second part of this thesis therefore concentrates on the conceptual introduction of MitoBench and MitoDB. The idea of MitoBench centers around the concept of an advanced analysis application that can be used by researchers to integrate their mitochondrial population genetics data with metadata from a variety of resources. In addition, the idea of MitoDB provides a centrally accessible database of mitochondrial DNA with metadata to serve as a data resource for future analysis projects within a mitochondrial population genetics context.

As sequencing costs are in steady decline and thus sample sizes for research projects are growing at equal speed, novel methods and frameworks for the analysis of aDNA are required. Overall, both the EAGER framework explained in this dissertation and the conceptual ideas of MitoBench and MitoDB have contributed in understanding human history better. They will hopefully continue to aid researchers in elucidating the changes past populations have experienced.

Acknowledgements

First and foremost I wanted to express my deepest gratitude to my advisors Prof. Dr. Kay Nieselt, group leader of the Integrative Transcriptomics group at the University of Tübingen and Prof. Dr. Johannes Krause, director of the Max-Planck Institute for the Science of Human History in Jena. Both Kay and Johannes were very supportive of my work throughout the whole time of my thesis, enabled me to work on numerous exciting projects and provided me with the possibility to also include own ideas described in this work. Furthermore, I always had the feeling to get professional and personal advice even in difficult times, for which I am grateful.

Much of my work here was inspired and also shaped by in-depth discussions with other students, providing ideas, comments and also contributions to several of my projects. First of all, I wanted to thank Dr. Günter Jäger and Dr. Alexander Herbig for their support throughout the whole scope of this thesis and providing creative ideas for various projects. Furthermore, I wanted to thank André Hennig, Alexander Seitz, Sven Fillinger, Alexander Immel, Stephen Clayton, Judith Neukamm, Marie Gauder and Julia Söllner for copious inspiring discussions, proof reading and several very productive collaborations in research projects while working on my Ph.D. projects.

Much of my work would not have been possible without feedback from my colleagues applying the methods and software I developed. Therefore I wanted to express my deepest appreciation to present and past colleagues at the Integrative Transcriptomics (IT), Palaeo - & Archaeogenetics (PAL) groups at the University of Tübingen and as well at the Max-Planck Institute for the Science of Human History in Jena (SHH): Andreas Friedrich, Aydın Can Polatkan, Kerttu Majander, Dr. Alissa Mitnik, Aida Andrades Valtueña, Dr. Kirsten Bos, Michal Feldman, James A. Fellows Yates, Dr. Cosimo Posth and Maria Spyrou. I also wanted to thank the whole Medical and Population Genetics group at Harvard University for providing me with the opportunity of an internship in their group: Dr. Qiaomei Fu, Swapan Mallick, Nick Patterson Ph.D., Pontus Skoglund Ph.D. and Prof. Dr. David Reich.

I am thankful for the help of Sabine Gebert-Rudolph (IT), Birgit Grieg (IT), Sonia Varandas (PAL), Sylvia Arnold-El Fehri (SHH), Johanna Allner (SHH) and Beate Kerpen (SHH) with administrative tasks, including paper submissions, event and trip planning as well as other bureaucratic challenges during the time of my Ph.D. For the work on EAGER, I am thankful to all collaborators and users that gave comments and provided feedback on the pipeline: Dr. Stephan Schiffels, Stephen Clayton (SHH), Christian Kniep (Docker Inc.), Prof. Dr. Ben Krause-Kyora and Dr. Marcel Nutsua (University of Kiel). I also wanted to personally thank Dr. Gabriel Renaud (Center for GeoGenetics, University of Copenhagen/Denmark) for

lots of fruitful discussions on contamination estimation.

I am very grateful to Assistant Prof. Dr. Dr. Verena Schünemann, Ella Reiter, Anja Furtwängler, Christian Urban, Dr. Beatrix Welte (University of Tübingen), Willem Paul van Pelt Ph.D. (University of Cambridge), Chuanchao Wang Ph.D., Dr. Wolfgang Haak, Dr. Stephan Schiffels (SHH) and Dr. Martyna Molak (Uniwersytet Warszawski, Warsaw/Poland) for the interesting collaboration and joint work on the ancient Egyptian mummy project.

For the interesting work on *Treponema pallidum* I wanted to thank Dr. Natasha Arora and Dr. Michal Strouhal.

I am delighted that I was moreover involved in the reconstruction of the genome of George Bähr, the architect of the Dresdner Frauenkirche and I am thankful for my collaborators to work in this project: Chuanchao Wang Ph.D., Tristan Begg, Kajo Kusen and Dr. Siegfried Gerlach.

Additionally, I wanted to thank all collaborators in the MitoBench and MitoDB project for the contributions, interesting discussions and help in providing a general framework for mitochondrial analysis work in the future, most notably: Judith Neukamm (PAL), Dr. Wolfgang Haak (SHH), Oleg Balanovsky Ph.D., Valery Zaporozhenko (Vavilov Institute of General Genetics, Moscow, Russia), Dr. Martin Bodner, Prof. Dr. Walther Parson (Gerichtsmedizin Universität Innsbruck, Austria), Prof. Dr. Antonio Torroni, Dr. Alessandro Achilli (University of Pavia, Italy), Prof. Dr. Martin B Richards, Dr. Maria Pala (University of Huddersfield, United Kingdom), Mannis van Oven (Erasmus MC, Rotterdam, Netherlands), Dr. Mark Stoneking, Enrico Macholt (MPI-EVA, Leipzig, Germany), Dr. Hansi Weissensteiner, Sebastian Schoenherr (Medical University of Innsbruck, Austria), Anna Olivieri Ph.D. (Istituto Superiore di Sanità, Rome, Italy) and Benjamin Dietrich (Database systems group, University of Tübingen) for his help with our database questions.

Much of my work would not have been possible with the help of Bachelor or Master students, namely: Christopher Jürges, Maximilian Hanussek, Judith Neukamm, Max-Emil Schön and Veronika Böttcher.

Most notably, I am grateful to my parents, Doris and Gerhard Peltzer, for their love and continued support throughout the whole course of my studies, which is why I dedicate this thesis to both of them. Lastly, I want to thank my wife Anika for her support especially during the last years of my Ph.D.

Contents

List of Figures	xi
List of Tables	xiii
List of Abbreviations	xv
1 Introduction	1
2 Background	5
2.1 Genetic Variation	5
2.2 DNA sequencing: Past & Future	7
2.2.1 Illumina sequencing & NGS analysis	8
2.2.2 NGS data applications & analysis	10
2.3 Specific characteristics of ancient DNA	12
2.3.1 DNA Damage: What is left	12
2.3.2 Environmental DNA: The 1% curse	15
2.3.3 Contamination: Background noise - or real signal?	16
2.3.4 Recent improvements	17
2.4 Bioinformatics challenges in aDNA	18
2.5 Bioinformatics challenges in genomics	19

Contents

3	EAGER: Efficient ancient genome reconstruction	21
3.1	Introduction	21
3.2	Development	25
3.2.1	Command line interface	25
3.2.2	Integrated tools and methods	26
3.2.3	Graphical user interface	29
3.2.4	Contributed tools and methods	31
3.2.5	Testing & deployment	41
3.3	Results	42
3.4	Feature comparison of EAGER and PALEOMIX	52
3.5	Availability & requirements	52
3.6	Application of EAGER in a forensic case	56
3.6.1	Introduction	56
3.6.2	Results & discussion	57
3.7	Discussion	61
4	mitoBench & mitoDB:	
	Modern tools for mitochondrial genome analysis	65
4.1	Introduction & motivation	65
4.2	Conceptual application design	69
4.2.1	MitoBench	69
4.2.2	MitoDB	72
4.3	Conclusion & outlook	76
5	Investigating north African population structure	79
5.1	Introduction	79
5.2	Challenges from a bioinformatics perspective	80
5.3	Methods	81
5.3.1	Sequence generation	81
5.3.2	Processing and authentication with EAGER	82
5.3.3	Sequence based analysis with mitoBench	82
5.3.4	Frequency based analysis with mitoBench	83
5.4	Results	83
5.5	Discussion	86

6 Discussion	91
6.1 EAGER, an efficient ancient genome reconstruction pipeline	93
6.2 MitoBench & mitoDB: Modern methods for mitochondrial genome analysis	95
6.3 EAGER and mitoBench revealing ancient Egyptian population history	96
6.4 Outlook	97
6.5 Conclusion	99
Bibliography	101
A Supplementary Information	121
B Publications	135
B.1 Articles	135
B.2 Posters, presentations & workshops	136
C Academic teaching experience	139
C.1 Supervised lectures and course	139
C.2 Supervised Bachelor/Master theses	140

Contents

List of Figures

1.1	The “hype cycle” of ancient DNA	3
2.1	IGV variant visualization	6
2.2	Basic Illumina sequencing approach	9
2.3	Array capture principle	11
2.4	aDNA damage and read length distribution	13
2.5	aDNA misincorporation chemistry	14
2.6	Visual example for DNA contamination types	17
3.1	PCR vs. NGS based size selection of DNA fragments	22
3.2	EAGER CLI Components Organigram	26
3.3	Main EAGER GUI	31
3.4	Usability features of EAGER GUI	32
3.5	CircularMapper concept	33
3.6	DeDup duplicate removal strategy outline	35
3.7	DeDup technical implementation details	36
3.8	DamageProfiler GUI	37
3.9	Library Complexity Estimation Plot	38
3.10	BWA-Mismatches Shiny App	40
3.11	Jenkins Build Infrastructure	41
3.12	aDNA project scope	44

List of Figures

3.13	DeDup coverage comparison	45
3.14	CircularMapper vs. BWA coverage plot	48
3.15	CircularMapper mean coverage plot	49
3.16	Pipeline execution workflow: Pathogen analysis	49
3.17	Pipeline execution workflow: mtDNA analysis	50
3.18	Pipeline execution workflow: WGS deep shotgun analysis	51
3.19	Front view of Dresden Frauenkirche	57
3.20	DNA Damage plot of George Bähr	58
3.21	Moore's law of aDNA	62
4.1	MitoBench scheme	69
4.2	MitoBench main GUI	71
4.3	Venn Diagram mitoDB design	73
4.4	Database layout of the MitoDB prototype	74
4.5	MitoDB web application prototype	75
4.6	MitoDB central dashboard prototype	77
5.1	Comparison of different tissue types in aDNA of Egyptian mummies	84
5.2	Stacked barchart of haplogroups investigated in ancient Egyptian mitochondrial genomes	85
5.3	MDS plot of F_{ST} distances on HVR-I regions	87
5.4	Geographic mapping of F_{ST} values between ancient Egyptian samples	87
5.5	PCA plot of haplogroup frequencies	88

List of Tables

3.1	DeDup vs. SAMTools performance on borderline cases	45
3.2	DeDup vs. Samtools coverage performance	45
3.3	Pathogen analysis runtimes of EAGER modules	47
3.4	Runtimes of EAGER modules on mtDNA analysis	51
3.5	Deep WGS analysis runtimes of EAGER modules	52
3.6	Basic evaluation criteria overview	53
3.7	Evaluation table of EAGER and Paleomix	53
3.8	EAGER Configuration Table	59
3.9	Phenotyping results of George Bähr	60
5.1	F_{ST} results of genetic distance computation with Arlequin [36]	86
A.1	Results of an aDNA analysis on George Bähr	122
A.2	DeDup Coverage comparison	123
A.3	CircularMapper Coverage Comparison	125
A.4	Criteria table for pipeline evaluation	127
A.5	Analysis results of EAGER on 90 mitochondrial genomes of Ancient Egyptian mummies	128
A.7	Modern reference genome references for F_{ST} analysis	133
A.6	ISO-3 country codes for Ancient Egypt project	134

List of Tables

List of Abbreviations

1000G	1000 Genome (project)
aDNA	Ancient DNA
CLI	Command Line Interface
DNA	Deoxyribonucleic acid
F_{ST}	Fixation index, a measure of population differentiation
GUI	Graphical user interface
HTS	High-throughput sequencing
LIMS	Laboratory Information System
mtDNA	Mitochondrial DNA
NGS	Next generation sequencing
NUMT	Nuclear mitochondrial DNA segment
ORDBMS	Object-relational database management system
PE	Paired end (sequencing)
PCR	Polymerase chain reaction
QC	Quality control
SE	Single end (sequencing)
SNP	Single nucleotide polymorphism
TBI/TAB	Tabbed document interface(s)
TPC	Test of population continuity
VCF	Variant Call Format
UDG	Uracil-deglycosylase (treatment)
UI	User interface
XML	Extensible Markup Language

LIST OF ABBREVIATIONS

CHAPTER 1

Introduction

Studying evolutionary events in organisms remains a primary field of research in current biology. In modern developmental biology, such events can be assessed by observations made on a genetic level and comparing these to closely related species, such as isolated species that diverged from an assumed common origin. Charles Darwin himself devised the concept of evolution in his book “On the origin of species” [29] by observing finches on the island of Galapagos and comparing their specific phenotypes found on several of these isolated islands. Although he did not perform these comparisons on a genetic level, the idea of comparing differences and hence understanding the concept of evolution in more detail than before nevertheless remains a key driver of research even with modern technologies in the life sciences. While direct observations and modern technology such as Next-generation sequencing (NGS) are now widely applied impressively in the study of environmental DNA [204], specific evolutionary events as past population splits and timing certain favorable mutations in a population, cannot be studied in modern populations effectively. With the rise of ancient DNA (aDNA) research in the last years, new and improved means to study even wholly extinct organisms at a previously unprecedented scale are now widely adopted in evolutionary biology and related fields [97]. The most important improvement that ancient DNA can provide here is the possibility to directly observe evolutionary changes on a genetic level, even for extinct species. Thus, aDNA provides an original and translucent view of evolutionary events, which is why the field got immense attraction when initiated in the mid-80s of the last century[67]. The first studies on aDNA appeared when the experimental polymerase chain reaction (PCR) technology was primarily adopted in the life sciences. Within the first experiments, researchers had extracted ancient DNA fragments of the Quagga [67] and an Egyptian mummy specimen [152]. These studies are widely accepted as the first attempts at extract-

Chapter 1. Introduction

ing extinct species DNA and thus “kicked off” the entire field of aDNA research also shown in Figure 1.1. However, they are nowadays also seen as exemplary cases of fundamental authentication issues the field of aDNA still has to cope with. Especially in the PCR-era, researchers thought that there is almost no age limit for the extraction of aDNA from bones, fossils, and other remains. Publications were claiming authentic aDNA from Dinosaurs millions of years old and were published alongside with reconstructed DNA results from fossilized flies in amber, dated to be 120-135 million years old [22]. Later, many of these initial results have been identified to originate from modern-day contamination, including initial findings of Egyptian mummy DNA and the weevil. In the latter case, the DNA fragments found were identified to be stemming from a fruit-fly (*Drosophila melanogaster*) experiment in the same department. The finding of these authentication issues in widely accepted and published cases caused a trough of disillusionment amongst many aDNA researchers. Some researchers even contested the authenticity of any DNA fragments obtained from bone material. However, after a couple of years, the field moved on, and several authors proposed new experimental protocols to limit the effects of potential contamination under laboratory conditions [168]. Unfortunately, the previous findings claiming a practically almost non-existent time limit for the retrieval of authentic DNA was found to be wrong as well. Even the oldest specimens accepted to be authentic are now assumed to be up to 780 thousand years old, at maximum [149]. Important studies unraveling details about human history were, for example, the discovery of a new human subspecies, the Denisova hominin [179] and that Neanderthal hominins shared the same variant of the FOXP2 gene with modern humans, having a direct influence on the ability to speak [96]. As will be described in Chapter 3.6, aDNA can also be used to investigate forensic cases and provide novel information on deceased individuals, even in cases where there is no historical record of the deceased individual [163]. Other publications investigated the evolution of pathogens [6, 202] and potential transmission of these to new continents [11], which can provide useful information on the evolutionary adaptations certain pathogens share with modern strains. This information can help to potentially find treatment options for pathogens, especially in a globalized world where such diseases are now spreading more rapidly and wider than before [191].

Outline

This thesis is structured into three main parts. The first part is comprised of Chapter 2 and provides the reader with a basic background of genomic variation and the applications of population genetics methods on this variation. In particular, the background for the analysis methods this thesis centers around is explained, focusing on the DNA sequencing technology used to infer the nucleotide sequences of organisms and the challenges this NGS technology imposes for bioinformatics

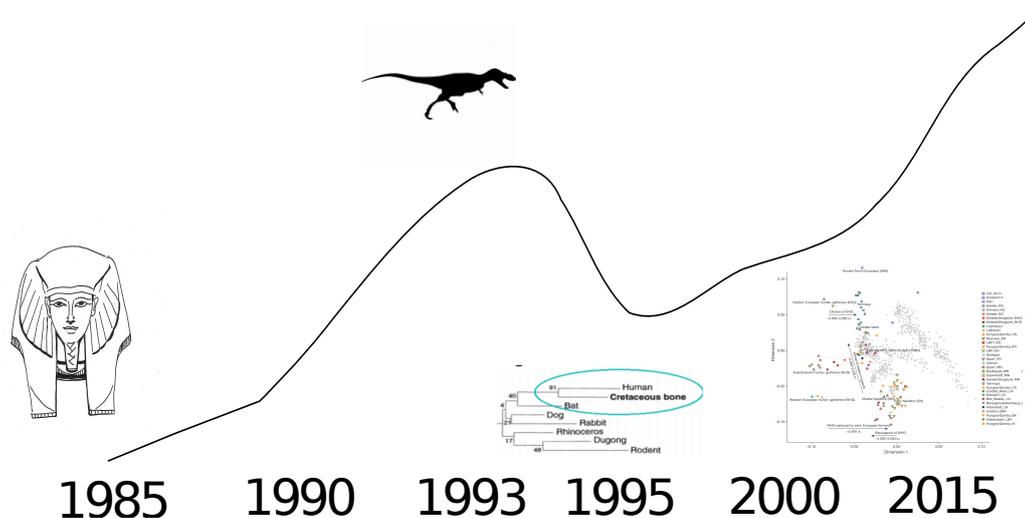


Figure 1.1: The “hype cycle” of ancient DNA research (adapted from [158]). After an initial innovation trigger in the 1980s, such as DNA allegedly found in ancient Egyptian mummy material [152] and a Quagga, the field started to analyze increasingly old material [22, 235]. Most of these claims were challenged and reverted in the next years in a so-called “trough of disillusionment”. Ancient DNA research then moved on to authentication and verification of results in the 2000s, reaching an ultimate plateau of productivity.

analyses in aDNA. Another part of this background chapter then focuses on the specific characteristics of ancient DNA, which is a central topic for the applications developed in this thesis and provides additional challenges for bioinformatics software development and application. Chapter 3 introduces the EAGER framework, the major ancient DNA processing pipeline that has been developed as a part of this thesis. Detailed information on all aspects of EAGER is provided in this chapter, including all tailor-made software applications that are combined within the entire framework for ancient DNA analysis. In the second part of Chapter 3, the pipeline is evaluated on a forensic use case to investigate the remains of George Bähr and demonstrate EAGERs capabilities on a real specimen.

The second part of this thesis starts with Chapter 4, where the mitoBench and mitoDB applications are introduced with a focus on the explorative features these applications provide for the analysis of mitochondrial data from human population genetics projects. Major achievements such as the user-friendly GUI and the main possibilities these applications provide to researchers in the population genetics fields are introduced and are compared to other methods used in the field until then. As an exemplary case for the application of EAGER and mitoBench, Chapter 5 then demonstrates the improved analysis workflows on both the automated part of modern NGS data analysis on ancient DNA from Egyptian mummy material and the applicability of mitoBench in the explorative data analysis parts of this

Chapter 1. Introduction

collaborative analysis project.

All the developmental and analytical work is then concluded with Chapter 6, providing an outlook on future implications of the introduced tools and methods with a discussion of the presented work and potential directions of future work building on top of it.

In summary, this thesis contains the EAGER and mitoBench/mitoDB applications as general frameworks improving the analysis of ancient DNA and mitochondrial DNA in population genetics. The described approaches target an area of active research and provide solutions for various issues, especially in reproducibility and usability. Furthermore, they facilitate better integration of state of the art tools than other available pipeline solutions in the field. As will be shown in the respective chapters, a wide application of the described frameworks is both easily possible and provides new means to study data from a variety of different projects within the scope of evolutionary biology and bioinformatics. Further extensions of the presented work in this thesis will be investigated by collaborators and other Ph.D. students to fill remaining open points in the integrative analysis platforms that both EAGER and the mitoBench/mitoDBproject exhibit.

CHAPTER 2

Background

This chapter provides a general background for the central topics of this thesis. First, a basic introduction to genetic variation on the DNA level is given, and some of the more commonly used methods to detect differences between populations or individuals in population genetics are explained. Next, DNA sequencing with modern NGS technology is introduced, first on an experimental level and later then followed by a general overview of the challenges of NGS technologies in data analysis. Afterwards, the specific characteristics of ancient DNA - around which most of the work in this thesis centers - are introduced, with a more tangible explanation of the bioinformatics challenges these characteristics impose on the data analysis for both biologists and bioinformaticians.

2.1 Genetic Variation

Of interest to researchers in evolutionary biology are typically not the common nucleotides among a certain species, but the variation that can be observed at a molecular level. With these molecular differences, researchers can, for example, try to predict variants directly in connection with certain diseases. Other possible uses of such variants include specific variants only found in certain populations, thus making populations differentiable and accessible to research in population genetics: Determining migration, admixture and other events for example in human prehistory. Single point mutations in the genome of an organism are the most commonly found variants and are called single nucleotide polymorphisms (SNPs)[131]. An exemplary screenshot of the application IGV [186] showing two such SNPs can be seen in Figure 2.1. As especially higher life forms, such as humans, reproduce sexually, thus receiving half of their genomic information from their mother and

Chapter 2. Background

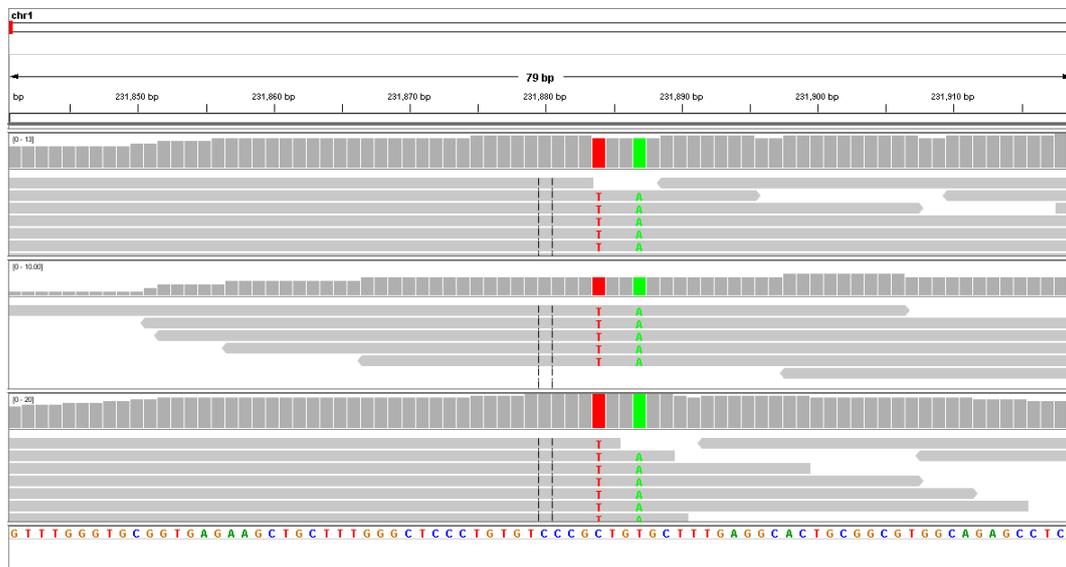


Figure 2.1: A screenshot generated with IGV [186], the integrated genome viewer application. Three genome tracks from different BAM files are shown (top, center, and bottom), all displaying the same SNPs (green and red here). Reads without substitutions are masked here in grey (the default of the application). The reference genome sequence can be seen at the bottom, showing a cytosine (C) and thymine (T) at the mutated positions normally.

a half from their father, they typically have diploid genomes, except for the sex chromosomes and the maternally inherited haploid mitochondrial genome. SNPs can, therefore, be found in either heterozygous (having two alleles, in other words, two different nucleotides) or homozygous (having twice the same allele) form. Due to their frequency in humans, SNPs are the genetic markers most commonly used to infer disease models, perform phenotypic prediction [231] and compare populations in population genetics [58, 106]. For financial reasons, researchers have developed a wide variety of different SNP panels (see Section 2.2.1 for details), that are nowadays used to determine known SNP targets for certain diseases or markers that are commonly used for population genetics.

Another form of genetic variation is referred to as insertions and deletions (INDELs). These can be introduced for example through DNA viruses (Eppstein Barr or similar DNA viruses) that use these approaches to survive in human cells. Deletions typically occur in cases where the recombination during meiosis fails or introduces errors. Other more large-scale variation examples are inversions, typically introduced during DNA replication and translocations, which describes cases where entire locations of a genome are found in a different part of the genome. A well-studied case of the latter are nuclear mitochondrial DNA segments (NUMTs), which are parts of the mitochondrial genome found in the nuclear genome of an individual [116].

2.2. DNA sequencing: Past & Future

Although all these different forms of genetic variation can play a role in diseases or disease risk, the majority of mutations on DNA level are not necessarily causing disease, for example when they are located in non-coding DNA regions [164]. Large parts of the human genome consist of such non-coding DNA. Recently, it has been discussed whether they have an effect for example on epigenetic activity [119]. The detection of variation is one of the most challenging fields in molecular biology and bioinformatics. Determining true variation in the presence of sequencing errors and other for example technical errors and limitations can cause inherent issues for the successful analysis of genetic data. A common example of such limitations can be found in the difficulty of determining large-scale inversions or translocations with short-read sequencing technology.

2.2 DNA sequencing: Past & Future

Modern biology would be almost impossible without the technological advances that started with the invention of Sanger sequencing and the chain termination method in the mid-1970's [193, 194]. The general target of DNA sequencing is the determination of the current sequential arrangement of the investigated organism's nucleotides. Thereby, we can obtain important knowledge about the basic principles of life itself: First, by identifying the nucleotide composition of the investigated organism and subsequently comparing these nucleotide sequences to those of other previously analyzed organisms to find similarities or differences. One of the biggest leaps forward in the last 20 years was the human genome project, where two large research groups were attempting to sequence the genome of *Homo sapiens*, which was still performed using the Sanger sequencing methodology invented in the 1970s [101, 229]. In 2005, the first updated methods came into place, largely phrased next-generation sequencing (NGS) with respect to the drastically changed methodology applied by these new technologies. After a couple of years, the number of NGS methods dropped significantly, and as of today, there are only a few providers still available at the market, among which the Illumina/Solexa sequencing by synthesis method is the most widely applied in the field. Sequencing by synthesis, in general, allows for the rapid sequencing of large amounts of DNA fragments in parallel, enabling a much faster and less cost-intensive sequencing process.

With up to date technology as the newest NovaSeq sequencers, the sequencing costs of a human genome already dropped to around USD 1,000 per genome, with expected further expense drops in the next couple of years [66]. Compared to a price of around USD-1 Billion for the first human genome [229], this is a significant decrease in price within a relatively short time frame of 16 years. The development of novel 3rd generation technologies such as PacBio SMRT [183] and Oxford NanoPore [16] sequencing is rapidly progressing, and the application of these technologies is in vogue at the moment. However, this thesis will focus on the state of

the art Illumina sequencing methodology, which is the currently preferred method used in the field of ancient DNA (aDNA) research.

2.2.1 Illumina sequencing & NGS analysis

Library preparation

The first step in Illumina sequencing is referred to as DNA library preparation. After DNA extraction, the DNA molecules need to be transferred to prepared sequencing libraries first. As ancient DNA is typically fragmented already due to post-mortem damage [195], there is no requirement to perform any ultrasound shearing, which is typically required to break the DNA molecules in smaller fragments that are susceptible to sequencing. During library preparation, DNA adapters are attached to both ends of the DNA fragments. These contain an anchor sequence that hybridizes with bound nucleotides on the flowcell during the actual sequencing process and the primer sequence with an optional index. The index sequence is used to perform multiplexing of several DNA libraries on a single sequencing run to save costs on HiSeq/NextSeq platforms [87, 134]. For ancient DNA projects, researchers typically use double-indexing protocols where two indexes are used per DNA library [88] to minimize cross-contamination with other samples sequenced in the same sequencing run. After the library preparation with adapter attachment to DNA fragments, the resulting DNA fragments are amplified using PCR amplification protocols, typically following a gel purification process and are then ready for the actual sequencing process.

Sequencing

The previously prepared sequencing libraries are put on a flowcell with oligonucleotide sequences that are complementary to the adapter sequences used for the DNA extract. Thus, the DNA fragments are immobilized on the surface of the flowcell in the instrument. Next, a bridge amplification process is started to amplify each fragment and produce clonal clusters to increase the signal strength of each sequence. During the actual sequencing by synthesis step, the library fragments bound to the surface of the flowcell serve as template sequences to which fluorescently labeled nucleotides are binding as illustrated in Figure 2.2. In each cycle, when fluorescently labeled nucleotides are added to the flowcell to bind to the template sequences, a laser is inducing the fluorescently labeled nucleotides bound to the template strands to excite light [54]. This light emission is measured with high frequency and resolution cameras. After that step, the remaining nucleotides are washed off, and the cycle is repeated for a pre-defined number of cycles (depending on the sequencing mode and protocol). Ultimately, this produces sequencing reads of the DNA fragments, which are then subsequently used in the NGS analysis.

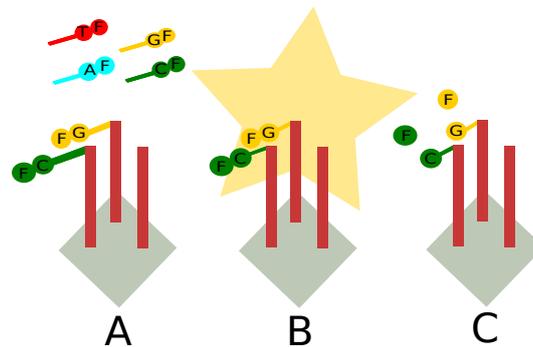


Figure 2.2: Sequencing by synthesis approach as used in Illumina sequencers [54]. A: Hybridization of fluorescently labeled nucleotides to the complementary base. B: A laser excites the fluorescence marked nucleotides and a color corresponding to the base incorporated is emitted. C: The fluorescence markers are cleaved off the nucleotides incorporated and are removed from the flowcells. The process (A-C) is repeated in several cycles depending on the utilized sequencing protocol. Figure adapted from Goodwin *et al.* [54]

Paired-End & Single-End Sequencing

The process described above produces single-end reads (SE), in other words, reads that have been sequenced from only one end of the DNA fragments in a sequencing library. Another sequencing protocol for Illumina sequencing is paired-end sequencing (PE), where both ends of a DNA fragment are sequenced. In this protocol, the reads form a read pair with a typically positive insert size between both reads, e.g., an area of the DNA fragment that is covered by neither of the two reads. For short DNA fragments, as commonly found in aDNA, the two PE reads to produce a negative insert size and show an overlap between both reads of the DNA fragment. Thus, reads can be merged with adapted applications and form a consensus/majority call consensus sequence in the overlap region to improve quality of the resulting merged reads. As sequencing instruments tend to have decreasing quality with higher cycle numbers [131], this can greatly improve the overall quality of the resulting merged reads and is therefore typically done in ancient DNA pipelines [162, 198, 199]. Concerning modern applications, PE reads also provide improved genome reconstruction possibilities in repetitive regions and hence better SNP and INDEL detection. Even though PE sequencing still has advantages compared to SE sequencing regarding quality, financial aspects make many labs switch to SE sequencing for most of their projects ¹.

¹Personal communication with Stephen Clayton, August 2017

Capture Enrichment methods

Although modern whole genome sequencing (WGS) with NGS technologies is now applicable in many cases and prices are still dropping [66], there are cases when a WGS analysis can still be inadvisable. First, the information gain that can be obtained with WGS approaches is sometimes not required, e.g., in exome analysis [25]. Next, there are applications when a WGS approach is not suitable, for example when the low DNA content of an investigated sample (the endogenous content) is increasing the overall costs to sequence enough of the target DNA. In both cases, capture and enrichment approaches are typically used. While applications in the medical community are targeting interesting SNPs for certain diseases [98], applications in the field of ancient DNA typically aim at enriching for endogenous target DNA [12, 202] and specific markers of interest in a population genetics perspective [58]. Most methods in ancient DNA research projects utilize in solution hybridization capture methods such as a bait-capture approach. A single-stranded DNA probe of a target organism is taken as bait, and synthetic biotinylated oligonucleotides are generated using a tiling approach [124]. The tiling method defines how often a single position of the target genome is covered by different bait probes. After denaturation of the probes, the baits can hybridize in a mixing phase with the DNA fragments in the DNA library. Using the strong biochemical binding affinities of streptavidin and biotin, the baits can then subsequently be pulled out of the library extract together with their bound DNA fragments. The general principle works in the same way as illustrated in Figure 2.3. After amplification, the enriched DNA library can then be sequenced in the same way as a normal DNA library. In this thesis, several target enrichment and capture methods have been used in the data acquisition phase of the respective projects. For the projects on George Bähr (Chapter 3.6) and the Egyptian Mummies (Chapter 5), the human mitochondrial capture method described by Maricic *et al.* [124] was used to obtain mitochondrial genomes for our investigated specimens. To further get high-density SNP capture information on George Bähr [163] and three of the ancient Egyptian mummies [201], the 390K and 1240K SNP capture panels defined in Fu *et al.* [44] and Haak *et al.* [58] were used. Even though these are the most commonly applied capture and enrichment panels in population genetics on ancient DNA [58, 210], there are several other specific capture and enrichment solutions available for pathogens such as *Mycobacterium leprae* [202], *Mycobacterium leprae* [11], *Yersinia pestis* [13] and *Treponema pallidum* [6]. Several companies such as MYBaits, for example, offer on-demand solutions for new capture and enrichment protocols [189].

2.2.2 NGS data applications & analysis

After sequencing, the obtained images with raw intensities require further processing to generate the nucleotide sequences for genomic assessment of the data. This

2.2. DNA sequencing: Past & Future

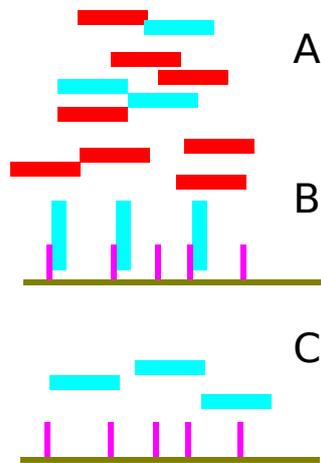


Figure 2.3: Basic principle of a capture-enrichment hybridization approach on an array. Of a given target organism, some DNA is fixed on an array or bead, and DNA fragments (A) can hybridize with these probes (B). Only reads that can hybridize with the probes are immobilized, thus after elution, only DNA fragments of the target organism are left (C) and can be amplified. The principle works in the same way for in-solution capture with the exception that instead of the fixed array biotinylated baits are fixed with on streptavidin-coated beads are used.

procedure is called demultiplexing and involves the processing of said images and subsequently binning of the multiplexed samples into individual FastQ files as a standardized data storage format for raw nucleotide information. There are several available software tools for that purpose, including Illumina's `bc12fastq` application and some competitor applications claiming improved data retrieval [180, 182]. FastQ files are text-based files with 4-line based information content. The first line contains a read identifier, information on the utilized sequencer and the sequencing mode the specific sequencer was used in (e.g., SE or PE sequencing protocol). After the second line, which holds the called nucleotides, a third line consists of a '+' character and some optional comment or repeated read identifier information. The fourth line then contains the respective quality information for the called nucleotides in line two. Typically, these are calibrated quality scores for each base that are obtained by sequencing a calibration library (typically a phage Φ X174 or specific *E. coli* strain) with the actual DNA load and using this as a golden standard for the quality of that sequencing run.

After demultiplexing, all NGS workflows have to follow certain steps to remove sequencing adapters. There are dozens of tools available for that purpose [162, 181, 199]. Some methods integrate quality trimming features with their adapter removal applications to account for the quality loss with increasing cycle numbers. After adapter removal and quality trimming, reads can be mapped against either a reference genome following a *mapping approach* or assembled *de novo* when for example no known reference genome is available. Subsequently, variants

Chapter 2. Background

are typically classified with variant calling applications to detect variation in the investigated samples. In addition to SNPs or small INDELs, NGS also facilitates the detection of larger variation such as copy number variations, or chromosomal rearrangements [39]. However, highly repetitive genomes still pose a significant challenge to NGS methods due to their inability to span repetitive regions with short reads and thus being unable to differentiate between individual repeats. In such cases, novel methods with long read technologies such as NanoPore sequencing or PacBio SMRT technologies are more suitable and are slowly taking over areas in genomics where these resolution dark spots are crucial for analysis success [54]. Although there are some shortcomings, large-scale projects in the era of personal genomics such as the recent Danish sequencing project with 150 sequenced Danish Genomes [123], 15,220 Icelanders [78] and most famously the UK Biobank 500K genomes project [21] are still demonstrating the wide applicability of short read methods. In other areas such as epidemiology and field work, NanoPore technology has largely taken over. With astonishing achievements such as real-time surveillance of flu strains [139, 140], Ebola [51, 171] and lately also Zika [38], the field of NGS applications is even further growing with newer experimental methods and sequencing equipment. Exciting times are ahead for biologists and bioinformaticians!

2.3 Specific characteristics of ancient DNA

2.3.1 DNA Damage: What is left

While most research projects focus on the reconstruction of modern genomes, ancient DNA projects focus on the reconstruction of (parts) of ancient genomes of long deceased organisms. Although there is some dispute on what ancient means with respect to age, a common definition bases this classification on the amount of DNA damage found in ancient DNA. Shortly after the death of an organism, the DNA repair mechanisms of the cell cease to work, and thus the DNA degrades. External factors such as water, heat, multiple sorts of radiation (UV, cosmic microwave background) but also intracellular metabolism products all contribute to this form of degradation of unprotected DNA. This form of natural degradation results in the DNA fragments to be broken down into much smaller fragments, with a typical shift of DNA fragments towards short lengths of 30-60 bp length as illustrated in Figure 2.4. Ultimately, this means that the amount of recoverable DNA is changed due to fragment size distributions less favorable for modern sequencing methods. Apart from these degradation processes, other chemical modifications are notably changing the “shape” of the DNA. In many cases, nicks are introduced during a hydrolytic attack on the phosphate backbone of the DNA [63, 71], increasing the instability of the DNA strand. Other modifications mostly related to hydrolytic attacks of the DNA fragment are com-

2.3. Specific characteristics of ancient DNA

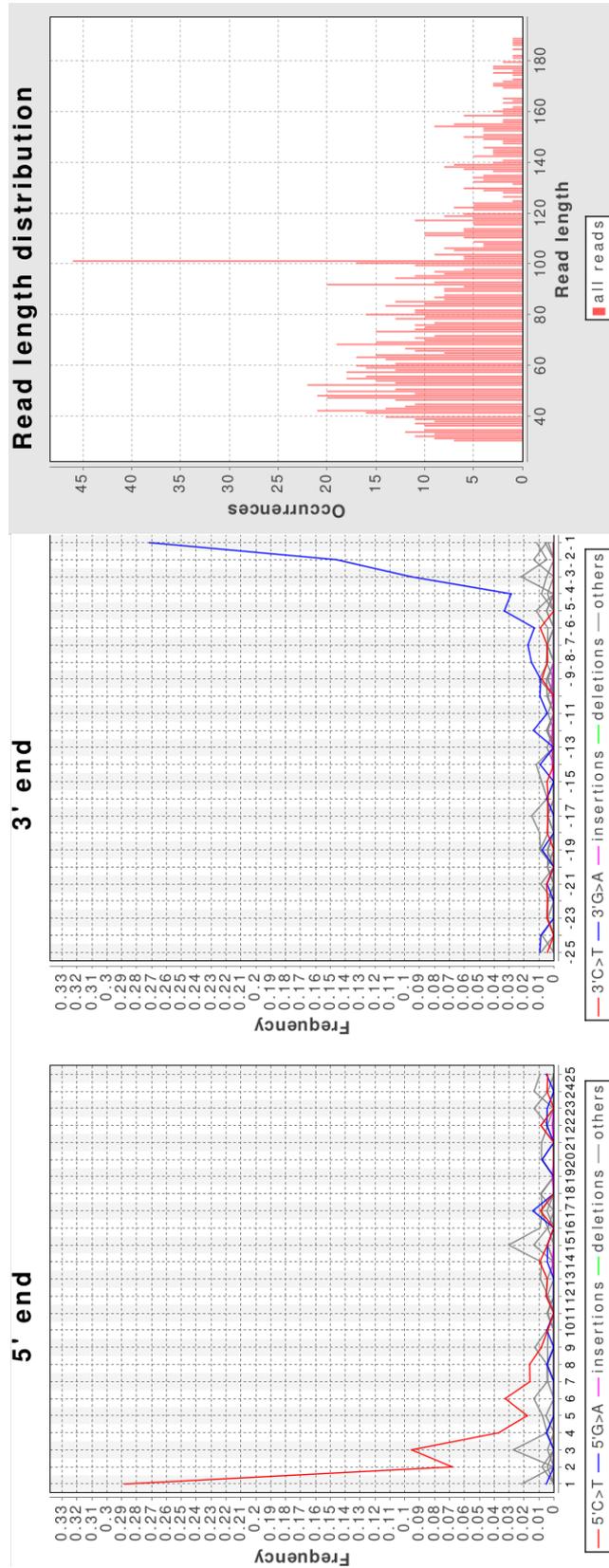


Figure 2.4: Typical aDNA misincorporations on both 5' (left) and 3' (center) of NGS reads. The total frequency of C-to-T and G-to-A misincorporations is increased at the terminal ends of the sequencing reads. Next, to the misincorporation plot, the typical read length distribution of an aDNA library is shown with the average shifted to the range of 30-60 bp length, significantly smaller than for modern NGS reads.

Chapter 2. Background

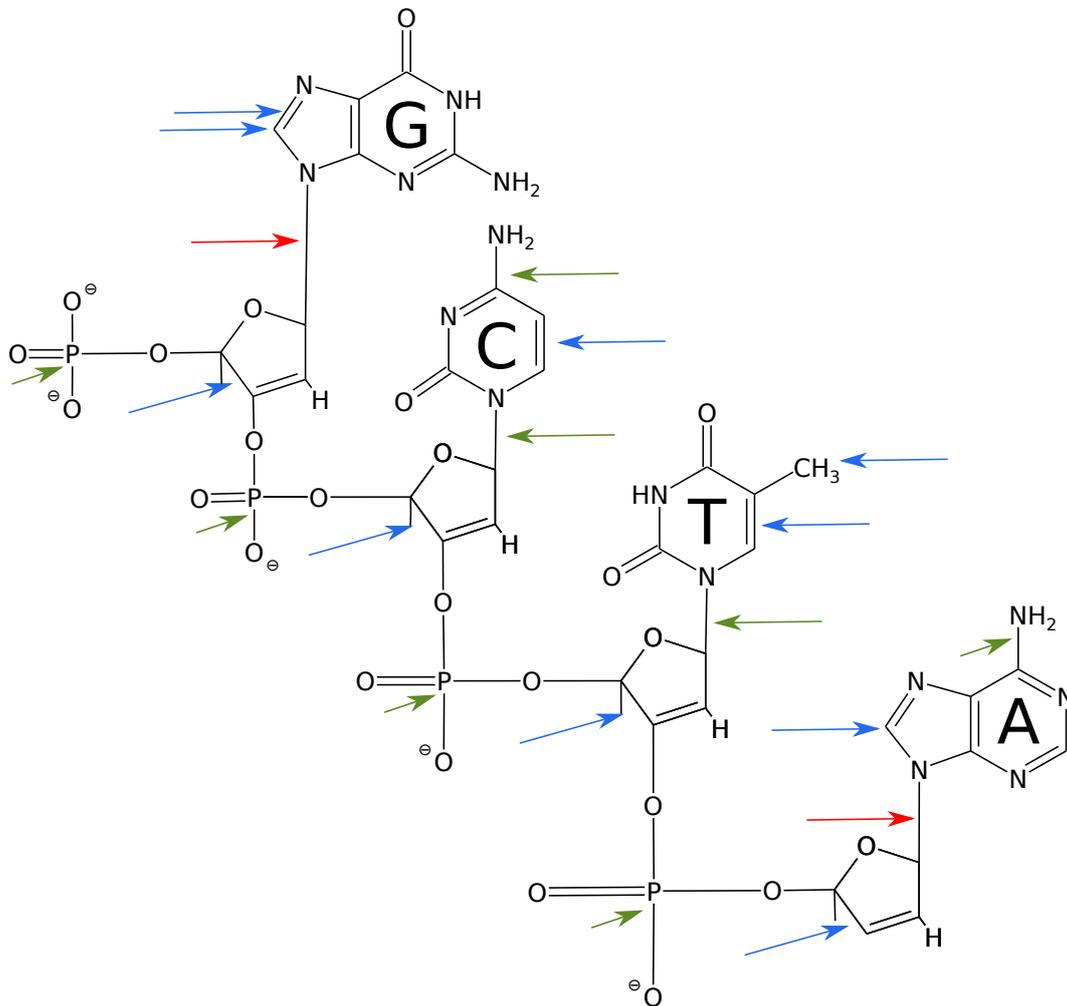


Figure 2.5: The various kinds of DNA damage that can happen on a chemical level. Oxidative damage (blue arrows), hydrolytic damage (green arrows) and loss of purines (red arrows) of here. Adapted Figure from Hofreiter *et al.* [90].

only found as double-strand breaks, base loss, and cross-links. Although these modifications happen quite frequently, the most abundant and observable DNA damage modification is the deamination of cytosine to uracil in the presence of water [17, 19, 69, 195] as depicted in Figure 2.5. In general, this should not be problematic, as the introduced uracil is treated by the DNA polymerase during PCR or library preparation for NGS as a Thymine. This results in an inserted adenine, subsequently producing a characteristic C-to-T and G-to-A modification at the ends of the ancient DNA reads as illustrated in Figure 2.4. While this specific pattern is nowadays also seen as useful for authentication purposes, it can produce significant issues in the downstream bioinformatics procedures due to the incorrectness of obtained sequencing reads, which are often poorly alignable to reference genomes for example. Especially in the context of genome assembly,

2.3. Specific characteristics of ancient DNA

where overlaps between reads are crucial for assembly success, the introduction of base misincorporations can cause severe issues and require further assessment of analysis results, with specific adjustments for ancient DNA. Several attempts on the experimental side, therefore, tried to overcome these issues by applying a specific UDG (uracil-deglycosylase) treatment [18]. This experimental procedure can be used to repair such deamination patterns at the ends of the DNA utilizing the same repair mechanisms as a living cell [188]. Although the approach works well, many research groups have now moved to partial UDG treatment [188], to keep a certain level of DNA damage for authentication purposes [133, 208].

2.3.2 Environmental DNA: The 1% curse

The second major issue in aDNA projects is the low number of endogenous DNA that is commonly found in ancient remains. Researchers refer to endogenous DNA as the part of the DNA isolated or assessed in a sample of an investigated target organism. The non-endogenous part, therefore, represents the amount of environmental DNA of several other species and organisms in an investigated sample. Unfortunately, DNA from microbes, fungal species, and other organisms typically makes up the largest proportion of DNA in samples as also illustrated in Figure 2.6. In many cases, the amount of endogenous DNA found is even below the level of 1%. This proves to be an increasing problem for several reasons: First, due to the low number of endogenous DNA fragments found in a sample, the financial effort to sequence enough of the target endogenous DNA is high, while producing lots of data for the non-endogenous part of the investigated sample simultaneously. Second, the direct assessment how unique the obtained DNA fragments are is affected drastically, too. Although prices for DNA sequencing with modern Illumina methods are still dropping dramatically [66], sequencing a sample with less than 1% of endogenous DNA in a whole genome shotgun approach is still not cost-efficient. A common solution for the issues with environmental DNA is capture and enrichment approaches. There exist several published and widely applied bait-capture / in-solution capture approaches which can be used to enrich for parts (e.g., human mitochondrial DNA) or even whole genomes (e.g., for *Mycobacterium tuberculosis* [202]) of specific organisms or species. Due to limitations in the capture array and chip design, these approaches only work with a limited number of selectively captured positions on a genome. For whole human genomes, this approach works well for published capture protocols such as the 390K, 1240K [58] and some commercially available ones (e.g., an archaic capture protocol [42]). However, there are potential issues with these capture approaches, especially when the investigated organism has diverged substantially and therefore the capture design does not capture the true variation that would have been present in the ancient genome [12, 68]. Other problems include genetic regions that are heavily interspersed with repeats, making the array design hard or even impossible [12].

2.3.3 Contamination: Background noise - or real signal?

Another important issue that can arise in ancient DNA projects is contamination. Contamination can occur in multiple forms: First of all, there is potential contamination with endogenous factors that do not necessarily manifest themselves as DNA itself. Most of the archaeological or forensic samples were in contact with soil and other environmental elements. Especially bone material is known to be prone to external factors and can function similar to a sponge, “soaking up” other material, particularly when found in humid conditions [81, 126]. While some of these external factors contribute to DNA damage in general, others are inhibiting experimental analysis such as library preparation. If the sampled material has been contaminated with environmental factors, for example, certain metal ions [126], which bind exclusively to DNA polymerase, the sample acts biochemically inert to downstream experimental protocols. On an experimental side, these problems are typically addressed by performing elution experiments to determine whether the inhibiting factor can be diluted until the DNA fragments are accessible by the DNA polymerase again. Unfortunately, this also reduces the amount of DNA accessible in the sample, as the dilution process affects both the contaminating as well as the endogenous DNA content. Some experimental techniques, such as humic acid treatment [137] and silica-based methods [82], have been proposed to remove the inhibiting factor, but these can only resolve the issue in some cases. Sometimes, samples are inhibited up to an extent where nothing can be done anymore [197].

The second form of contamination that researchers in aDNA projects have to deal with is human contamination of samples. There are multiple occasions when such contamination can be introduced. Most prominently, this form of contamination is brought into samples during excavation, sample retrieval for example from archives and sample cleaning after excavation in an archaeological context. However, also the subsequent archival handling, cataloging or other museum-related interactions with sample material can introduce human contamination. A simple skin contact without gloves, a lost hair or even just skin flakes of shed skin can introduce substantial amounts of human contamination to samples. Even though protocols have been improved to minimize DNA contamination at all costs, there is still the possibility of introducing human contamination during DNA extraction or sample handling. The key problem with human contamination is again linked to the typically low endogenous DNA content of old samples. If samples are then contaminated with modern DNA introduced during handling, extraction or otherwise, the modern DNA can obscure the analysis results obtained on authentic ancient DNA. If the contamination is not taken into account for example in downstream variant analysis or population genetics, results can be altered dramatically and thus ruin an entire analysis project. As exactly this issue of modern human contamination was affecting lots of samples in the early days of ancient DNA research [152], researchers came up with updated protocols to tackle these issues at an early stage [168]. All major research groups that are working on ancient DNA

2.3. Specific characteristics of ancient DNA

projects utilize separated clean rooms with UV radiation, bleach bone material before extraction and use clean room suits. Other measures include the inclusion of DNA library blanks and extraction blanks to keep track of how much “free form” DNA is around in the laboratory to keep this influence as low as possible. In some cases, even archaeologists are nowadays using protective gear and gloves to keep contamination low during excavation of samples.

Contamination with closely related organisms, for example in ancient pathogen projects, is less well investigated. The presence of certain closely related species of genus *Mycobacterium* could potentially also alter results obtained for ancient cases of *Mycobacterium tuberculosis*. With an estimated number of close to 100 different species and some of these spread in common soil [61], contamination with ancient *Mycobacterium* organisms could infer with downstream analysis today.

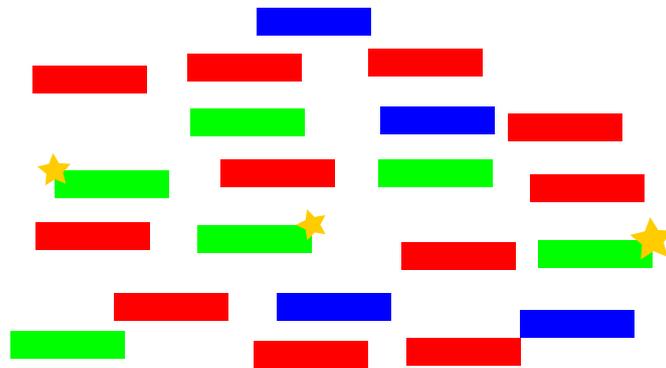


Figure 2.6: The various types of contamination that are found commonly in aDNA research projects. Green reads show actual endogenous DNA fragments, some of them with DNA damage (stars). Blue are contaminant reads from fungal or microbial sources. Red is DNA from present-day human contamination. Targeting in such an example only human endogenous content can be difficult and challenging, both from an experimental and an in-silico point of view.

2.3.4 Recent improvements

Although the described issues are still present in modern projects on ancient DNA, there have been several improvements in the last years that promise to resolve at least parts of the previously mentioned issues on a more general level. First, the experimental methods have been improving dramatically since the early days of aDNA research. Especially laboratory and experimental protocols have been adapted to keep introduced contamination as low as possible [168]. Further improvements in the treatment of bone material for example by the utilization of UV radiation and bleach to decrease external contamination with other source DNA have also proven to be successful [83]. Other more current methods try to decrease the amount of microbial and human contamination before sequencing [92]

by using specialized biochemical treatments during library preparation. More recently, improvements in DNA extraction protocols - mainly by focusing on the human petrous bone have shown promising results [166]. With increased rates of endogenous DNA, sometimes by several orders of magnitude higher endogenous yield than in teeth or long-bone material [166], these methods now even pushed the borders of authentic ancient DNA retrieval to more humid and hot areas of the globe [196, 209, 210]. Previous attempts were mostly failing due to too low endogenous DNA content of the investigated samples on samples found under such hot and humid conditions [166]. Further progress made on top of the initial DNA extraction method [205] aim at improving sample handling with less destructive petrous bone sampling approaches, thus increasing the agreeability of the sampling method with curators of large collections. Combined with efforts to push the recoverable part of ancient DNA from samples with improved single-stranded protocols, for example, [49], enabled the reconstruction of otherwise single-stranded samples in Europe already [132]. Overall these new experimental methods increase the confidence that the methodological and experimental progress can still provide results that were infeasible a couple of years ago.

2.4 Bioinformatics challenges in aDNA

Bioinformatics plays a key role in both modern and ancient genomics projects. Even for modern DNA data generated with NGS technologies, there are substantial challenges involved in reconstructing entire genomes, annotate these and perform downstream variation calling [30]. Accurate methods to map generated sequencing reads to a reference genome have to solve complex algorithmic challenges [109, 111]. For ancient DNA, these methods have to be able to deal with misincorporation patterns and on average even shorter read lengths [195] as shown in Section 2.3. Microbial contamination or other (non-organism specific) contamination is usually not critical for reconstructing the genome, whereas in cases with closely related species' DNA in the non-endogenous content of the investigated sample, this can cause severe issues. Variant detection methods such as the GATK [30] are susceptible to contamination issues, and their results can be heavily skewed, especially if the ancient sample contains a high fraction of modern DNA, which is often the case. However, the characteristic misincorporation patterns are typically not an issue in variant calling, as the likelihood of having several reads with the same misincorporation pattern being mistaken for a variant is low. Furthermore, such reads would usually be seen as duplicates of each other during duplicate removal procedures, further reducing the likelihood of variant calling errors being introduced due to misincorporations alone. Specifically the issues with modern human contamination have been tackled with several in-silico methods such as *contaMix* [43, 45], *Schmutzi* [181], *ANGSD* [93] or *DICE* [173]. *ContaMix* performs an automated comparison of found mitochondrial mutations and compares

2.5. *Bioinformatics challenges in genomics*

the set of detected variants with a database of known putative contaminants to determine whether there are multiple conflicting “mutation profiles” present in the investigated sample. Schmutzi improves upon this using a Bayesian framework likelihood approach, also integrating DNA deamination patterns and separating the putative contaminant mitochondrial DNA from the endogenous part. ANGSD provides an X-chromosomal authentication method which estimates the amount of X-chromosomal heterozygosity found in the investigated sample to determine contamination. DICE uses population genetics approaches to determine whether the investigated sample shows a shift in a demographic model to detect contamination. Other researchers utilize the damage patterns or misincorporations of authentic ancient DNA to filter for NGS sequencing reads that show these and remove all reads that do not show these patterns [208]. PMDTools can be utilized to select for reads showing significant DNA damage and thus misincorporation. However this approach works well, the number of reads in samples where PMDTools is applied can be reduced to an extent where further downstream investigation becomes infeasible. Other researchers try to further minimize the effect of misincorporation patterns in population genetics by restricting their analysis to transversions, thus removing any potentially skewed biases of aDNA misincorporations. Unfortunately, this often reduces the number of usable variants for analysis methods such as PCA or f statistics by orders of magnitude and can thus result in too sparse results due to lack of resolution.

2.5 Bioinformatics challenges in genomics

Apart from the specific challenges for bioinformatics in aDNA projects, there are multiple other challenges present in genomics for bioinformatics. One of the biggest issues for novel methods, applications, and tools in bioinformatics is the massively growing amount of sequence data generated. Since technological advances enable a cheaper and much faster analysis of, e.g., human genomes, this also imposes challenges for bioinformatics to be able to keep up the required speed on the analysis side.

Furthermore, this rapidly growing pile of data generates issues with the integration of previously published and analyzed data. Specifically, when novel datasets have to be re-analyzed in a context with older data and using newer methods, bioinformaticians typically find themselves in a “dependency hell”.

In such cases, simple scripts do not suffice anymore to be able to keep up the required speed in the analysis of data. Containerization approaches such as Docker [130] or Singularity [100] can at least provide ways to ensure that software dependencies are reproducible, while this has also to be supported on the workflow side.

The second general challenge for bioinformatics is to provide accessible methods.

Chapter 2. Background

Accessible in this context means, that users without a broader bioinformatics background should be able to apply bioinformatics tools and methods at least. Aforementioned can be achieved by providing up to date documentation, user manuals and descriptive “How To’s”. Another approach to empower users to utilize bioinformatics tools is a general trend to provide graphical user interfaces (GUIs) or even web user interfaces (webUI’s) to eradicate the requirement of a local installation of tools.

The general underlying concept of this thesis was centered around generating novel methods and applications to improve aDNA analysis. During the development of EAGER, much dedication went into designing a pipeline to make the pipeline execution accessible for users without a bioinformatics background. The same motivation went into the design process of mitoBench and mitoDB, taking the experiences in maintaining EAGER during the last three years into account.

CHAPTER 3

EAGER: Efficient ancient genome reconstruction

Text and figures in this chapter were adapted with modifications from our work previously published in Genome Biology [162].

3.1 Introduction

While the first attempts at retrieving aDNA from ancient material were made using PCR based technologies in 1984 [67], these methods nowadays have been widely replaced by modern NGS applications [90]. One reason is that methods such as PCR exhibit significant drawbacks when applied to aDNA samples, as they target longer DNA fragments primarily for fundamental technical reasons [170, 200, 218]. As aDNA has typically shorter fragment lengths [95, 144, 177](Figure 3.1) even under almost optimal conserving conditions such as for the Tyrolean iceman [80], the amount of observable contamination in a sequencing sample is technically increased by utilizing PCR based applications. Aforementioned is mainly related to selecting the largest fragments obtainable in the sample, while aDNA shows typically smaller fragment lengths and is therefore subsequently less favored by PCR based methods. The second reason for the rise of NGS based methods in the field of palaeogenetics is the possibility to cost-efficiently sequence even whole genomes of organisms [90]. While previous approaches such as PCR or Sanger-based sequencing were typically used to target specific genes or for example hypervariable regions in the mitochondrial DNA (mtDNA), these new methods are now enabling researchers to sequence mitochondria and even whole human genomes. Combined with unrivaled speed and cost efficiency, this led to a large increase in data being generated in the last couple of years [58, 106]. The development of hybridization techniques to capture interesting markers on genomes for

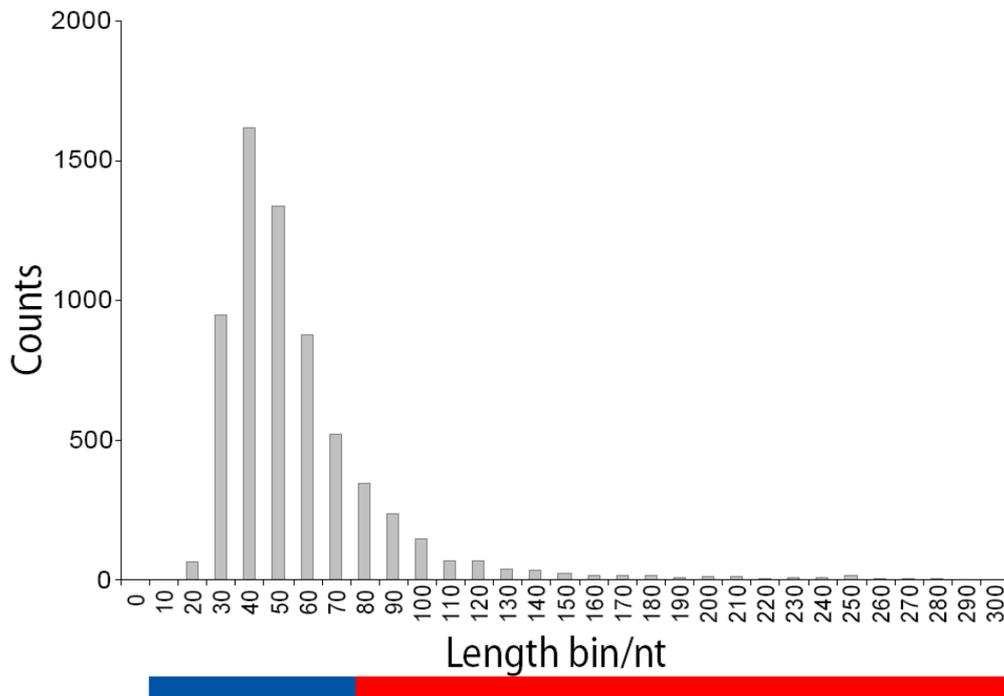


Figure 3.1: Theoretical fragment size distribution of an aDNA sequencing project. As can be seen, the majority of aDNA fragments show a peak at fragment lengths of 30-60 nucleotides. Marked in red are fragment lengths that can be captured with PCR methods, while NGS methods are much more suited to capture specifically the shorter DNA fragments, here marked in blue. Note, that real samples might have different distributions of fragment lengths, e.g., due to modern DNA contamination present. Figure adapted from lecture notes in Palaeogenetics [200].

a variety of organisms [106, 124, 202], enabled researchers to target even samples with low levels of DNA available. In combination with the usage of HTS methods, this permitted the analysis of large cohorts of organisms. As the whole field has moved on from smaller genomic marker regions like mitochondrial genomes [57] to high-density SNP capture [58] or even whole genomes, methods being used in the successful analysis of genomic data from ancient specimens need to be able to scale to account for these requirements. Intrinsically, this poses various problems to computational biology, both in technical aspects such as software development, algorithms and in software design to improve the usability of *in silico* methods. Modern methods in aDNA analysis need to be able to scale well regarding processing power, requiring efficient algorithms capable of dealing with complex analysis procedures under limited resource availability. During the early years of NGS, the majority of established methods used, e.g., in Sanger sequencing or PCR based analysis procedures, could not be applied to the millions of short reads produced

3.1. Introduction

by NGS methods. In many of these cases, these limitations arose from memory consumption or processing power restrictions [162, 198]. On the other hand, while these can be seen as purely technical issues, there exist several characteristic issues arising from the interaction that researchers undergo with complex analysis systems required for the analysis of large datasets in computational biology. In increasingly interdisciplinary fields such as aDNA research, not all researchers have a background in bioinformatics or computational biology. An issue arising from this can be the cumbersome analysis of datasets with complex computational methods, that are hard to apply even with such a background. Unfortunately, a substantial portion of bioinformatics software is poorly documented and lacks even basic usage guidelines, disabling researchers from other fields to apply them [114]. A further aspect that confines many researchers is the missing availability of user-friendly interfaces such as an advanced command line interface (CLI) with a help function, for a more fail-proof user interaction. Even fewer examples of computational biology software include graphical user interfaces (GUI), which enable an even more intuitive and user-friendly access to more complex analysis procedures.

An aspect of growing importance in modern data science and analysis is furthermore seen in the possibility to reproduce analysis results, e.g. after initial publication. A drawback of increasing cohort sizes in aDNA research is, that typical projects require previous analysis results to be taken into account. Although this is important in terms of availability of comparison data for various reasons, this also requires either a reprocessing of a whole previous project or at least incorporation of downstream analysis results from a multitude of different sources. This also means, that results previously published need to be reproduced to a certain extent, meaning that researchers should have easy methods at hand to reproduce their previous findings, ideally in an automated fashion [174]. While laboratory notebooks and to a certain extent also digital documents or laboratory information systems (LIMS) aim at improving the situation on the experimental side [115], methods in computational biology often rely on complex dependency management systems. Unfortunately, this renders a reproducible analysis workflow in bioinformatics cumbersome, especially when a multitude of different tools, software methods, and smaller scripts are involved in a complex analysis workflow. Various approaches to tackle these issues have been proposed, sometimes integrating a whole analysis workflow in a workbench like environment, for example in a Jupyter notebook [174]. While these approaches can be useful for interactive and explorative data analysis, the applicability for large-scale and fully automated analysis workflows remains limited.

Several workflow systems have been published recently to tackle exactly these remaining issues while providing efficient methods to enable entirely reproducible research for large-scale analysis projects [31, 230]. Even though these prove to be useful for making the analysis procedure itself reproducible, methods used by these workflow frameworks are not kept in a specific version by default. Considering

that results can differ widely between two versions of a single tool, this restricts reproducibility of the workflow itself. Extensions of the proposed workflow frameworks, therefore, aim at integrating the workflows with Docker containers to freeze respective analysis tools for future use and employ these in a versioned approach to be able to entirely reproduce an analysis procedure both regarding the actual workflow, as well as the utilized software dependencies [31].

Methods to aid in the analysis of aDNA have first been introduced by Martin Kircher [87], who published a guide of best practices for aDNA reconstruction in general. Building upon this, other groups have implemented general analysis frameworks such as PALEOMIX [198], incorporating improvements aiming at more scalable and automated analysis workflows for aDNA. The initial workflow described by Kircher was designed to reconstruct only small genomes of bacteria and mitochondrial data from archaic hominins and could therefore not be quickly applied to large-scale datasets from NGS methodologies. Even the much more advanced PALEOMIX workflow, although providing technically improved methods for large-scale genome analysis, was relatively difficult to use and required manual creation of configuration XML files, impairing the usability for researchers without an advanced background in computational biology.

Aiming at resolving and tackling all of the mentioned issues, the EAGER project had been started in 2013. EAGER, short for “Efficient Ancient Genome Reconstruction” is a pipeline for the reconstruction of ancient genomes, integrating several of state of the art methods for the reconstruction of ancient genomes, combined with updated and improved methods that have been developed within the EAGER project. Aforementioned ensures that the pipeline provides a state of the art solution in functionality, addressing entire aDNA research projects to be analyzed within a single software solution. Users are provided with functionality to run various research projects on human, bacterial and other organisms’ genome data without having to learn the basic usage of all the methods in the background. With additional features such as the possibility to restart incomplete analysis runs, default parameter sets for certain parts of analysis procedures following community recommendations, the pipeline aims at providing a central data analysis workflow for a variety of aDNA analysis procedures. It was designed to help researchers from a broad background spectrum to perform their analysis. To further add upon the state of the art functionality of the pipeline, EAGER features a GUI that can be used to efficiently configure an analysis procedure on up to hundreds of samples with a few clicks on a regular desktop computer. Furthermore, the pipeline comes with an extensive user manual, describing several possible use cases for the pipeline as well as explaining the meaning of all analysis modules used in the workflow. An analysis report with exports to various file formats typically used in computational biology (e.g. Excel, CSV, HTML, PDF) is produced upon successful analysis runs, making a quick assessment of analysis results possible. To fully enable researchers without larger hardware resources to run their analyses too, EAGER is distributed in various formats, ranging from executables to enable local or cluster installation

as well as Docker containers [31, 130] enabling truly reproducible and versioned research. All versions of the pipeline can be accessed on Docker Hub, thus making pipeline downgrades possible. Users can, for example, downgrade to the pipeline version available at a certain time point, making the re-analysis of a whole project with a defined software environment and version rather easy. The execution of methods running with EAGER is organized in a way that ensures a highly parallelized execution of tasks, thus making the pipeline applicable to even large-scale datasets with hundreds of gigabytes raw data input.

3.2 Development

Both the GUI and the command line interface (CLI) are implemented using the Java programming language.

3.2.1 Command line interface

To be able to separate the configuration of the pipeline from the actual execution, EAGER has two basic components: The CLI component for pipeline execution and the GUI which can be used to generate a configuration. The command line interface is responsible for the execution of individual modules, as depicted in Figure 3.2. The CLI component `RunEAGER` runs the pipeline itself by calling the required tools and methods on the analysis workstation or cluster. To enable an easily extensible pipeline development, the idea of a more general framework for the execution of individual tools and methods was used. We generated abstract module definitions `AModule`, which contain a set of parameters, configuration instructions and naming conventions for their respective output. For example, a module always possesses an input file and an output file. Every module in the pipeline is based on these predefined sets of variables, thus ensuring consistency within the pipeline. These modules are then pooled in `ModulePools`. For example, a mapping process usually requires an index reference genome. Therefore, these strongly dependent individual modules of indexing are bundled together in a separate module pool to ensure they are executed after each other in all cases. With these module pools, we aim at making the reuse of parts of the pipeline possible.

Individual modules and module pools are then subject to execution in the CLI component by utilizing the Java `ProcessBuilder` method. To be able to trace back which modules have been executed, how long these have been running and what kind of standard (error) output they produced, the CLI component produces a text file called `EAGER.log`. The executed processes with the exact parameters along with runtime information, a time stamp of their execution and any information written to `STDOUT` and `STDERR` are logged as well. To keep this easily observable, the log file contains this information in an interleaved format, thus making it easy to trace back which module or tool caused potential errors for example.

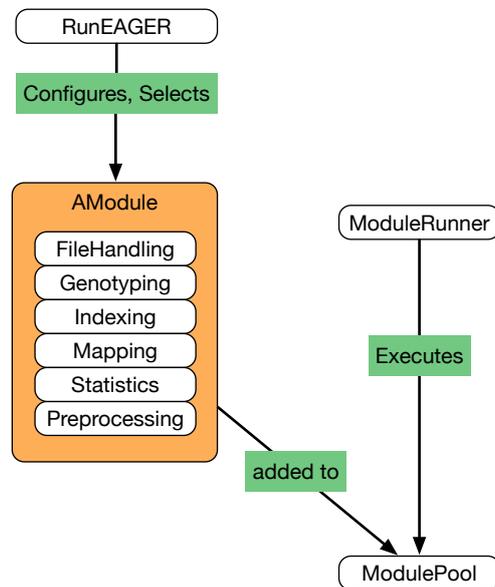


Figure 3.2: EAGER CLI components as implemented in the pipeline. The RunEAGER component configures and selects required modules for an analysis procedure. Modules are abstract modules, each implementing the same set of standard methods, e.g., their parameters. These are added to a ModulePool that is executed by the ModuleRunner iteratively one after each other. In general, the model would allow for the execution of individual tasks simultaneously, but some design decisions are prohibiting this, e.g., if processes use more than a single CPU core having multiple of these in parallel can cause issues on cluster systems. This general abstraction provides the possibility to efficiently write new modules for new methods or adapt existing ones without further changes in different parts of the pipeline.

Following up on the GUI's capabilities to search for input files, the CLI component finds XML configuration files produced by the GUI in a provided path. This enables a rapid execution of all configuration files in a given path, without the need to specify them directly. For example, a user can specify a folder containing multiple samples instead of selecting them individually, greatly simplifying the analysis of multiple samples. The CLI component features functionality to search the file system for configuration files.

3.2.2 Integrated tools and methods

The EAGER pipeline provides access to several state of the art methods used for the analysis of aDNA samples. The following sections will provide a more detailed explanation of the available tools. Additionally, contributed methods that aim at improving the analysis results are introduced in these sections.

Preprocessing

After NGS sequencing, the created reads are typically analyzed in a preprocessing step to ensure the data has high technical quality and follows standard expectations regarding format, compression, and similar metrics. The EAGER pipeline includes several methods to preprocess raw sequencing data. These include FastQC [5] for an automated raw quality assessment of NGS data. Furthermore, EAGER provides two methods for adapter removal and read merging, Clip&Merge [162] and AdapterRemoval v2 [199]. Ancient DNA typically shows characteristic post-mortem damage, arising shortly after the death of an organism and steadily decreasing read lengths while simultaneously increasing misincorporations due to deamination [2, 89, 195]. This renders the analysis of aDNA data difficult with modern NGS methods, specifically when mapping reads to a reference genome, and even more for de novo genome assembly methods. A common approach in the analysis of aDNA is, therefore, the merging of paired-end reads to improve the overall quality of reads from aDNA [162]. A further aspect here is, that unmerged reads with negative insert sizes between read pairs cannot be mapped by common modern read mapping algorithms such as BWA. To compensate for this, the Clip&Merge application has been developed within the scope of the EAGER project to combine adapter clipping, read merging of paired-end reads and read trimming of non-merged reads based on a user-defined quality threshold. To achieve this, Günter Jäger has developed a clipping strategy in his Ph.D. thesis [76] that was motivated by the technique implemented in the FASTX-Toolkit [222], additionally speeding up the analysis by utilizing multi-core CPU capabilities to parallelize the clipping and merging procedures. The method identifies adapter sequences at the ends of the reads by using a local alignment based on the Smith-Waterman algorithm between the adapter sequence and the read investigated [215]. After successful testing, the AdapterRemoval v2 method was integrated into the pipeline as an alternative to Clip&Merge. Except for the removal of additional barcodes, the method can be used for adapter removal, paired-end read merging and quality trimming of reads that show inferior base qualities towards the respective read ends. By default, the EAGER pipeline utilizes the Clip&Merge method. However, this can be adjusted by the user in the GUI.

Mapping

The process of read mapping describes the alignment of short DNA reads to a reference genome. In NGS projects, this is typically one of the computationally most demanding steps. There are hundreds of different algorithms available to map sequencing reads to reference genomes, but only some are commonly used in aDNA projects. EAGER features BWA [109], BWA-mem [108], Bowtie 2 [102] and Stampy [120] as the algorithms of choice for read mapping to reference genomes. Other tools that are part of the mapping category are SAMTools [109] and the Picard-Tools (<http://broadinstitute.github.io/picard/>) that are used in

the background for conversion, extraction, and file handling purposes in various intermediate steps in the pipeline.

Genotyping

Another important step in NGS analysis is the identification and verification of variants on mapped read data. This process is called genotyping, or variant calling and EAGER includes several applications in the pipeline for performing variant calling on mapped read data. The Genome Analysis Toolkit (GATK) UnifiedGenotyper and HaplotypeCaller [129] are the two most widely used genotyping applications and are both supported in EAGER. Additionally, the ANGSD framework [93] for generating genotype likelihoods especially for low coverage genome data is available. Users can select various variant filters by applying the GATK VariantFilter method [129] to filter called variants accordingly, or use VCF2Genome [162] to generate a genome sequence including the called variants. EAGER can be used to perform a full genotyping of a given sample using the method, including both available genotypers (the UnifiedGenotyper and the HaplotypeCaller) along with a variant filtration method to perform downstream analysis of called variants inside the pipeline. EAGER follows the GATK Best Practice's Guidelines[29], including INDEL realignment but excluding the Base Score Recalibration procedures. Following up on genotyping samples, the VCF2Genome method has been developed within the scope of Alexander Herbig's Ph.D. thesis [65] and later extended to be compatible with more recent versions of GATK (≥ 3.4). This method can be used to generate a consensus sequence of a provided VCF file produced with the UnifiedGenotyper or HaplotypeCaller. Various filtering requirements can be added to ensure that the generated consensus sequence fulfills certain criteria, for example, minimum coverage, the frequency of the observed allele at a specific position of the genome and a minimum genotype base quality.

Downstream statistics and reports

Assessing whether a sequencing experiment was successful and whether the obtained results are trustworthy is crucial for any project, not only on ancient DNA. This typically requires running several quality control applications that generate, e.g., coverage metrics for the investigated organisms to detect for example how much endogenous DNA is present in a sample. EAGER features several methods and tools for summarizing important attributes and analysis results after preprocessing, mapping and genotyping. An increasingly important part of aDNA analysis consists of the authentication of ancient samples. Typically this is addressed with damage pattern analysis and fragment length analysis. EAGER provides two methods, mapDamage 2.0 [77] and DamageProfiler [141](current version in preparation for publication[142]), to perform damage pattern analysis to authenticate samples. For detailed contamination assessment of mitochondrial data based on both damage patterns and haplotype profiles, the pipeline also integrates the method

Schmutzi [181], a framework that can be used to compute an improved endogenous human mitochondrial genome sequence by taking the estimated contamination values into account. For enabling cost-efficient screening analysis with EAGER, the Preseq [27] method is available in the pipeline. Researchers can utilize this method to estimate the library complexity in initial screening rounds. This is especially useful, for the decision process which samples have the potential to provide more DNA and should be sequenced deeper after initial DNA screening protocols. To summarize most of the mapping results in a single report, we furthermore included the tool QualiMap 2 [50, 146] in the EAGER pipeline, generating mapping, read length and other important metrics for a full analysis run. On top of this, the pipeline has an integrated ReportTable tool that can summarize several individual analysis runs in a single report, thus enabling direct comparable analysis overviews over large numbers of individual samples in an efficient way.

3.2.3 Graphical user interface

The EAGER pipeline features a graphical user interface (GUI) that can be used to generate configuration files in a simpler way than using a command line interface. During the initial design process of the interface, we decided to make the GUI as user-friendly as possible. This was particularly addressed by concealing more complex and advanced options as illustrated in Figure 3.3, while still preserving advanced functionality accessible for more advanced users. The GUI was developed using the Java programming language and the SWING widget toolkit, integrated in Java since version 1.2. The main interface was built using the IntelliJ IDEA development environment and the integrated GUI support tools. The choice for Java was motivated by the possibility to run the GUI on all operating systems.

In general, the GUI is split into three parts (Figure 3.3). First, the input section of the GUI provides the user with interactive buttons that can be used to select raw sequencing input files in FastQ format and a reference genome in FastA format. The selection of an output folder is also a requirement in this part of the GUI.

Next, the GUI offers the possibility to define computational constraints, such as the number of CPU cores the analysis is executed on and the amount of memory available for that. For cluster environments with specific constraints on temporary file sizes and storage, the GUI additionally allows to specify a special temporary file directory for intermediate results. The parameters selected in the GUI are applied to all subsequently selected analysis modules where applicable, ensuring a consistent usage of all computational resources.

Finally, the user can see the available modules in the EAGER pipeline in order of appearance in the analysis process. An interactive configuration of all analysis steps is possible here, as modules can be selected or deselected where desired. The pipeline offers well established parameters for typical analysis procedures in the field of aDNA research [56, 87], for example for raw data processing and read

mapping. However, users can define their own parameters in more detail by using the “Advanced” buttons next to each particular module to configure the respective module using their own set of parameters. Ultimately, the GUI provides an accessible way to generate the required configuration files in the last section. After all parts of the GUI have been filled out, users can click on the “Generate Config File” button to generate a configuration based on previously selected input files and analysis methods.

An important feature of the graphical interface is that it provides additional measures within itself to make analysis procedures as easy as possible for the user. This is achieved already on a low level by showing tooltips for all components in the GUI to provide the user with direct feedback on the meaning of these specific buttons, drop down menus and modules. Although this can not replace a more detailed user documentation, it enables especially first time users to get a quick overview of modules and configuration possibilities in the GUI. A less prominently advertised feature of the GUI is the possibility to run multiple samples in a single configuration, by selecting a folder as input with some samples with different identifiers in the file names. If the file naming patterns follow a certain pre-defined pattern as defined in the standard Illumina sequencing manuals, the folder structure of the input is automatically created in the selected output folder, making scalable analysis with even hundreds of different samples possible. When users want to rerun the same samples against multiple different reference genomes, the GUI furthermore supports to generate a configuration once and then selecting a different reference genome and output folder to create a second configuration for all the samples again, without having to configure everything else from scratch. To aid in such cases, the GUI marks the input fields that have not been changed in orange, while newly adjusted options are highlighted in green (Figure 3.4). These usability improvements can reduce the amount of time required for running an analysis procedure on a set of individual samples significantly. By for example running several samples against the same reference genome without having to re-select a reference genome, the GUI provides measures to achieve such an improved behavior. In addition to these features, the GUI automatically disables or enables modules upon certain constraints in some cases. For example, the “Quality Trimming” module is only selected, when neither of the two available adapter clipping and read merging methods are applied, as both of these support quality trimming of NGS reads, too. In such cases, this module would only increase the running time of the overall process unnecessarily. As some of the downstream analysis methods, for example, VCF2Genome, require a specific input format, the pipeline automatically configures itself in order to provide this specific required input format. Another example of this enhanced behavior of the application, is, that certain intermediate conversion steps, including reference genome indexing, generation of sequence dictionaries and file conversions e.g., from SAM to BAM format are done automatically for the user. This ensures both that the pipeline can run and that

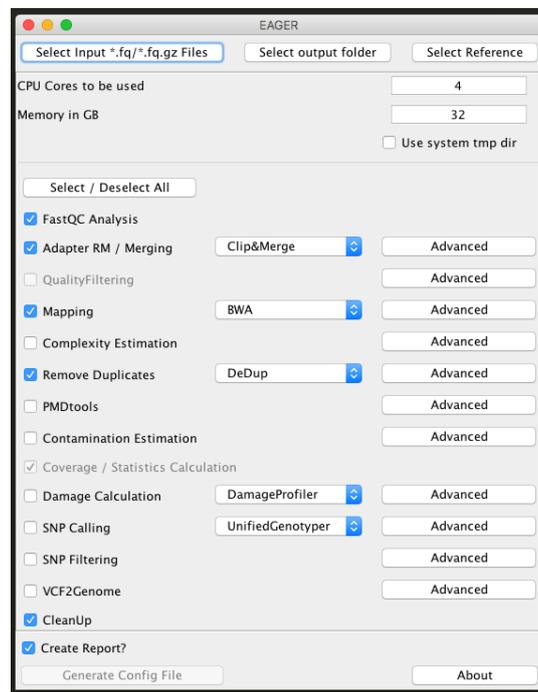


Figure 3.3: Main GUI of EAGER in Version 1.92.40 as of May, 3rd 2017. The GUI can be split into three sections: On top, the input section (red box). Next, the computational section where computational parameters for the execution can be set (blue box) and ultimately, the module section for selecting the required analysis modules in the pipeline (green box).

meaningful output is generated.

The GUI generates an extensible markup language (XML) based snapshot of a communication class, defined as a separate project “EAGER-lib” and containing all the access variables. The application runner “EAGER-CLI” can load this XML file later to reconstruct the communication object on pipeline execution, enabling a separation between configuration of the pipeline and actual execution. This is especially useful when local machines or head nodes on a cluster are used for configuration of a task and cluster submission systems, or larger computational resources are used for the actual pipeline execution.

3.2.4 Contributed tools and methods

Apart from methods already available, the EAGER pipeline introduced several new methods and tools for a successful analysis of aDNA. These methods complement or replace other published tools for preprocessing, mapping, PCR duplicate removal and other parts of the analysis procedure. One of these methods is Clip&Merge, developed by Günter Jäger [76, 162]. Clip&Merge can perform an efficient adapter



Figure 3.4: Exemplary usability features of EAGER: On top, the input dialog buttons before opening any files are shown. Once a user selected input FastQ files and a results folder, the respective buttons are marked in green to reflect a current change as shown at the bottom.

clipping of sequence reads, also integrating a method to simultaneously merge paired-end reads with negative insert sizes into a single collapsed read.

CircularMapper

Many of the available mapping methods are optimized for mapping NGS reads to a linear reference genome. Especially for bacterial genomes and mitochondria but also for some viruses and plants the respective genomes are circular. Modern mapping algorithms such as BWA [109] or Bowtie2 [102] try to map sequencing reads entirely against such a reference genome and mark reads that cannot be mapped in their entirety as unmapped. Even improved methods such as BWA-mem, utilizing modern concepts such as soft-clipping/masking reads, have not resolved the problem of incomplete reads mapping in entirety [108]. In the special case of circular genomes, this poses a problematic situation for reads ranging over both ends of the circular reference genome, as they cannot be mapped in their entirety to the reference genome. This can result in lower coverage for regions at both ends of the circular reference genome, thus hampering downstream analysis tasks such as haplotyping or genome reconstruction. To resolve these issues, the CircularMapper method has been developed in the scope of the EAGER project.

The basic idea of the method as illustrated in Figure 3.5, is to elongate a provided reference genome artificially by adding k bases at the start of the reference genome to its end. Subsequently, reads that would usually be marked as unmapped can now be mapped in their entirety, thus increasing coverage at the ends of the genome. The overall process to achieve this is split into two distinct parts: First, an elongated reference genome is synthetically created with `CircularGenerator`, by adding the first k bases of the reference genome to the end of the genome. Following up on this, all reads are mapped against the modified and elongated reference genome. The value of k is typically set to $k = 300$ (such that Illumina MiSeq reads of length 300 can be processed) but can be chosen freely, thus enabling adjustments for much longer reads (e.g., from other technologies). In general, the elongation value k should be chosen to be at least the maximal read length observed in the preprocessed sequencing dataset used as input. The second component of

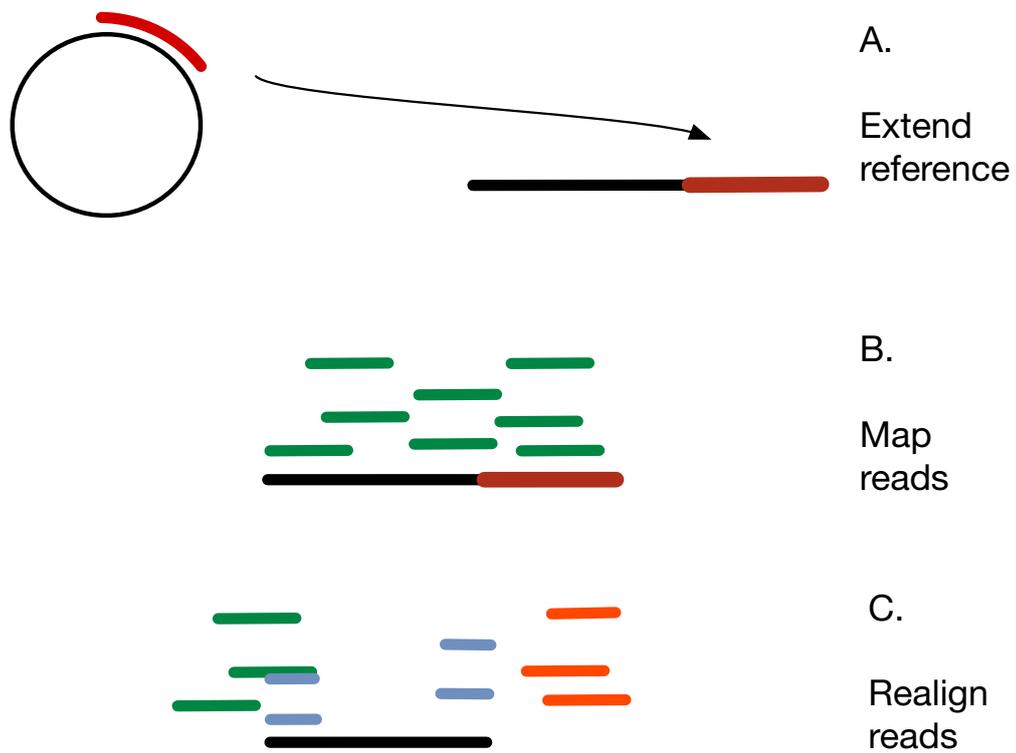


Figure 3.5: Conceptual idea of the CircularMapper method. A: The first k bases (red) of a circular reference genome (black) are extracted to extend the current reference genome. B: Sequencing reads (green) are mapped to the modified reference genome. C: Reads are categorized and realigned to fit the original unmodified reference genome.

CircularMapper `RealignSAMFile` distinguishes between three categories of reads. The first category of reads that maps to the interior region of the genome, thus these reads do not require any further processing. The second category of reads are the reads that fall either in the extension region or the first k bases region that was used to extend the genome. For these, the XA tag information in the mapped BAM file is used to match them with their secondary mapping position, and the reads are realigned to their correct origin in the first k bases. This elimination of a secondary mapping position, which is in fact not truly a secondary mapping position due to our synthetically modified reference genome, is more time efficient than remapping all reads mapping ambiguously and afterward combining the output with reads in categories one and two. The realignment is however important, as reads that are found to fit two or more regions on a reference genome are typically marked as ambiguous by the mapping algorithm. Since we can be sure that these result from the modifications introduced by CircularMapper, a repositioning within the first k bases of an unmodified reference genome is entirely correct for all reads that show exactly two alternative positions of high mapping quality.

Finally, reads in the third category have a starting position within the unmodified reference genome, and their end position is in the modified region. These are considered to be overlapping reads, spanning the circular overlap of the reference genome. `RealignSAMFile` splits these reads according to their overlap and places them in their correct positions.

For mammal genomes, where the mitochondrion is the only part of the genome to be organized as a circular chromosome, the method can perform this extension and split approach on a subset of chromosomes (e.g., just the mitochondrion). This ensures that only the mitochondrial reference is modified and modified and all other chromosomes are unaffected by the methodology presented here.

DeDup

As the PCR duplicate removal procedures available are unable to correctly remove duplicates from merged paired-end read data, we developed an improved duplicate removal method called DeDup. One of the most prominent issues of aDNA research is that the investigated samples typically show low amount of endogenous DNA. Therefore, enrichment and amplification techniques are usually applied to increase the number of DNA that can be retrieved from the given DNA fragments [202]. Unfortunately, this also increases the number of sequencing duplicates resulting from the same DNA fragments that are multiplied several times during for example PCR or more specific enrichment methods. As the coverage of a specific genomic locus is important for almost all downstream analysis types, the misled impression of high coverage at a specific locus by high duplication rates can communicate a false-positive trust in a specific region that might not be justified at all [162]. To regulate this, in-silico methods are employed to remove duplicated sequencing reads. For modern NGS data, several methods to achieve this have

been proposed, including `rmdup` in SAMTools [109] and `MarkDuplicates` in the Picard software set [129]. Both of these methods work well on regular paired-end sequencing data or single-end data, where the 3' of the forward and the 5' ends of the reverse reads are well known. However, since both of these tools only consider the 5' positions of merged paired-end reads, the intrinsic assumption regarding equal 3' ends fails for merged reads of aDNA, where resulting reads might have to be merged previously. As a result, both `rmdup` and `MarkDuplicates` remove reads that stem from different fragments and start at the same 5' position. To compensate this, the new DeDup method implements an idea described by Green *et al.* [56], in which both the 5' and 3' ends of the reads are taken into account for duplicate removal. For reads that have exactly the same 5' start and 3' end

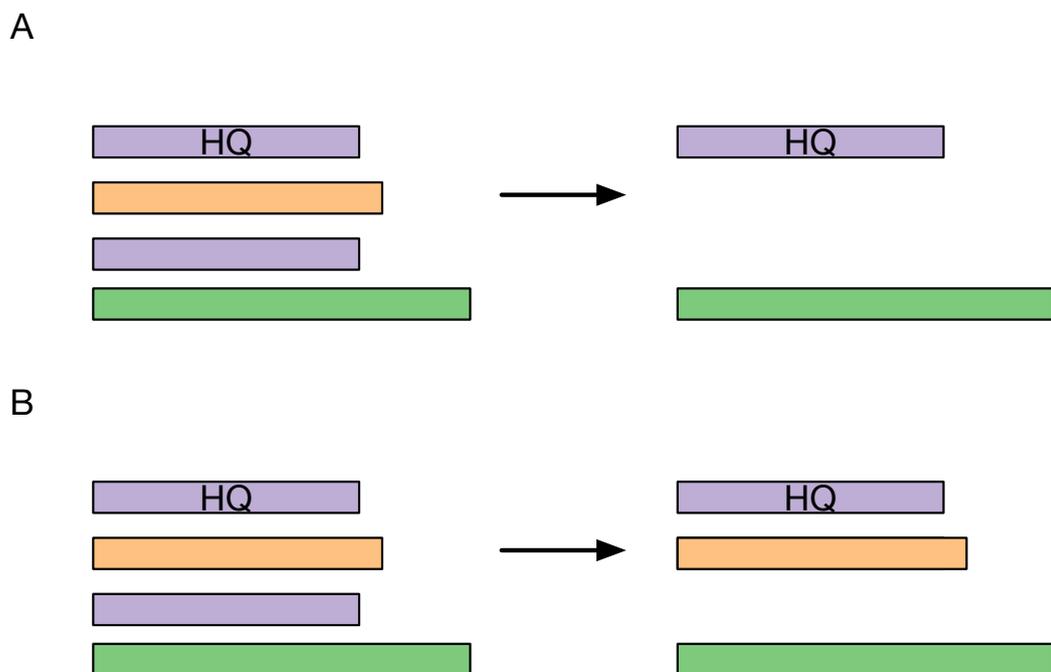


Figure 3.6: Duplicate removal strategies of SAMTools `rmdup` (A) and DeDup (B). Violet reads represents true duplicates, with one higher quality (HQ) read that does not get removed in both cases. Longer reads with the same starting position (green) are getting removed by SAMTools if their quality is inferior to a read with the same 3' starting position. Assuming that all these reads are merged paired-end (PE) reads, the DeDup application is taking both 3' and 5' positions of the respective reads into account and can keep three of four reads, while truly solely removing PCR duplicates in such a case.

positions, the read with the higher sum of base qualities is kept. Reads that lost their partner, for example due to being too short after sequencing adapter removal are treated as single end reads respectively. As an improvement over other duplicate removal tools, the DeDup tool also produces more detailed output statistics

and a duplication count histogram that can be subsequently used for estimating the library complexity of the investigated sample efficiently. This reduces further the amount of time required for an entire analysis run, as other methods require to read whole BAM files in order to estimate the library complexity, which can largely affect runtimes on especially larger samples.

An intrinsic difficulty in the applied approach is the problematic handling of read starting and end positions within the SAM and BAM format. For forward reads, the 3' starting position is defined, whereas for reverse reads only the 5' starting position is defined, which is a result of adapter clipping and quality trimming, resulting in different lengths of reads, stemming from the same DNA fragment. Only for merged reads, both positions are well defined and can be trusted entirely. Unfortunately, this increases the complexity of duplicate removal significantly, as shown in Figure 3.7. If a BAM file contains solely merged reads or forward reads, the read handling is trivial and can be done by traversing the sorted BAM file on position. If the file however contains reverse reads, this approach is incapable to detect duplicates between merged reads and reverse reads, as start position of a reverse read is traversed later in such a case. Therefore, DeDup handles data with a Java PriorityQueue and flushes this queue regularly, making the handling of such "overlapping" cases possible. Additionally, this would allow to enable DeDup to handle paired-end read data as well.

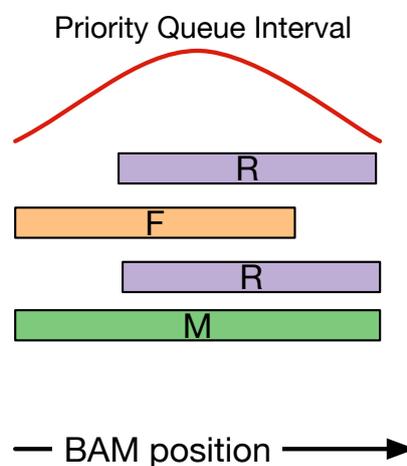


Figure 3.7: Technical implementation details of the DeDup application. As illustrated here, the duplicate removal procedure cannot easily read the file with an implemented file reader, as potential duplicates of the same DNA fragment might still be present in the file. Thus, a Priority Queue data structure is used, that caches all forward reads (F) with a defined 5' starting position and all reverse reads (R) with a defined 3' starting position, comparing them with merged reads (M) for which both start and end positions are defined. This is illustrated with the priority queue interval, which is subsequently traversed once the end of a merged read is reached, to resolve duplicates properly.

Other tools

To aid users in their analysis, interpretation and other parts of pipeline interaction, several smaller tools were developed within the scope of EAGER. As these are mostly smaller contributions, for example, some plotting scripts, additions to provide users with some feedback or reduce technical efforts, these are just listed here for the sake of completeness.

DamageProfiler

A crucial step in aDNA research is the authentication of sequenced samples to be indeed of ancient origin. Methods such as mapDamage2 [77] were typically used for this purpose in EAGER. As mapDamage, unfortunately, requires a large set of R dependencies installed in particular versions, work on DamageProfiler was started by Judith Neukamm in 2015 to produce a more consistent replacement tool [141, 142]. DamageProfiler is capable of creating damage pattern profiles, average fragment histograms and all of the typically required authenticity analysis types for aDNA, including a simple GUI for easier interaction as shown in Figure 3.8. Additionally, the method does not require any other tools or methods to be installed and is by several orders of magnitude faster than mapDamage, especially on larger input datasets and thus integrated as a replacement method for mapDamage 2 in the EAGER pipeline.

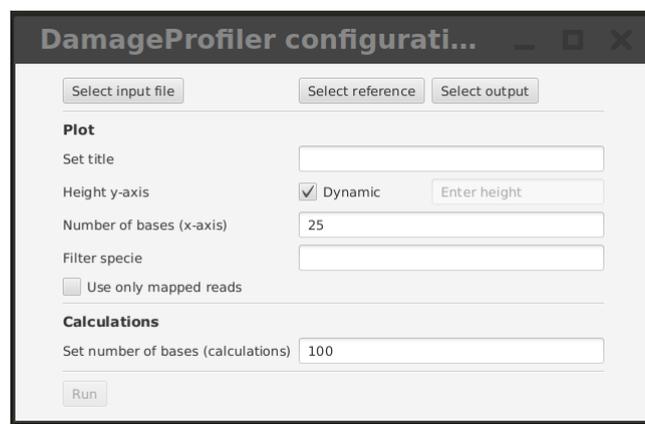


Figure 3.8: GUI of DamageProfiler. Users can configure various parameters in a plain interface, also selecting the input BAM file, reference genome and selecting an output folder. When executed directly with parameters, the application can be used in a command line mode (not shown here).

Library complexity plotter (LCP)

Determining the complexity of a given sequencing library is crucial for the cost-efficient analysis of genomes. Preseq [27] is currently accepted as the state of the

art tool for estimating library complexity of a sequencing library. While the application generates a count table of observed library complexity for given DNA libraries and can also optionally create an extrapolation of expected library complexity, the method itself does not enable researchers to visualize their results directly. Within EAGER, which is often used for such screening purposes on aDNA libraries, this motivated the development of the Library Complexity Plotter (LCP) application (Figure 3.9). LCP is a small Java application that can plot these count and estimate tables to provide the user with a more intuitive way of estimating the library complexity for such an experiment.

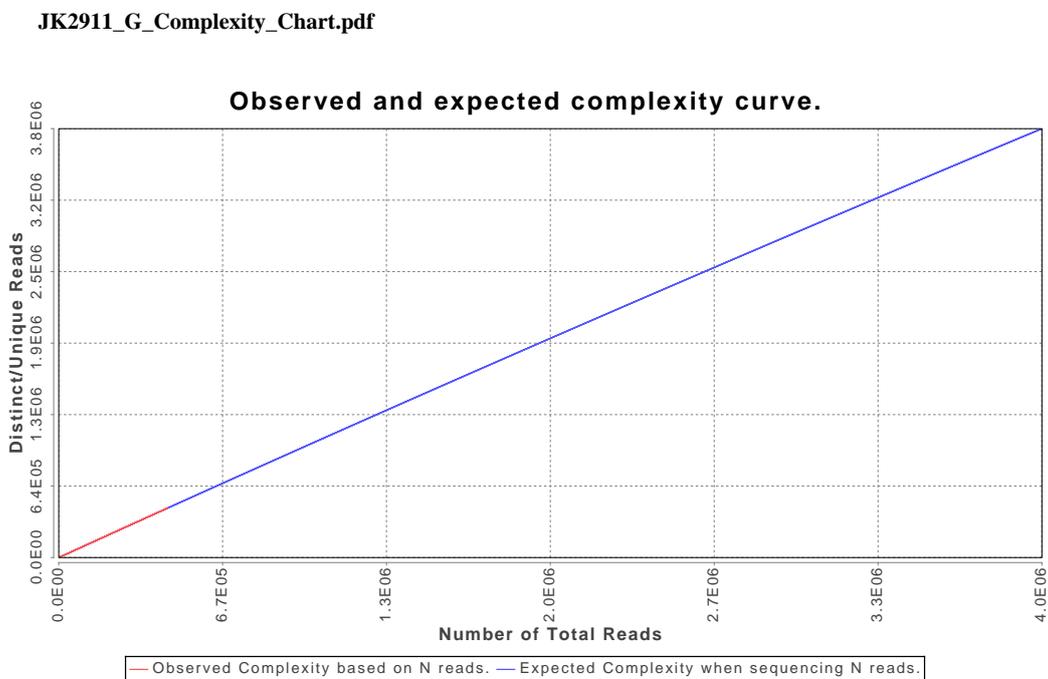


Figure 3.9: Plot showing the observed and expected library complexity of the sample JK2911 after initial whole genome screening. The X-axis displays the total number of sequenced reads. On the Y-axis, the total number of unique reads (e.g., reads that have not been seen before and provide novel information on the genetic composition of the investigated sample). The red line shows the number of reads truly observed on which basis the estimation (marked in blue) is performed in Preseq. In this exemplary case, a linear growth can be expected without reaching saturation of the library when sequencing 4.0×10^6 reads. In many other cases, a logarithmic behavior can be observed with such a plot, showing a saturation when almost no further sequenced reads provide new information content.

BWA mismatches calculator

Choosing appropriate mapping parameters is crucial not only in ancient DNA analysis. Increasing, for example, the number of expected mismatches in a read is a common parameter adjustment typically done in aDNA projects to account for misincorporations at the ends of reads. While some methods such as Bowtie2 [102] provide an intuitive way to choose such an “allowed misincorporations per read” parameter, other mappers such as the more common BWA `aln` do not allow that. Unfortunately, the parameter description is not even documented in the BWA documentation or man pages, so users are completely left with figuring out what the parameter represents on their own¹. To aid users with the choice of mapping parameters, a Shiny web application [24] has been developed in this thesis, helping users to choose the right kind of mismatch parameter specifically for their datasets. Users can interactively set the average read length they obtained on their datasets in initial quality checking and set the n parameter of BWA accordingly in the interface². An example is shown in Figure 3.10. A simple graph then visualizes the allowed number of mismatches for the analysis, which is more intuitive than the λ parameter³ set as n in BWA. Furthermore, a benefit from having the BWA-Mismatches app available in the web is that the application is directly accessible with the general EAGER documentation and user tutorials and can be accessed by everyone with a modern web browser.

VCF2Genome

The VCF2Genome application for reconstructing a consensus genome from a provided VCF file has been published already by Alexander Herbig [65] and was subsequently published in the EAGER paper [162]. Lately, some upstream technologies have been changed, requiring significant code basis changes in VCF2Genome. With updated versions of GATK (> 3.7 , available since 2017), VCF specification format 4.2 is now widely used by GATK, and the former version of VCF2Genome was unable to parse these files. As a small contribution to the method described in Alexander Herbig’s thesis, the old version was ported to utilize the standardized `htsjdk` development library [110], to enable the application to read current VCF 4.2 format VCF files again. Additionally, the entire source code has been modularized, subsequently improving the possibilities for changes and thus maintenance of the application, while preserving identical results when applied to older VCF files that have been generated for example two years ago. The accordance of obtained results to older versions of VCF2Genome is guaranteed by a set of software tests with previously published data.

¹See for example <https://www.biostars.org/p/16221/>, accessed October 20, 2017

²<https://apeltzer.shinyapps.io/BWAmismatches>

³<http://bio-bwa.sourceforge.net/bwa.shtml>

BWA Mismatch Rate estimator

Info

This is a small tool to assist you in choosing an appropriate mapping parameter ($-n$) for the BWA mapping algorithm (bwa aln). Please specify both your average read length and additionally your selected relative mismatch rate parameter. You will then see a curve for a range of 1-10 mismatches the likelihood of a read with n mismatches occurring by chance. If this is lower than your set threshold, the read gets discarded, if its higher, the read is kept. Example: For $l=100$ and $n=0.04$, we keep reads with up to 4 mismatches and discard reads with more than that.

Your chosen parameter $-n$ in BWA aln.

0.04

Average Read Length

100

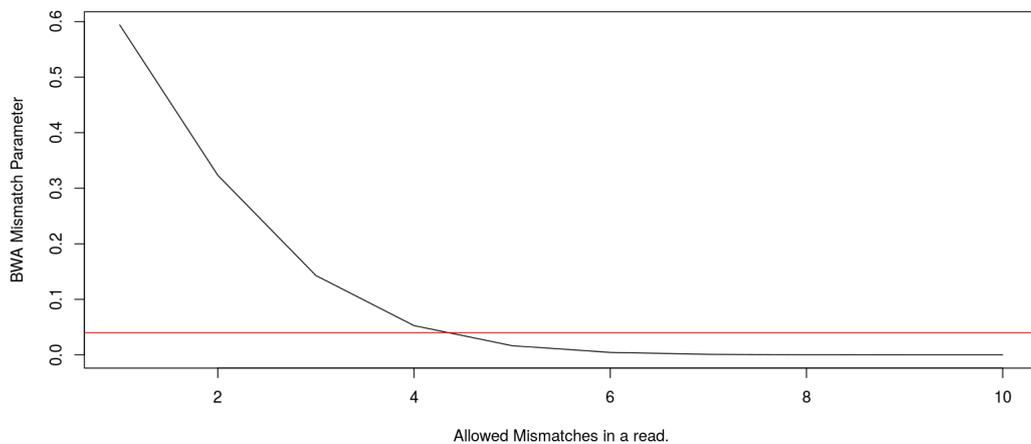


Figure 3.10: User interface of the Shiny Web application for determining an optimal $-n$ parameter for BWA. The user interface is kept simple with an informative textbox and two textboxes to enter the respectively chosen parameter for the $-n$ parameter in BWA, and the observed average read length in the investigated sample. A simple plot then informs the user, how many mismatches are allowed for a read of the chosen length with the parameter $-n$.

3.2.5 Testing & deployment

Writing software for complex analysis procedures as in the EAGER project requires extensive testing of the utilized methods and tools to ensure consistency and reproducibility. Complex pipelines that integrate multiple foreign methods and add their additional tools and methods cannot be tested solely by applying simple Unit tests. Unit tests are typically applied to single tools and are used to test the outcome of a single method or function within a single application. Therefore, developers rely on continuous integration frameworks, such as Jenkins [79] or Travis CI [225]. Within the scope of modern software development, continuous integration means that the framework automatically (re-)builds the software during the development process and applies *integration tests* to ensure consistency of the introduced changes. Here, all parts of a complex application are tested in a real environment unlike Unit tests, which are typically synthetically testing a method behavior without integrating dependencies. As shown in Figure 3.11, the integration process within

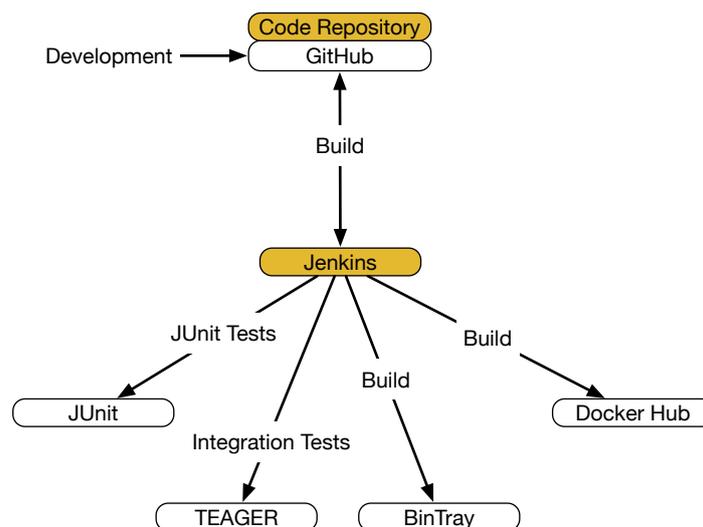


Figure 3.11: Diagram showing the development process of the EAGER pipeline. Created code within the development process is stored in a Git Repository on GitHub, which automatically triggers automated integration tests and testing of utilized tools and methods. If this is successful, a work version of the current executable is uploaded to Bintray and made available to all users. Furthermore, a Docker image is built upon successful testing and made publicly available on Docker Hub for users, too.

the development process of the EAGER pipeline is kept fairly simple. Once the code is written and checked into the central code repository on GitHub, the continuous integration framework Jenkins automatically triggers code tests, integration tests, and the software deployment. Within the scope of this thesis, Unit tests for the tools Clip&Merge, DeDup, CircularMapper, and VCF2Genome have been created. These have proven to be useful in the development process. Testing the real behavior of the pipeline cannot be tested using this approach, which is why

the TEAGER project was started, to test the actual pipeline execution. TEAGER is capable to run consistency checks on a selected set of golden standard analysis datasets. This allows checking for inconsistencies when running the golden standard analysis datasets with a newer or updated version of the EAGER pipeline. Results created by the pipeline are summarized and compared to expected results from previous (manual) pipeline executions. The developer subsequently receives a short report with potential causes for inconsistencies found by the TEAGER framework. To keep these integration tests completely separate from any production environment, the continuous integration framework Jenkins automatically sets up an encapsulated testing environment for these tests using a Docker container. Ultimately, the docker container containing the most current software versions runs all integration tests separated from any other running jobs on the productive system, thus ensuring that currently running other analysis procedures are unaffected by software development. Whenever a component of EAGER is updated, the automated integration tests are started again from scratch, ensuring consistency of the whole pipeline at all times.

In general, the EAGER pipeline and all contributed tools are built with the Gradle (<https://gradle.org/>) build tool. Upon successful testing, resulting binaries are then uploaded to Bintray with a version tag. This enables users also to downgrade EAGER versions, e.g., to rerun an analysis that has been created earlier for publication. Relying on Gradle ensures, that software libraries used in the development process are also kept up to date. Without automatic software building tools such as Gradle or Maven, this would require manual interaction during the development process, making the whole process more error prone. Furthermore, this ensures that bugfixes in upstream projects (e.g. the htsjdk library [110] for reading and writing SAM/BAM output) are automatically integrated in the pipeline during the developmental process, without further interaction required.

3.3 Results

The EAGER pipeline is implemented to enable an efficient analysis of aDNA data both for small labs and large research institutes. In order to achieve this, the pipeline executes single modular processes iteratively after each other. If single subprocesses allow for parallelization, these utilize the maximum number of CPUs and memory that were specified by the user during the configuration in the GUI. This ensures that processes do not directly deadlock each other by consuming more resources than the system provides. Apart from this, the pipeline uses efficient and typically compressed storage formats, to keep both CPU and disk storage usage as low as possible.

Evaluation

EAGER has been widely adopted in the field of aDNA research and has received more than 35 citations⁴ in the first two years after publication. Although this certainly substantiates the applicability of the method itself, the EAGER pipeline has been applied to several datasets from various projects to provide a more detailed overview of the pipeline and its possibilities within the scope of this thesis. For the majority of aDNA projects, the typical project scope can be summarized as shown in Figure 3.12. In most cases, an initial whole genome shotgun screening is done after laboratory work to determine characteristics such as endogenous DNA content and library complexity of the investigated samples. After this initial decision process, samples are either sent for deeper whole genome shotgun sequencing or enriched for a specific target, e.g. a mitochondrial capture, a nuclear capture (390K, 1240K [58, 106]) or a specific pathogen capture [11, 202]. For all of these use cases, EAGER provides methods, automated procedures and reporting functions to enable a successful analysis. In the following sections, several different types of data were analyzed using EAGER to elucidate various aDNA analysis projects in their entirety. All evaluations have been performed on an Intel®Xeon®CPU E5-2698 v3 @ 2.30GHz machine with 64 CPU cores and 356 GB of RAM. Analysis results were stored on a 48 TB storage system accessible via a 10 GBit/s network connection, capable of disk I/O speed with 550 MB/s both read and write. The runtime evaluation was performed with the integrated timer function of the EAGER pipeline, which automatically saves the execution time in the logfile of an analysis run on a per-sample basis. An important aspect when interpreting the resulting runtimes is, that not all of the methods are parallelized and therefore do not benefit in the same way of more CPU cores. Even in theory, some of the methods would not benefit from a parallelization of the algorithms in the background, e.g. when a sequential process is the only available way of running an analysis. Examples of such behaviour are the DeDup, DamageProfiler and VCF2Genome modules. While the performance of PALEOMIX [198] was evaluated in 2016 in the initial EAGER publication [162], we decided to not perform such an evaluation with both methods within this thesis. First, the methods and tools utilized in both pipelines are identical, except for independently developed tools such as DeDup or CircularMapper. An evaluation of both pipelines would merely show the differences these methods generate. Thus, this thesis focuses on evaluating the contributed methods that are unique and compares their performance with several metrics against direct competitor tools. Apart from this, the feature set of EAGER and PALEOMIX is described in detail in an additional evaluation, with a focus on usability, reproducibility and other metrics for scientific software, as defined by Koschmieder *et al.* [94].

⁴<http://bit.ly/2xSvwsU>, accessed January 30th 2018

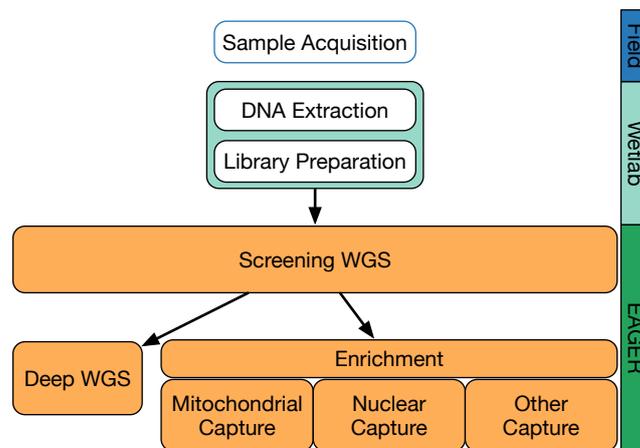


Figure 3.12: A general project workflow in aDNA projects. The sample acquisition is marked in blue and typically done, e.g., by archaeologists. The following wet lab work (here light green) includes DNA extraction and DNA library preparation procedures in a clean room facility. Coloured in green are NGS sequencing and enrichment methods whose output can be analyzed using the EAGER pipeline, covering all major aspects of aDNA NGS analysis.

DeDup

The DeDup application to remove PCR duplicates after read mapping was evaluated to determine differences to the standard SAMtools `rmdup` application. Both applications were evaluated on a dataset of 110 Syphilis genomes from Pinto *et al.* [167] and Arora *et al.* [6]. The comparison was performed by using DeDup as duplicate removal procedure, and coverage was calculated with SAMtools `depth` on all samples.

Following up on this, SAMtools `rmdup` was run on the same input with default settings. The average coverage obtained over all samples was 56.34 X for `rmdup`, while DeDup was able to achieve an average coverage of 110.77 X on the samples (see Table 3.2). Especially on high coverage samples, the coverage values for DeDup were higher than with SAMtools `rmdup` as can be seen in Figure 3.13. The full set of results can be seen in Supplementary Table A.2. An explanation for this behavior is that the likelihood to observe a duplicate with similar starting position is increasingly more likely with higher coverages. Thus `rmdup` will remove more reads in high-coverage than in low-coverage samples. However, there are borderline cases, where DeDup still increases the coverage. As can be seen in Table 3.1, the samples C34 and S7 are below a threshold of 1 X if the `rmdup` method is used, while DeDup retains enough reads to achieve > 1 X on these samples. Thus, the application of DeDup could enable researchers to keep the respective samples in their analysis. The application of DeDup could enable researchers to keep the respective sample in their analysis.

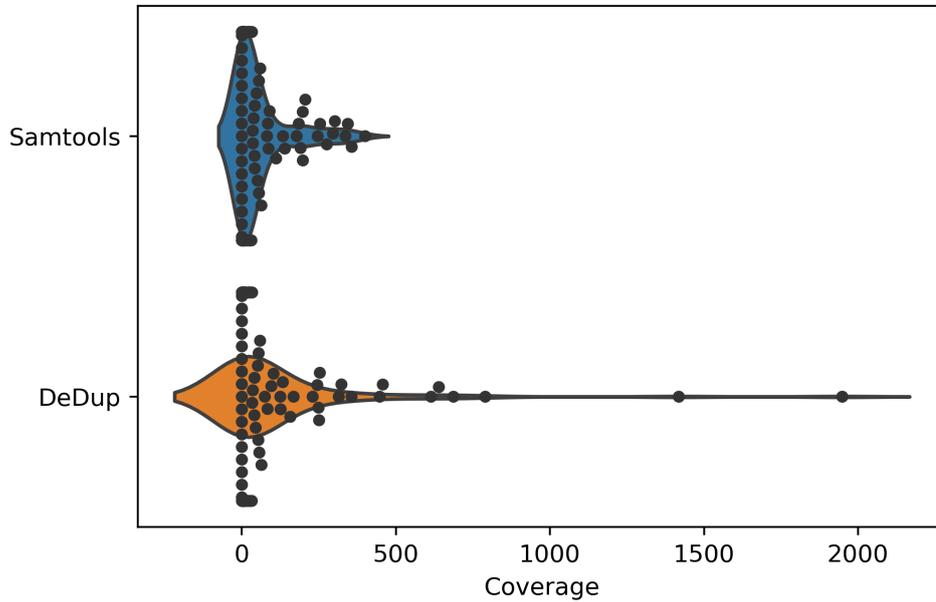


Figure 3.13: Violinplot comparing coverages (in X , on X -axis) on all investigated samples of the Syphilis project [6, 167]. The orange violin shows the coverages achieved with the DeDup application. The blue violin shows the coverages that the SAMToolsrmdup application achieved. For further details see the Supplementary Table A.2.

Table 3.1: DeDup and SAMTools rmdup performance on three borderline cases C34, SRR3584838 and S7.

Sample	Coverage (in X)	
	rmdup	DeDup
C34	0.44	1.15
S7	0.87	1.63

Table 3.2: DeDup and SAMTools rmdup performance on 110 Syphilis samples. Shown are mean and median coverage (in X) obtained by the methods.

	SAMtools rmdup	DeDup
Mean Coverage (X)	56.34	110.77
Median Coverage (X)	8.82	11.12

CircularMapper

The CircularMapper method was evaluated on a randomly chosen dataset of 52 Ancient Egyptian mummy mitochondrial captures from Schuenemann *et al.* [201]. The EAGER pipeline was executed on all 52 samples with the CircularMapper option, and the generated BAM files were analyzed with Bedtools [172] to obtain a coverage histogram. Similarly, the EAGER pipeline was used with the BWA option to generate BAM files without the circular improvement procedure, and the coverage histogram was analyzed with Bedtools to compare these histograms. A coverage graph of an exemplary sample (JK2869) was created using the R Bioconductor package *gviz* [59], to investigate the differences in coverage on the exemplary sample as shown in Figure 3.14. The coverages in the start and ends of the circular reference genome are notably increased, specifically for the first 10-15 BP. Additionally, the mean coverages on the first and last 50 bases of each sample were computed using SAMTools *depth* as shown in Supplementary Table A.3. As can be seen, the achieved coverages with CircularMapper are in all cases higher than the ones obtained with BWA, as illustrated in Figure 3.15, for example.

Computed over all 52 samples, the runtime of the CircularMapper approach added an average of 2.1% additional mapping runtime for each sample.

Runtime evaluation

Pathogen capture analysis To elucidate the runtime performance of EAGER in the context of a pathogen analysis project, a dataset of 110 Syphilis samples published in Arora *et al.* [6] and Pinto *et al.* [167] was analyzed using EAGER as in Arora *et al.* publication. The pipeline was configured to run a default quality check, the mapping with BWA-mem and genome consensus reconstruction with VCF2Genome on all samples. BWA-mem was evaluated within our Syphilis project [6] to provide increased performance with respect to BWA. The overall process is shown in Figure 3.16. The samples in the analysis procedure range from a minimum of 30 MB (Megabytes) up to a maximum of 8,500 MB with an average sample size of 1,221 MB. In general, the analysis of a single Syphilis took on average 12.7 minutes (762 seconds), including all individual modules (see Table 3.3). As a general trend, the mean performance of the pipeline runtime is driven by the AdapterRemoval and IndelRealignment modules, while typical analysis metrics generation modules such as SAMTools *Flagstat*, *Index* and *DamageProfiler* only need a much lower fraction of the overall runtime of the analysis procedure. The integration of data from a different publication [167] can be achieved by downloading the respective RAW sequencing files and providing these to EAGER with the same settings as used before for our in-house samples.

3.3. Results

Table 3.3: Runtime of modules within the EAGER pipeline on a dataset of $n = 110$ Syphilis DNA samples. Listed are the minimum, maximum, mean and the median runtime of a selected module on all investigated samples along with the standard deviation. Due to the variance in input size of the analyzed samples, the high standard deviations within modules are expected.

Module name	Runtime in seconds				
	min	max	mean	median	std. dev.
FastQC	12	587	118	71.5	110.69
AdapterRemoval	8	2,105	373.9	220	411
BWA MEM	1	189	40.54	31	34.72
SAMTools View	0	137	20.93	13	24.54
Picard RG	3	479	76.78	46	88.68
SAMTools Sort	0	175	18.52	10.5	25.49
SAMTools Flagstat	0	33	5.18	3	6.24
DamageProfiler	1	290	22.22	10	38.86
DeDup	2	5,922	157	41	647.49
SAMTools Index	0	31	4.69	3	5.46
QualiMap	4	102	20.91	13	18.66
GATK IndelRealigner	5	756	97.14	61	118.879
GATK UnifiedGenotyper	11	37	16.725	17	3.7
VCF2Genome	3	5	4.02	4	0.62
ReportGenerator	1	29	4.41	3	5.09

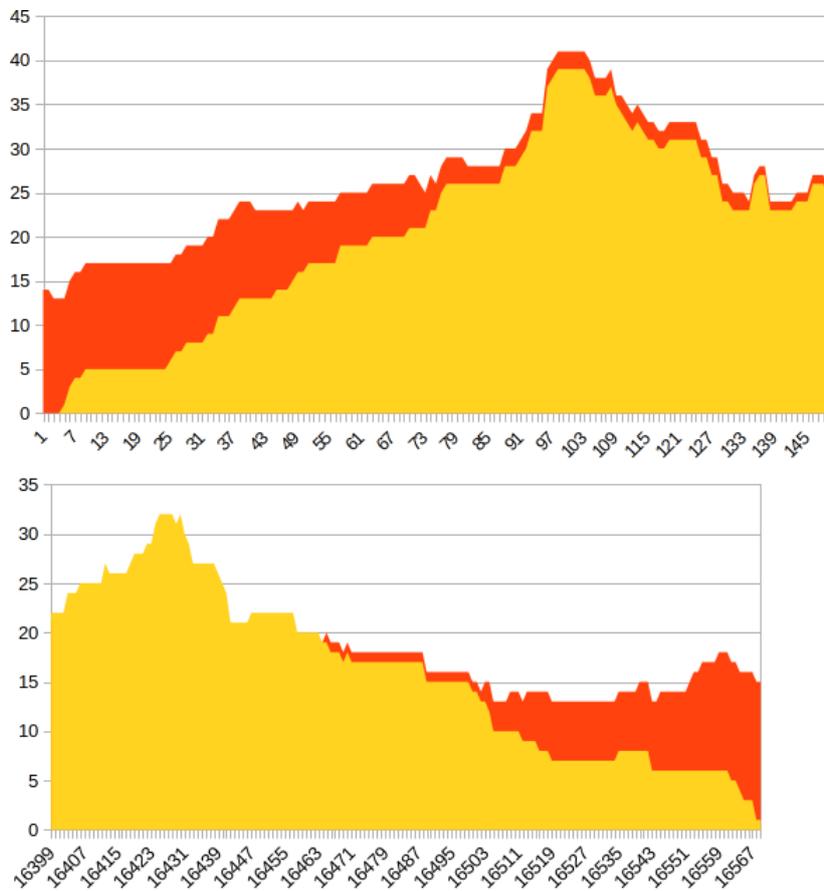


Figure 3.14: Coverage plot of BWA (orange) and CircularMapper (red) on the mitochondrial capture sample JK2869 [201]. Note that the differences in coverage in the selected region are due to the realignment process of CircularMapper, which reduces the coverage obtained in some of the elongated areas, as the reads are realigned around the start and end position of the circular genome. On top, the coverage (in X) at the first 145 bases is shown. At the bottom, the coverage (in X) on positions 16,399 to 16,569 of the mitochondrial genome is shown. As can be seen, the CircularMapper approach achieves superior performance at specifically the ends of the mitochondrial genome while behaving identically to BWA at non-end positions of the respective genome.

Mitochondrial capture analysis To test the runtime performance of EAGER within the scope of mtDNA analysis, a dataset of 151 mitochondrial capture libraries, as published in Schuenemann *et al.* [201], was analyzed using EAGER. The pipeline was configured to run the required initial QC modules (FastQC, Adapter-Removal) and subsequent mapping modules with CircularMapper for improved genome reconstruction, authentication (DamageProfiler) and contamination estimation (schmutzi) methods to reconstruct mitochondrial consensus sequences from the investigated samples (see Figure 3.17). Again, QualiMap was used to as-

3.3. Results

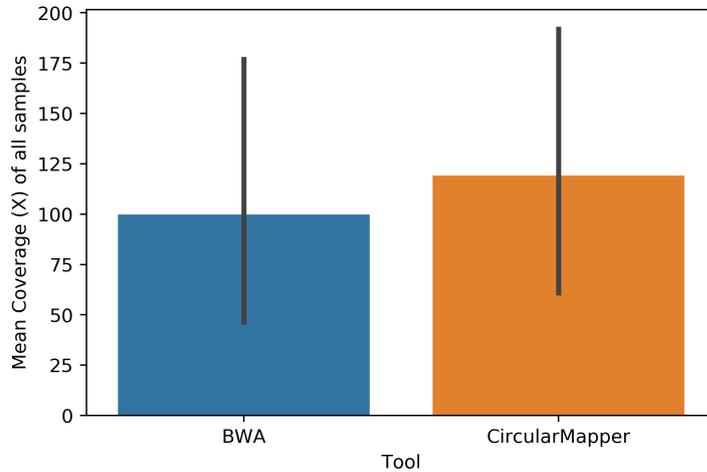


Figure 3.15: Mean coverages obtained on all 52 samples with CircularMapper or BWA. The shown mean values were computed on the first and last 50 bases of each sample using the Python Pandas and Seaborn libraries. As can be seen, the CircularMapper achieves higher mean coverages on all 52 samples than BWA.

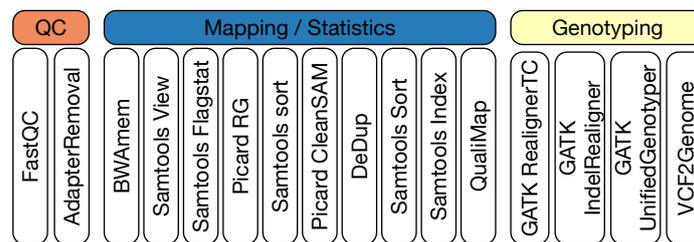


Figure 3.16: Pathogen analysis workflow as used in the Syphilis project [6]. The process can be split into three distinct components, initial quality control (QC, in orange), mapping and statistics (in dark blue) and genotyping (in yellow). The quality control contains the FastQC module and Adapter Removal tools. The mapping and statistics modules contain BWA-mem for read mapping, several SAMtools modules for data conversion and filtering, Picard RG for adding Read group information, DeDup for PCR duplicate removal and QualiMap for generating mapping statistics. In the genotyping block, the GATK modules generate an improved IndelRealigner VCF file that is the subsequently filtered with VCF2Genome to generate a consensus FastA file for downstream analysis after EAGER.

sess basic statistics such as coverage in this analysis. The minimum raw data input size was 22 MB up to a maximum of 3,700 MB, with an average of 453 MB for a sample. Qualitatively, this analysis was performed identically within the Ancient Egyptian mummy project to produce 90 mitochondrial genomes of ancient Egyptian specimen [201]. Therefore this evaluation solely consists of a performance evaluation regarding runtime, demonstrating the applicability of the EAGER pipeline on

mitochondrial DNA data.

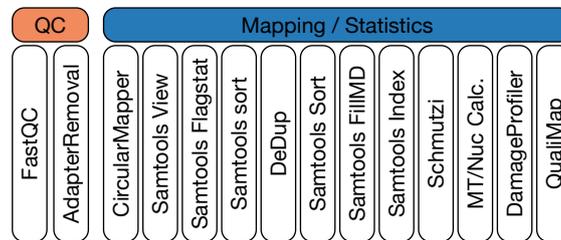


Figure 3.17: MTDNA analysis workflow as used in the Egyptian Mummy project [201]. Here, the workflow can be split into two distinct components, initial quality control (QC, in orange) and the mapping and statistics (dark blue) parts. To improve the results on circular genomes, the CircularMapper module was used along with several SAMtools modules for data conversion and filtering. To assess contamination of the sequenced libraries with foreign mitochondrial profiles, the schmutzi application module was used in this analysis, too. Duplicates were removed with DeDup, and QualiMap was used for generating mapping statistics.

Deep whole-genome sequencing analysis As we have seen, EAGER can be used for bacterial genome reconstruction and mitochondrial genome reconstruction both in an ancient and modern context. As a further test case for a general applicability of the EAGER pipeline in large-scale data projects, a whole genome sequencing project with a total raw input data size of 221 GB (Gigabytes) was processed, following the workflow in Figure 3.18. The pipeline was configured to run initial quality control methods without adapter clipping, as there were no adapter sequences present anymore. Mapping sequencing reads to the reference genome (hg19) was performed using BWA aln / samse approach, following already published recommendations [56, 106]. Additionally PCR duplicate removal was performed with DeDup and mapping statistics were created with QualiMap. To authenticate the sample to be of ancient origin, the DamageProfiler application was used within EAGER to generate damage profiles. Similar to other analysis standard operating procedures, the QualiMap application was utilized to generate basic output statistics for the reconstructed genomic data. In total, a 25X coverage human genome was reconstructed from the selected raw sequencing input data. While most aDNA sequencing projects typically produce far smaller project sizes, EAGER is capable of dealing even with datasets ranging into the hundreds of gigabytes in raw data size. This exemplary evaluation, therefore, demonstrates that EAGER can be applied in small and large-scale sequencing projects in general while providing the same state of the art functionality and usability to all kinds of users.

3.3. Results

Table 3.4: Runtime of modules within the EAGER pipeline on a dataset of $n = 151$ mtDNA captured DNA samples. Listed are the minimum, maximum, mean and the median runtime of a selected module on all investigated sample along with the standard deviation. Due to the variance in input size of the analyzed samples, the high standard deviations within modules are expected.

Modulename	Runtime in seconds				
	min	max	mean	median	std. dev.
AdapterRemoval	0	700	75.3	54.5	103.34
BWA Align	0	64	6.96	5	9.67
BWA Samse	0	107	11.255	7	16.2
CircularMapper	0	187	19.22	13	27.52
CleanSAM	0	95	13.3	10	14.56
ContaminationEstimator (DMG)	2	48	10.42	8	9.45
ContaminationEstimatorMTdefault	188	6,829	925	372	1,339.14
DamageProfiler	1	12	4.23	4	2.02
DeDup	0	1,264	46.86	11	180.35
FastQC	9	336	58.19	48	48.69
MT/Nuc Ratio Calculation	1	9	3.13	3	1.79
QualiMap	4	13	6.89	7	2.17
ReportGenerator	2	47	8.93	5	9.02
SAMTools Flagstat	0	50	4.75	3	7.45
SAMTools Fillmd	0	49	10.344	9	9.029
SAMTools Index	0	9	0.93	1	1.4
SAMTools Sort	0	54	4.93	3	8.55
SAMTools View	0	20	1.65	1	2.87

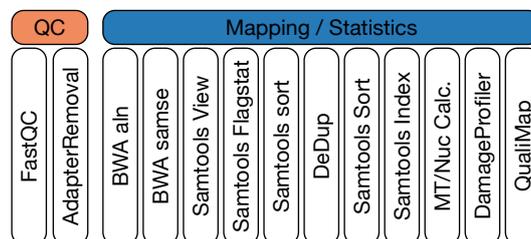


Figure 3.18: Exemplary workflow for a whole genome deep shotgun sequencing experiment. The workflow can be described as two distinct components, the initial quality control (QC, in orange) and the mapping and statistics part (dark blue).

Table 3.5: Deep WGS runtimes of EAGER modules

Modulename	Runtime in seconds
BWA Align	49,500
BWA Samse	540,508
DamageProfiler	18,714
DeDup	21,262
FastQCdefault	17,025
MT/Nuc Calculator	1,660
QualiMap	2,916
ReportGenerator	4
SAMTools Flagstat	896
SAMTools Index	918
SAMTools Sort	4,212
SAMTools View	9,446

3.4 Feature comparison of EAGER and PALEOMIX

As the majority of tools and methods in EAGER and PALEOMIX are identical, specifically with respect to offered mapping algorithms for most use cases, the evaluation of EAGER and PALEOMIX has been done on a comparative level. We followed established criteria by Koschmieder *et al.* [94] and generated a set of additional categories under which both pipelines were compared (Table 3.6). As Koschmieder *et al.* focused on Microarray analysis methods, we furthermore added more specific categories for aDNA analysis. Within each of the proposed main categories, we evaluated a number of important features to inform users on what kind of performance can be expected when using one of the two available pipelines respectively.

The general results in Table 3.7 show the results of a detailed evaluation of EAGER and PALEOMIX. The criteria used in the evaluation can be found in a more detailed form in Supplementary Table A.4. Tools that only implement specific functionality (e.g. ATLAS[112]) were not considered in the evaluation.

3.5 Availability & requirements

The EAGER pipeline and all tools and methods created within the scope of the project are available in various types. We were able to run the pipeline successfully on Linux workstations with four CPU cores and 4-8GB of RAM already, making the pipeline accessible even for small research groups with limited resources. As

3.5. Availability & requirements

Table 3.6: Basic pipeline evaluation criteria for EAGER and PALEOMIX. A more detailed explanation of all criteria head nodes in Supplementary Table A.4. Shown are the evaluation category, a short overview description and the number of criteria within the respective category.

Category	Description	Criteria
Basics	Basic information (project homepage, ...)	8
System Properties	Type of Tool, Availability ...	8
Standards	Supported Standards	2
Pre-processing	Available pre-processing methods	2
Analysis	Available analysis methods (mapping, duplicate removal, quality metrics)	7
Software testing	Continous integration testing of pipeline	1
Reproducibility	Support for methods improving reproducibility (e.g. containers)	1
Output	Result (web site, GUI, report) and output format	3

Table 3.7: Evaluation results of EAGER and PALEOMIX with respect to criteria as defined in Supplementary Table A.4.

Category	Criteria	EAGER	PALEOMIX
Basics	Project homepage	yes	yes
Basics	Organisation	yes	yes
Basics	People Involved	yes	yes
Basics	Brief description	yes	yes
Basics	Year of Publication	2016	2014
Basics	Number of references	37	81
Basics	Analysis	yes	yes
Basics	Commercial	Academic	Academic
System properties	Type of installation	Server, Local, HPC, Cloud	Server, Local, HPC
System properties	Mode of access	GUI / CLI	CLI
System properties	Source code available	yes	yes
System properties	How to obtain	GitHub (GPLv3)	GitHub (GPLv3)
System properties	Data storage	SAM/BAM/CRAM/VCF	SAM/BAM/VCF
System properties	Operating system	Linux/BSD/macOS/Windows* (*via Container)	Linux/macOS
System properties	Software requirements	Java + dependencies	Python + dependencies
System properties	Maintained	Yes, 2018	Unclear, 2017
Standards	FastQ/SAM/BAM	Yes	Yes
Standards	VCF	Yes	Yes
Pre-processing	Pre-processing methods	AdapterRemoval, Clip&Merge, FastXTools	AdapterRemoval
Pre-processing	Quality control methods	FastQC	-
Analysis	Mappers	BWA, BWA-mem, Bowtie 2, CircularMapper, Stampy	BWA, Bowtie2
Analysis	PCR Duplicate removal	Picard MarkDuplicates, DeDup	Picard MarkDuplicates
Analysis	Workflow support	No/Yes	No/Yes
Analysis	DNA Damage metrics	Yes (mapDamage 2 + DamageProfiler)	Yes (mapDamage 2)
Analysis	DNA contamination assessment	Yes (schmutzi)	No
Analysis	DNA mapping metrics	Yes	Yes
Analysis	Other analysis types	Preseq, GATK: Genotyping, VCF2Genome	SAMTools: Genotyping, Phylogeny reconstruction
Analysis	Parallelized? (Process/Sequential)	Yes/Yes	Partial/Yes
Software testing	Tested (CI, Integration)	Yes/Yes	No/Yes
Reproducibility	Container support	Yes (Docker + Singularity)	No
Output	Result	BAM	BAM
Output	Export formats	SAM/BAM/VCF/FASTA	SAM/BAM/VCF/FASTA
Output	Additional output features	Interactive report HTML, PDF, XLSX, Pipeline execution report	Pipeline execution report

the pipeline is implemented in a scalable way, by using parallelization where applicable, we were able to utilize the computational resources of systems up to 64 CPU cores and 500GB RAM each, as well as a standard grid infrastructure to run several dozen to hundreds of samples in parallel. For interested users, we provide a simple VirtualBox image with the whole pipeline integrated for testing purposes via our homepage⁵. However, this version relies on virtualization, thus being relatively slow in execution, as the underlying operating system is emulated in a virtual environment and should not be used as a productive installation. On top of this, EAGER is available as a stand-alone version, too. The pipeline's components can

⁵<http://bit.ly/eagervbox>

be downloaded from www.bintray.org, requiring a Linux operating system and a Java environment. All the dependent tools need to be installed by hand when using this stand-alone version of the pipeline. To make the installation easier for advanced users, we provide a set of shell scripts on GitHub that can be edited by the user for his convenience. Lastly, EAGER was initially available as a Docker image with accompanying helper tools written in the Go programming language. We furthermore provide access to a Singularity container [100]. Both the Docker and Singularity concepts share the basic idea to separate the operating system executing a particular application from the actual implementation. Although this has been possible using full virtualization technologies such as Xen, KVM, and QEMU, the Docker concept employs several new technologies such as `cgroups` and `namespaces` to run directly on the host's hardware without the typical overhead that other virtualization technologies impose [32].

Unfortunately, the Docker concept proved to be more difficult in maintenance, especially since parts of the pipeline are solely available as GUI applications, thus requiring complicated implementation tunnels for tunneling the GUI outside of the container with the pipeline. The concept of Singularity provides a much better-suited framework for such tasks, and therefore the application container was switched to using Singularity for pipeline containerization instead. A further improved aspect here is that Singularity applications can be executed, e.g., on HPC cluster environments with complex user access control, which is often not the case for Docker images, due to security concerns [100]. Users can furthermore download the entire pipeline with all dependencies and keep the downloaded image with a project on their environment, being able to reproduce an analysis workflow in the future with a specific pipeline version. To even further enable users to utilize these benefits, the Singularity container provides a simple method to produce a list of used software tools within the pipeline with their respective version inside the container. This is especially useful for publication matters, where a detailed statement which software tools in which version was used is required. Having both the possibility to store the container and utilized software versions quickly is critical for modern data science purposes.

Support for Singularity can be installed on cloud providers such as Amazon AWS, Microsoft Azure or different platforms, thus making the whole EAGER pipeline portable to such cloud providers. Following up on best practices guides in Bioinformatics [115], the whole pipeline and all accompanying tools contributed within the scope of the project are available on GitHub as open source projects under the GNU Public Licence v3 (see ⁶ for details). Additional information such as the source code for all applications, including DeDup, CircularMapper, VCF2Genome, Clip&Merge and an installation script for stand-alone installation is available there, too. Other researchers in the field are welcome to contribute new ideas, open bug reports or feature requests and contribute to further development of the pipeline.

⁶www.github.com/apeltzer/EAGER-GUI

3.5. Availability & requirements

To complement this, extensive documentation for all the contributed tools and the EAGER pipeline is provided on the Read The Docs platform⁷. This includes extensive installation and general usage instructions, descriptions of all modules within the pipeline and additional guides for the interpretation of analysis results. On top of this, several hands-on tutorials on exemplary data were created, to provide simple hands-on experiences for end users eager to learn how to apply EAGER. The tutorials try to introduce the dedicated users to understand how the pipeline handles particular use cases and is completed with a step-by-step manual including screenshots of the whole analysis procedure. As we received feedback often, a frequently asked questions (FAQ) section has been added to the documentation too, providing advice for certain types of interaction errors that occur more frequently. Although the primary documentation is available as an HTML webpage and can be interactively searched, for example, users can also download the documentation entirely for offline usage in PDF or ePUB format.

⁷<http://eager.readthedocs.io>, last accessed Oct 10, 2017

3.6 Application of EAGER in a forensic case: Inferring genetic origins and phenotypic traits of George Bähr, the architect of the Dresden Frauenkirche

Text and Figures in this section were adapted with minor modifications from our work published in Scientific Reports[163].

3.6.1 Introduction

At the beginning of aDNA research, many of the projects had a focus on determining evolutionary information, such as signals of adaptation to altitude [73], or to climate [232] and for example the evolution of a gene playing a major role in language development [96]. In those early days, the laboratory protocols and methods utilized did not allow for a cost-efficient sequencing of more than a few individuals, more or less strictly limiting research questions in that area to those that can be answered with aDNA from single individuals. With the rise of modern and cost-efficient SNP capture enrichment methods, the number of available genomes from ancient specimen drastically increased between the first complete ancient human genome in 2010 and 2017 by several orders of magnitude [212] to a few hundred available genomes. Apart from questions in population genetics, the possibility to cost-efficiently sequence individual genomes from ancient humans also has led to projects within the context of forensics in the last couple of years [10, 33, 86, 103, 156, 165, 187]. Within the context of forensic studies, typical questions are more tailored towards the identification of individuals, determination of a cause of death or other diagnostic questions.

Previous analysis types included either partial mitochondrial information, such as the hypervariable region I, II or the D-Loop of the mitochondria only. Within the forensic case of George Bähr, we had the opportunity to additionally obtain a high-density SNP capture dataset apart from the full mitochondrial capture. As this requires the processing of several different data types, this forensic case represents a suitable case for the application of EAGER. The main intention in the case of George Bähr was to investigate how much additional information can be retrieved by applying modern in-solution SNP capture methods and utilizing population genetics and medical genetics tools for genetic identification purposes. This required analyzing raw sequencing data from various sources, including an initial shallow whole genome sequencing, a mitochondrial capture dataset, and a high-density SNP capture dataset. George Bähr himself is widely renowned for his work as an architect of the Dresdner Frauenkirche (Figure 3.19), one of the most important monuments in German history due to destruction during the last weeks of World War II and its reconstruction after the German reunification in 2005. Bähr was

3.6. Application of EAGER in a forensic case



Figure 3.19: A current view on Dresden Frauenkirche after its reconstruction. The church is known as one of the most iconic cultural sites, especially after World War II and the reconstruction of the church after German reunification and was designed by George Bähr. Picture reprinted with permission from Wikimedia[91].

born in 1666 in Fürstenwalde [40, 52, pp. 22 ff], close to Dresden and moved to Dresden in 1690. In 1705, he was appointed Master Carpenter of the city of Dresden [121, pp. 171, Kat-Nr. 45]. Unfortunately, he was unable to see the outcome of his most prominent piece of work, as he died from supposedly pulmonary edema in 1738, shortly before the church was ultimately finished [52, pp. 199 ff]. There is almost no written material available on him other than basic family background and ancestry information. Even historians investigating his private and professional history were unable to reconstruct much more than this fundamental and shallow information on him [40, 136]. When Dresden Frauenkirche was reconstructed between 1990 and ultimately finished in 2005, parts of the skeleton of George Bähr were found. Subsequently, a genetic analysis project was conducted to shed more light on him with current methods of aDNA analysis.

3.6.2 Results & discussion

To determine whether deeper enrichment or sequencing was even applicable on the DNA samples obtained from the skeleton of George Bähr, an initial shallow shotgun sequencing (WGS) library was created. After sequencing, EAGER was used as defined in Table 3.8 to assess basic parameters of the sequencing library such as DNA content, the complexity of the library for future deeper sequencing and DNA damage. As the normal amount of DNA yield of this shallow sequencing run was expected to be low, only a straightforward analysis procedure was configured without performing genotyping or more complex downstream analysis.

The endogenous DNA content of the shallow WGS library was confirmed to 62.2% and DNA damage patterns of 7.5% on the 3' and 5' ends were observed, as also shown in Table A.1 and in Figure 3.20. Regarding library complexity, a cluster factor of 1.005 also assured that deeper sequencing attempts would probably produce even more DNA fragments of unique origin. Thus, a mitochondrial DNA

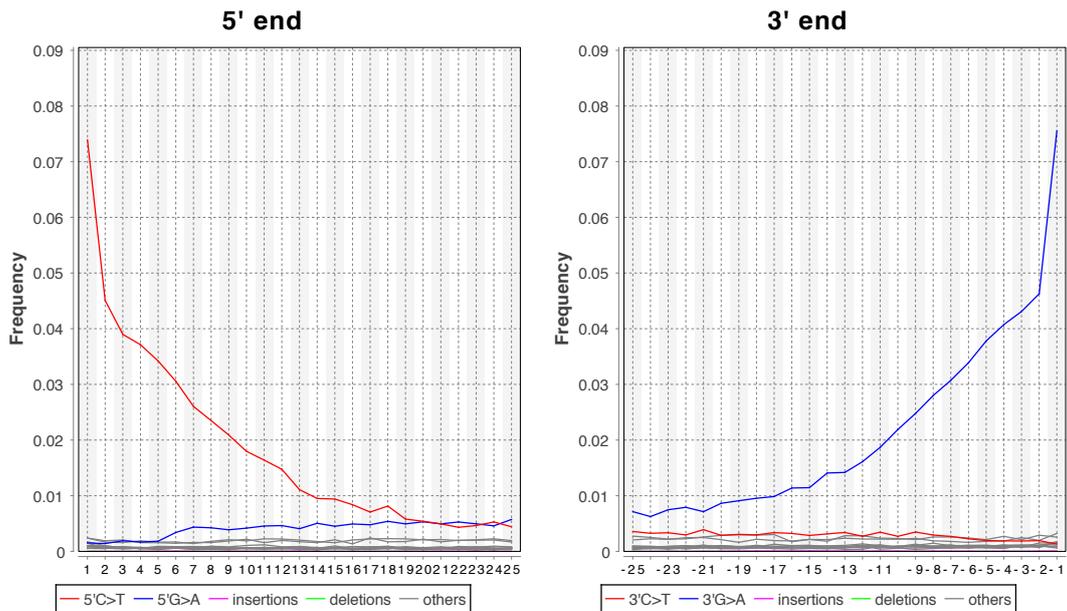


Figure 3.20: DNA damage patterns for 5' and 3' ends of aligned reads using the initial screening WGS shotgun data. The plots have been created with DamageProfiler within EAGER.

capture experiment was performed to generate a full mitochondrial genome for George Bähr. The EAGER pipeline was configured as shown in Table 3.8. Similar to the previous WGS screening analysis, methods for initial quality control (QC) and adapter removal with read merging have been chosen. To improve mapping results for the circular mitochondrial genome, the CircularMapper method was applied in this case. After removing duplicates with DeDup, a contamination estimation with schmutzi was performed and subsequently followed by DNA damage calculation. As shown in Table A.1, a mean coverage of 395.34X on the mitochondrial genome was achieved, accompanied by DNA damage of 7.4% on both 3' and 5' ends of the reads. Additionally, the mitochondrial contamination estimation with schmutzi converged at an estimate of 1 – 2% mitochondrial DNA contamination, which is well below a typically applied threshold of < 3% [181]. Besides, two 390K in-solution SNP capture libraries were sequenced to perform a more detailed population, phenotypic trait- and disease-related investigation. The pipeline was configured as illustrated in Table 3.8 again, resulting in a mean coverage of 29.19 X on the selected set of 390 K SNP positions.

3.6. Application of EAGER in a forensic case

Table 3.8: Summary of analysis parameters with EAGER for all three analyzed data types in the George Bähr project. Whole genome shallow shotgun data was analyzed with slightly different analysis parameters than the mitochondrial capture and 390 K capture data. All data types were clipped and merged using the Clip&Merge (C&M) application, CircularMapper was used to map reads in the mitochondrial capture analysis.

EAGER Parameter	Data Type		
	WGS	mtCapture	390 K
Organism Type	Human	Human	Human
Age of Dataset	Ancient	Ancient	Ancient
Treated Data	No	No	UDG+
Pairment Type	PE	PE	PE
Capture Data	No	Yes	Yes
Calculate on Target	No	Yes	Yes
QC	Yes	Yes	Yes
Adapter RM	C&M	C&M	C&M
Mapping	BWA	CircularMapper	BWA
RMDup	DeDup	DeDup	DeDup
Damage	DamageProfiler	DamageProfiler	DamageProfiler
Clean Up	Yes	Yes	Yes
Report	Yes	Yes	Yes

Downstream analysis

From the EAGER analysis output, a set of typical downstream analysis methods were applied to answer specific questions within the scope of the George Bähr project. First, a molecular sex determination analysis was performed using the tool published in Mittnik *et al.* [135] to determine whether the investigated skeleton was indeed male. Fortunately, this could be confirmed by the analysis, reassuring us that the sequenced skeleton indeed belonged to George Bähr. Furthermore, an X-chromosomal contamination check was performed on the 390K SNP capture data using ANGSD[93] to test for potential X-chromosomal contamination, which was reported to be low with 0.003%.

As EAGER already produced a consensus sequence of the endogenous mitochondrial genome, Haplogrep 2 [233] was applied to find the maternal haplogroup of George Bähr. The haplogroup of George Bähr was found to be H35, a common subclade of haplogroup H in Central Europe [7]. Based on the 390K capture data, the Y-chromosomal haplogroup of George Bähr was determined to be R1b1a2a1a2, also a common Y chromosome clade of paternal lineages across most of Western Europe, showing a frequency peak in the upper Danube and Paris area [138]. A detailed principal component analysis (PCA), f_3 , f_4 and ADMIXTURE analysis

of the determined 317,990 SNP positions covered, confirmed these initial findings that George Bähr was of central European origin and most likely had no direct ancestors from outside Europe. Of much bigger interest for researchers from the forensic community are however SNPs which are directly in relation to the phenotypic appearance in individuals. Based on the information that was obtained using EAGER and subsequently hIRISplex [231], George Bähr had brown eyes and light skin, which is a common combination in modern individuals from the same respective area in Germany [220]. This analysis was performed by investigating important genetic loci on George Bähr’s genome manually using IGV [186] and entering the obtained SNP information into hIRISplex [231] to predict George Bähr’s phenotype. We furthermore used the SNPedia based web-service Promethease [23] and uploaded a VCF file of George Bähr’s SNPs to investigate potential other disease risks. After manual curation of identified loci of interest, we can confirm that George Bähr was most likely able to digest milk as he was heterozygous for the *RS4988235* mutation on the *LCT* gene [227, 232], as shown in Table 3.9. Al-

Table 3.9: Phenotyping results of George Bähr. To ensure consistency, the analysis was limited to only include high-quality bases ($q > 30$), and duplicates were removed after merging of both sequencing libraries. The SNP *RS4988235* is responsible for lactase persistence in Europe [9, 35]. Both SNPs at *SLC24A5* and *SLC45A2* are considered to be responsible for light skin pigmentation [216], whereas the SNP at *HERC2* is the primary determinant of light eye color in present-day Europeans [34, 219]. The SNP at *EDAR* affects tooth morphology and hair thickness [47, 85].

SNP	Gene				
	<i>LCT</i> rs4988235	<i>SLC45A2</i> rs16891982	<i>SLC45A5</i> rs1426654	<i>EDAR</i> rs3827760	<i>HERC2</i> rs12913832
Ancestral	G	C	G	A	A
Derived	A	G	A	G	G
Coverage	39×	114×	8×	43×	46×
Derived allele frequency	50%	100%	100%	0%	57%

though the forensically driven project to determine George Bährs phenotype based on genetic data, is so far one of the few examples where this has been attempted, this project was an excellent trial run for the application of the EAGER pipeline. Especially for collaborators in the project, this meant they were able to simply apply the pipeline on their own, obtaining useful first insights into the datasets generated. Unfortunately, there is not much of a historical record on George Bähr available, which is why the obtained results are hard to confirm elsewhere. An important finding of the project was furthermore that the possibility to retrieve more than 300,000 SNPs of a human genome of ancient origin also enables researchers to reconstruct the phenotype of a historical individual at a previously unprecedented scale. However, newer methods such as the 1240K SNP capture, providing the possibility to genotype ancient individuals at 1.24 million SNP

positions simultaneously with much higher resolution, can even improve the situation for future projects. A potential shortcoming of the utilized approach in these forensic projects could, however, be, that the determination and interpretation of connections between genotype and phenotype are still one of the central dogmas of genetics and not resolved entirely. First works on estimating height, appearance and other characteristics have been done in the last couple of years [128], but are still in their early days. However, EAGER can be utilized even by forensic scientists to process their NGS data without the requirement to ask a Bioinformatician for most of their analysis for help, as could be shown within the scope of this project. Comparing the results of the analysis project with former projects, such as the analysis of mitochondrial DNA from other ancient specimens, the approach used to analyze George Bähr works well. We were able to analyze a set of 317,990 SNPs on the autosomal level, accompanied by a complete mitochondrial genome, thus improving the genomic resolution drastically when compared to previous forensic attempts at reconstructing other specimens. While this does not yield better or improved results on the mitochondrial level, the possibilities concerning phenotypic imputation such as eye and skin color imputation are rendered possible with this approach. We conclude that in historical cases, the SNP capture approach utilized to predict George Bährs appearance could be used in the future to infer more interesting information about specific persons of interest, without the financial limitations that would typically occur by applying WGS approaches.

3.7 Discussion

Since its early years, research on aDNA has established itself as a new field in evolutionary biology. Improved methodology in fieldwork and laboratory protocols provide material for an even more detailed analysis than ever before. In the first years, when next-generation sequencing methods were established in the field of aDNA, only single genomes could be generated within individual projects. During the last years, the number of ancient human genomes sequenced grew substantially (see Figure 3.21). It can be assumed, that by early 2018 the total number of 1,000 ancient human genomes will be available for researchers in population genetics of ancient human specimen. This could potentially be utilized to resolve ancient human population admixture events on an even finer scale than ever before. However, this also means that the requirements for computational methods to analyze such large amounts of data are steadily increasing, too. On the one hand, the computational improvements have to keep up with laboratory and methodological improvements, being flexible in their application and enabling researchers to integrate data from a variety of sources. Another important aspect in current projects is the widely adopted value of aDNA for various applications in evolutionary biology: Making a software workflow not only available for few big institutions, but also providing access to smaller laboratories, collaborators or

Chapter 3. EAGER: Efficient ancient genome reconstruction

even individual researchers, is increasingly important. EAGER is one of the first attempts at providing a modern, flexible and maintained framework for a variety of analysis workflows in aDNA research. As of now, a total number of 28 institutions and laboratories all over the globe use the pipeline in a variety of application cases.

As shown earlier, EAGER is implemented in a modular way. This enables re-

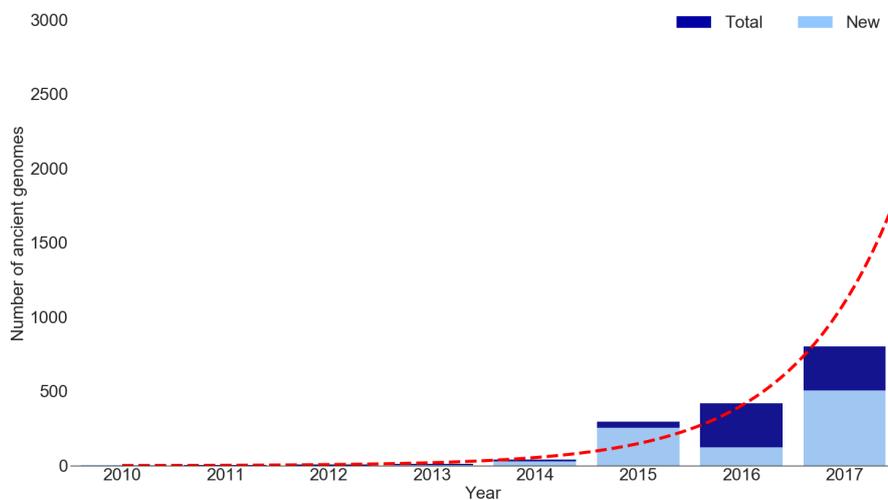


Figure 3.21: “Moore’s law of aDNA”. The total number of available ancient human genomes (WGS or high-density SNP captured), modified and adapted from David Reich’s⁸ SMBE 2017 talk, Austin/Texas, July 4th, 2017. Published genomes included from the following publications: Rasmussen *et al.* [177], Green *et al.* [56]. Rasmussen *et al.* [176]. Keller *et al.* [80]. Fu *et al.* [44]. Lazaridis *et al.* [106], Skoglund *et al.* [206], Raghavan *et al.* [175] Haak *et al.* [58], Rasmussen *et al.* [3]. Lazaridis *et al.* [105], Pagani *et al.* [154], Malaspina *et al.* [122]. Lipson *et al.* [113], Mathieson *et al.* [127], Olalde *et al.* [147]. Listing not necessarily complete. Light blue, new genomes added in the respective year. Dark blue, the total number of genomes available in that year. Shown in red, an exponential growth function showing the behavior or Moore’s law.

searchers to apply the pipeline in various research projects, ranging from fast screening runs to more complex (deep) whole genome sequencing experiments. Furthermore, the pipeline features modules for human as well as bacterial types of data in several input file formats (FastQ, SAM, BAM), thus making the pipeline even more flexible. Within the last years, the pipeline has been applied to several projects [46, 169, 200]. Furthermore of great importance are the improved methods that have been implemented within the scope of the pipeline, to extract more data out of important samples, compared to default applications that are for

example not tuned well to aDNA characteristics.

In such a context, EAGER features two improved applications (DeDup and CircularMapper) to improve the general output of duplicate removal and mappings on circular genomes such as (human) mitochondria or bacterial genomes. Integrated into a sophisticated workflow, this even further manifests the important role that such a pipeline plays in aDNA projects. Regarding computational efficiency, EAGER features several modern tools and methods such as BWA, Bowtie2, SAM-Tools and the GATK, thus enabling a fast and efficient analysis of aDNA. Most research projects in the aDNA community rely on using these methods for their analysis [56–58, 106].

Wherever this is applicable; the pipeline follows industry standards in file formats and output. Especially in large-scale collaboration projects, the outputs of the pipeline are commonly shared with collaboration partners on a project basis, which requires that results are stored and made accessible in standardized output formats. EAGER is capable of producing BAM, VCF and FastA output, accompanied by mapping statistics in XLSX, HTML and CSV formats. As these are standardized data formats, researchers can easily compare their results with other researcher's results or share their results with other institutions.

An aspect that can be as well motivated by more collaboration in the aDNA community is furthermore the usability of utilized software, combined with the possibility to share workflows with other researchers in the field. Before efforts like EAGER or PALEOMIX [198], most research groups were relying on their own 'self-cooked' analysis procedures, rendering data integration with other groups (or even within groups) extremely difficult. As we have seen in the detailed comparison of both pipelines, EAGER is going several steps further than PALEOMIX, providing users with several improved measures to establish their analysis procedures and rely on well-tested and established analysis workflows. Furthermore, EAGER integrates several more methods, e.g., a wider variety of mapping algorithms, improved PCR deduplication strategy, an improved method for circular genome handling and several different genotyping methods such as ANGSD [93] and GATK [129] in comparison to PALEOMIX. Additionally, the pipeline is well documented and features several up to date module descriptions, tutorials and Youtube videos explaining the setup, installation, and usage of the pipeline in an easily (and openly) accessible way. Combined with the efforts to provide EAGER as Singularity and Docker images to enable researchers to run and install the pipeline without complex installation procedures, is furthermore lowering the barrier for smaller research groups to run their aDNA projects powered with EAGER. Based on these two technologies, EAGER can for example even be used in software as a service type of business model (SaaS) on cloud providers supporting Singularity and Docker images. Considering the costly demands of computational hardware, this can provide smaller groups with the ability to rent infrastructure, thus greatly reducing the investments required for a typical local analysis workflow. EAGER is currently maintained, tested

Chapter 3. EAGER: Efficient ancient genome reconstruction

and extended at two institutions, the Max-Planck-Institute for the Science of Human History in Jena and the Integrative Transcriptomics group in Tübingen and will be further developed by these groups beyond the scope of this dissertation. This ensures that the pipeline will hopefully be maintained in the upcoming future and thus be of help in aDNA analysis projects.

As has been shown in this chapter, EAGER is a consistently updated and improved framework for aDNA reconstruction, enabling researchers to work on aDNA in a previously unprecedented straightforwardness. It features state of the art methods for the analysis of aDNA, along with two new methods for improving read mapping on circular genomes and an improved PCR deduplication method and integrates well-tested [110] and established methods [129]. This ensures that researchers can perform their analysis in a well defined and tested way, improving the outcome of their analysis projects over formerly applied shell scripts. Combined with the graphical user interface, extensive documentation, the cloud compatibility and its well-tested and established workflows EAGER can be seen as an increasingly important asset for many aDNA researchers.

CHAPTER 4

mitoBench & mitoDB: Modern tools for mitochondrial genome analysis

4.1 Introduction & motivation

While evolutionary biology covers many diverse topics, one of the most thriving areas of itself is seen in population genetics of ancient humans [207]. Though there are many interesting topics to be studied on populations of mammals such as wolfs [55] or mammoths [160] for example, the bulk of current research in population genetics focuses on human population genetics [58, 105, 106, 201]. As can already be deduced from a famous Charles Darwin quote: “If we admit a first cause, the mind still craves to know whence it came and how it arose.”¹, humans were and will foremost be interested in their origins. The last years have seen the advancement of experimental NGS methods [131] and following up on that the development of computational and theoretical methods to leverage the information stored in DNA [145, 159]. Since then, researchers have been investigating the origins of humanity on a variety of research questions using aDNA, characterizing the ancestry of multiple populations to an increasingly larger extent. Using aDNA, scientists in paleogenetics today are capable of exploring the ancestry of populations that went extinct thousands of years ago and can, therefore, recover information that archeologists on their own cannot detect. Migration patterns, population admixture and other events that contribute to the shape of today’s societies can be deduced and tied to artifacts found by archaeologists [104]. An important milestone that has been unraveled using population genetics methods was the discovery that modern European seem to be an admixture of three ancient

¹Charles Darwin, “Life and Letters of Charles Darwin, 1873”

*Chapter 4. mitoBench & mitoDB:
Modern tools for mitochondrial genome analysis*

ancestral populations (hunter-gatherers, farmers from the middle east and ancient north Eurasians (ANE)) [106].

While many research projects today focus on utilizing SNP capture panels and WGS [3, 58, 105, 106, 209, 210], there is still a large community focusing on mitochondrial DNA [14, 15, 48, 148]. The benefits of using mitochondrial DNA are clear: Mitochondrial DNA requires less memory and storage on the computational side, as the mitochondrial genome is 16kB in size, compared to 3.2Gb for the nuclear human genome. Furthermore, due to this size constraints, there have been more efforts to sample and sequence mitochondrial genomes, which has lead to an extensive available dataset of comparable modern data for analysis questions. Lastly, cheap target capture and sequencing methods specifically for mitochondrial DNA are available [124].

Independent of the DNA being sequenced, being able to compare genetic data sets between ancient and modern populations, requires computational methods such as F_{ST} [70, 72], that can calculate genetic distances between individuals. Computational methods that can be used to determine a variety of distance metrics for genetic data are for example Arlequin [36] or GENEPOP [190]. While being still frequently applied, Arlequin for example does not offer a modern user interface with intuitive and user friendly options to aid users in running their own analysis. Another shortcoming of Arlequin specifically is, that the currently available tools do not support direct file imports that are commonly in use such as FastA. The typical user would like to run his analysis using a specialized pipeline, such as EAGER, and then import the ultimate result directly into the subsequent population genetics analysis software. Arlequin does not support any other file formats to be imported other than their own "ARP" format. This requires users of Arlequin to be able to generate "ARP" files with a text editor, which can be a cumbersome and error-prone process when dealing with multiple samples at the same time. A further shortcoming of Arlequin is that there exist only few methods to aid users in exploratory data analysis. While this is understandable due to the long-term availability of Arlequin, which was created during a time where this was not of high concern, modern software methods should also have a focus on helping users to perform such an exploratory data analysis, wherever possible. Modern exploratory applications provide visualization methods that help researchers to determine e.g. differences between populations or groups of individuals. Generally, such differences are easier to spot when appropriate visualization methods are utilized in a software method [41], than solely e.g. displaying a table of numeric data with some meta information. The simple plotting functionality that is available in Arlequin does not fulfill these requirements and is even further limited by the necessity to run certain R scripts to plot results in the end. A more advanced analysis tool could provide users with the possibility to import their pre-analyzed data easily and subsequently allow for easy data exploration, ideally even with visual feedback.

Another big issue in population genetics is the acquisition of comparable mitochon-

4.1. Introduction & motivation

drial genomes for a specific analysis question. Typically, many research questions address questions like determining differences, e.g., between individuals, groups of individuals or entire populations, for which such comparable mitochondrial datasets are a requirement. This situation is not just true for aDNA but holds for modern DNA as well. Although there exist databases that provide genetic data sets to the public, most of these focus on whole genome data, such as the 1000 genome project (1000G) database [1] or the online ancient genome repository (OAGR) [223]. A big drawback of these large-scale databases is also that researchers interested in mitochondrial genomes would need to download WGS datasets and then extract the required mitochondrial genome reads from the selected specimen, which is again a cumbersome process. Some smaller databases, such as MitoMap [118] provide access to mitochondrial genomes, but have a focus on mitochondrial diseases and do not list population specific meta information explicitly with the uploaded specimens, making them hard to apply for population genetics analysis questions.

To sum up the current state of the art, the acquisition of mitochondrial genomes from a variety of resources is an arduous process. Therefore, researchers created their private collections for storing collected data in their ways. In a recent survey amongst 15 other scientists in the field², the large majority kept their private collections in Microsoft Excel sheets ("XLS"), or as "CSV" or "TXT" files. Increasingly problematic for performing any analysis in this way is the possibility to introduce substantial errors when using Excel as has been demonstrated in a recent publication by Ziemann *et al.* [236]. A commonly introduced error when using Excel is, for example, the automatic renaming of text to date formats or the conversion of gene names to date formats and vice versa.

All these issues are further exacerbated by the situation that most scientists would also like to integrate multiple data sources with each other. For example, population geneticists would like to deduce the maternal haplogroup information of their individuals to investigate potential maternal origins of their individuals [151]. In archaeological contexts, researchers, for example, determine radiocarbon (C14) dates that they would like to store together with their samples in a database. As research questions can vary quite drastically, the kind and form of metadata that scientists would like to store with their genetic samples can vary widely, too. Storing generic information types exacerbates the situation rapidly, as solely focusing on genetic information handling can hinder later explorative data analysis. Although methods and data resources such as Arlequin [36], MitoMap [118] or more recently the MitoSuite [75] aims at resolving these, the current progress in achieving this can be summarized as vastly incomplete. Simple research questions such as investigating whether a genetic variant can only be found amongst people speaking a certain language dialect can only be efficiently detected when the meta information for languages is standardized, well-defined and stored in direct relation to the genetic information investigated. A further big issue here is the in-

²Personal communication, within the current mitoBench and mitoDB projects.

*Chapter 4. mitoBench & mitoDB:
Modern tools for mitochondrial genome analysis*

coherent data standardization in human mitochondrial genomics: While there are approaches to follow certain guidelines, such as using Glottolog [60] for language and dialect classification, most researchers generating their private collections currently do not follow general guidelines or official standards. With the current lack of potent software methods to aid researchers in the field of human mitochondrial genomics in data curation, acquisition, and analysis, the situation tends to stay at the current level.

The mitoBench tool concept has been developed during this dissertation to address the need for an integrated solution to handle and visualize mitochondrial DNA data in population genetics. MitoBench builds on top of previous (unpublished) command line scripts that have been used in the analysis of ancient mitochondrial DNA in the scope of the Egyptian mummy project [201]. The new application concept is not just an extension of this previous work but improves upon these simple scripts with expanded data handling, conversion, and visualization concepts that were not present in the command line scripts initially used for the Schueneemann *et al.* project. Neither of the available competitor tools such as Arlequin [36], GENEPOP [190], nor the newer MitoSuite [75] provide similar functionality. Work on mitoBench has been started in the Java programming language with the Java FX GUI toolkit as the basis for the entire interface. This aims at making the application as platform agnostic as possible and only requires a working Oracle Java runtime environment to run. As argued before, the computational analysis methods part is not the only problematic part of modern population genetics on human mitochondrial DNA. Having a reliable database for comparable data is crucial for any analysis project. Unfortunately, currently available databases either have a medical focus such as Mitomap [118] or aim at WGS data such as the 1000G project [1]. Therefore, the mitoDB database and web UI prototypes have been created to resolve data acquisition issues when comparable mitochondrial genome data is required in population genetics research projects. The mitoDB database has been developed using the PostgreSQL database object-relational database management system (ORDBMS) [223] for the actual backend and the Java programming language with the Vaadin framework [226] for the web UI. In the following, a detailed presentation of the underlying concepts of both applications mitoBench and mitoDB is given. Additionally, an ancient Egyptian mummy project is analyzed using the current mitoBench prototype as a real data set in Chapter 5 to demonstrate the improved analysis process when utilizing the mitoBench application compared to competitor tools.

4.2 Conceptual application design

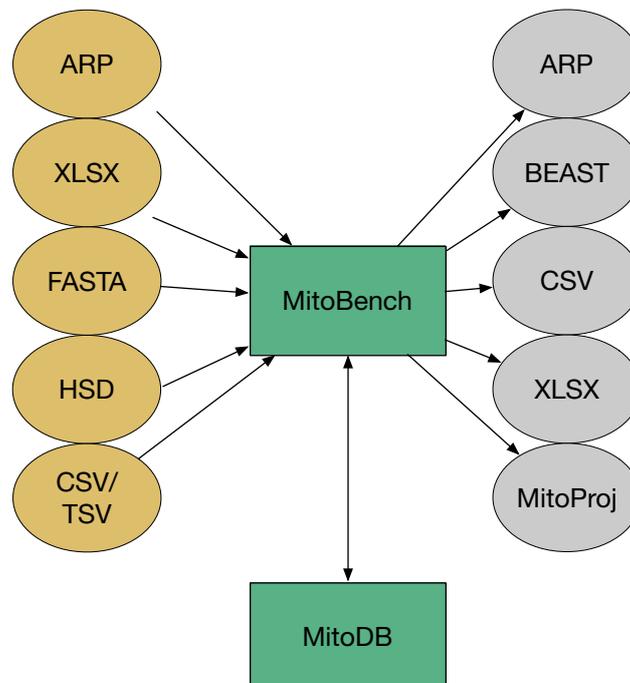
4.2.1 MitoBench

The conceptual idea of mitoBench was based on the experiences obtained when analyzing the Schuenemann *et al.* dataset of 151 ancient Egyptian mummies. In general, the application was modeled to offer a variety of features and possibilities to interact with mitochondrial genome data. The basic requirements for the concept of mitoBench were:

- the ability to combine and collect multiple file formats in a single tool.
- the possibility to create several output formats for downstream analysis tools.
- to visualize imported genomic data and meta information, aiding researchers to explore the datasets they create.
- to be as flexible as possible in the application, thus not enforcing, e.g., usage on human genomes solely.

These requirements are also shown in Figure 4.1. In the following, the ideas and

Figure 4.1: Basic import and export scheme for mitoBench. The application was designed to support multiple common file formats both for import and export. Furthermore, the direct connection between mitoBench and mitoDB is shown, that enables direct interaction between both tool and database.



concepts behind mitoBench are explained and detailed.

User interface

To be able to fulfill the demands of population geneticists interested in analyzing their data with an integrated application, mitoBench has been designed as a GUI-based method. The idea is, to provide an application that can be used to view mitochondrial genome data in various ways and is structured in four main components. An overview of the main GUI prototype is shown in Figure 4.2. The basic component of mitoBench that can be used to handle data imported from various sources in a table format is implemented as a JavaFX table, providing direct interaction possibilities such as sorting, filtering and rearrangement of columns. Furthermore, there are two visualization components planned for the main interface: The leftmost panel can be used to display plots or other visualization results, whereas the rightmost panel is used to show a table of calculation results or simple counts tables. To further improve the user interface, both visualization panels offer tabbed document interfaces (TBI/TABs), commonly known from modern web browsers to enable quick switching between different documents, or in this case results of a certain analysis. The last of the four components consists of the menu bar on top of the main interface. The menu bar presents various functions, including data import and export and provides access to visualization methods or database access functions of mitoBench.

Import and conversion features

The first conceptual idea of mitoBench was to be able to import data from a variety of different sources. Additionally, the combination of these data sets together in a single integrated fashion was of the highest concern. MitoBench supports (multi-) FastA, HSD (Haplogrep 2 [233]), ARP (Arlequin [36]), XLSX (Microsoft Excel) as well as a generic tab separated value (TSV) format for both import and export functionality. All required input details are specified in the documentation for the user. Importing files from multiple different files and combining output from these together requires the user to use unique identifiers for each sample. Once imported, the export functionality of mitoBench enables users to export to the same number of output formats directly. For output formats requiring the definition of groups of samples (e.g., Arlequins ARP format), users are automatically asked to specify such groups. MitoBench provides both context menus and functionality in the main menu on top of the application to define the desired grouping for such cases. All these components enable users to import their data efficiently and still be able to use different tools for downstream analysis procedures with ease, while still benefiting from the organizational features that mitoBench provides.

Data Integration & Reproducibility

A big issue when working on population genetics projects is the correct, consistent and efficient handling of multiple samples. Applications such as Arlequin require

4.2. Conceptual application design

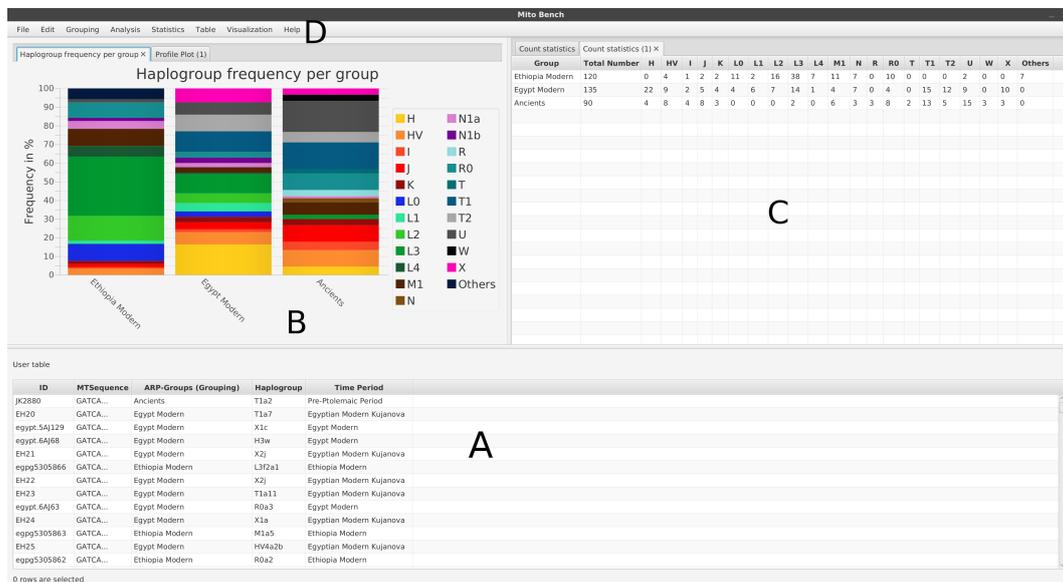


Figure 4.2: The mitoBench graphical user interface. The conceptualized interface consists of four distinct main components. **A:** The main user table to display the currently imported project data, with mitochondrial DNA sequence, identifiers, grouping information, haplogroups and C14 radiocarbon dates. **B:** The visualization panel, displaying an interactive visualization depending on the users choices. **C:** The counts or numerical panel, showing statistical or numerical results from frequency calculations, counts or other types of statistical analysis. Both the visualization panel and the numerical panel can show multiple tabs to keep multiple analysis results accessible while exploring data. **D:** The main menu, providing access to more advanced features of the mitoBench application.

users to manually convert various file formats, calculate interesting metrics and combine results manually to come to a conclusion. Our concept behind mitoBench provides the possibility to store the current table content in a project file with a “MITOPROJ” extension. The created file is inherently just a text-based storage format but includes all the data in a running project session of mitoBench. This ensures that once the project is stored in a file, it can be sent to e.g. collaborators or re-used to reinitialize a project at some point. This idea improves general data handling, since users do not have to import multiple e.g. FastA files and the corresponding metadata, but can instead use the project file. Especially when long-running projects are analyzed, where sequencing efforts of more samples are a common practice, this eases data handling inherently. Another conceptual idea behind mitoBench is to automatically produce a log file to ensure that users can trace back what kind of analysis has been done in an analysis session. Especially in cases where data is explored interactively, this resolves certain issues such as the inability to tell what has been done with a data set explicitly. Subsequently, there are plans to ask the user in a mitoBench session whether to store such a log file

when closing the application.

Software testing & Documentation

To ensure that the application is maintained well, the mitoBench application will be tested throughout the entire development process using JUnit 5 and TestFX [224]. Specifically, the components for importing and exporting will be tested automatically whenever the application is built using the Travis CI continuous integration service. Additionally, plans to test the mitoBench GUI exist by utilizing the TestFX framework. Several test cases that simulate clicks and interaction in the various windows can be triggered and the expected behavior will be tested automatically. This ensures that the software behaves as intended on all kinds of platforms, providing benefits during software development to investigate potential errors when they arise in an early stage.

During the conceptualization of mitoBench a state-of-the-art documentation was created online at <http://mitobench.readthedocs.io/>, which will be updated in the future to reflect newer additions and changes. Using the same platform as for the EAGER project, users have the choice between HTML, PDF or ePUB documentation enabling them to learn how to use the application. We furthermore provide a FAQ section in the documentation, along with some exemplary data for evaluation purposes.

4.2.2 MitoDB

MitoDB offers several possibilities to get access to stored human mitochondrial data for population genetics. In the following, the concept and process behind the design of mitoDB are explained in detail. Furthermore, an introduction into the current features of mitoDB is provided.

Database & design decisions

Defining a suitable database structure was one of the biggest challenges during the initial development phase of mitoDB. To successfully provide an appropriate and functional database layout, we conducted a user study, including all members of the mitoDB consortium. Users were asked to provide an exemplary data set of their own (private) mitochondrial genome collection and meta information for the selected samples. All users provided Excel sheets with mitochondrial DNA samples and accompanying meta information, with typically around 100-200 samples per collaborating group. After screening these exemplary data sets, we generated a proposal for the actual database, including all the information of the participating groups. In many cases, there was substantial overlap between the layouts that individual groups were using, for example regarding how genetic information was stored in the used tables respectively. However, the research contexts in the

4.2. Conceptual application design

participating groups are differing considerably. Thus other parts of the individual excel based “databases” were distinct as illustrated in a Venn diagram (Figure 4.3). More specifically, specialized meta information fields containing for example a language context for investigated samples were typically not found in other groups submissions. After generating the proposal, we conferred with experts on database

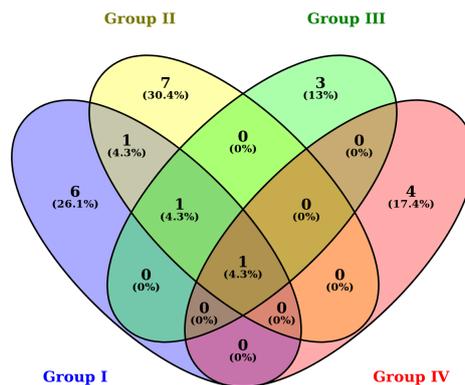


Figure 4.3: Venn diagram visualizing the differences between current “database” layouts in four research groups within the mitoBench and mitoDB consortium. As can be seen, the overlap between groups is minimal, while differences in the way how data is annotated with meta-information prevail between groups.

systems³ on how to generate a technically sound and efficient database system for our purposes. The general guidelines of the mitoDB, therefore, follow these best practices guidelines we discussed during several meetings. To store the data for mitoDB, much of the original data tables were abstracted and stored in several table layouts, as shown in Figure 4.4. The database has been built and designed with PostgreSQL, such that simultaneous access by hundreds of users with thousands of queries is possible in theory. Furthermore, the design allows running more than one database backend server, making a further scaling of the database layout possible on demand in the future if this is required. The layout shown in Figure 4.4 described the general layout of the mitoDB. In total, the database consists of 9 tables with a logical abstraction between individual tables. The overall database is structured on the basis of samples, where each sample (stored in the “Samples” table) can have multiple relations to other tables and reference additional meta-data. We intended to separate data that is required in all cases from data that is optional. As each sample can have references to multiple other tables in the database, optional data can be queried easily on demand. This simple abstraction

³Benjamin Dietrich, Database systems group, University of Tübingen, Personal communication

Chapter 4. *mitoBench & mitoDB:*
Modern tools for mitochondrial genome analysis

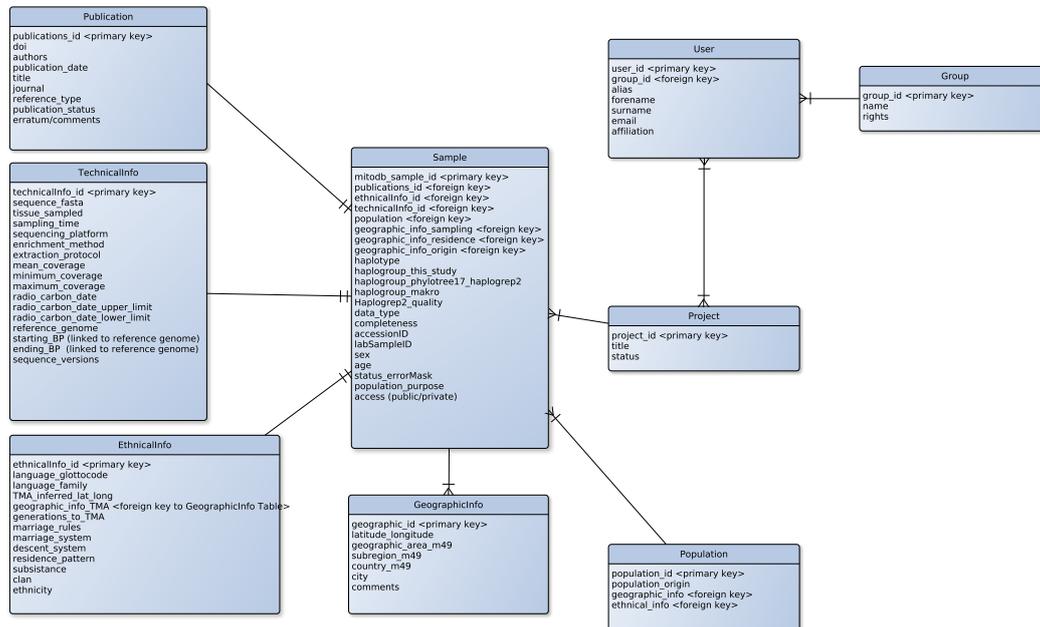
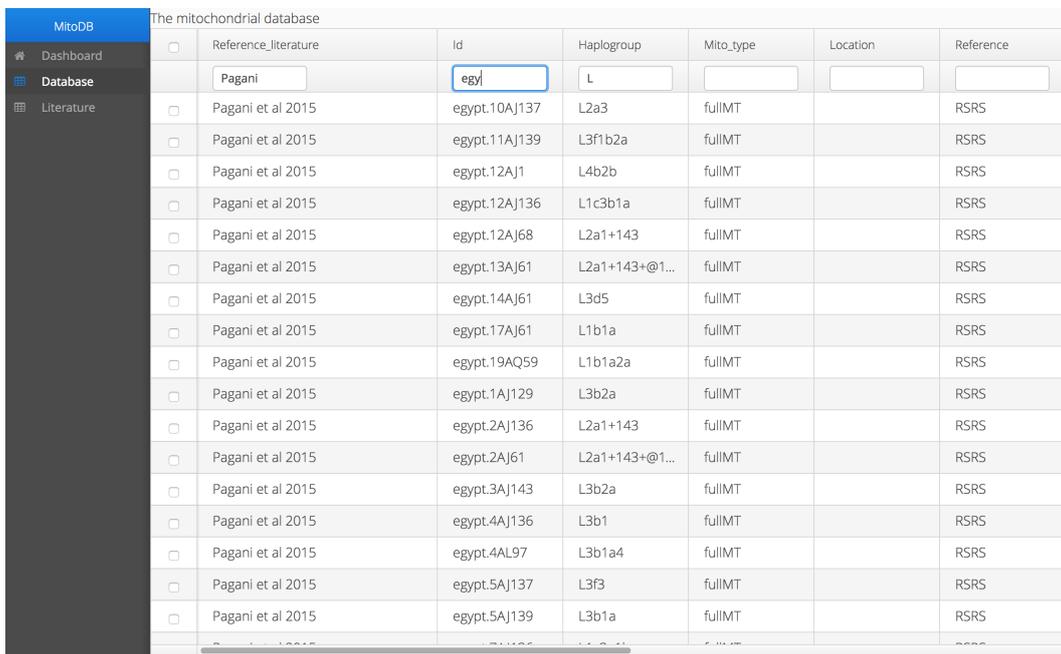


Figure 4.4: General mitoDB layout, displaying the available individual tables of the database. Relations between individual tables are shown with relational arrows. Note, that this is a current state of the database layout and may change in future versions of the database.

ensures that the return times of a query are kept at a minimum. For example, users will typically request information on samples and populations but might only query the actual mitochondrial sequence data in the last step of an analysis query, when some computational work on the mtDNA sequence has to be performed. Furthermore, this abstraction allows for easier UI creation as solely the main sample table content (see Figure 4.4) is displayed by default in the mitoDB web UI (Figure 4.5). A further aspect of the initial database design decision process was the data curation process in mitoDB. Medical databases such as EMPOP [157] utilize a sophisticated data polishing process, whenever new data sets are added to the database. The problematic part of these curation processes, however, is, that many of them are only semi-automatic. This situation ideally requires a responsible curator to take care of any new submissions to the database, whenever new data sets are becoming available for example due to new publications.

Within mitoDB, the design decisions only define low-level database servicing methods that are fully automatic. For this purpose, mitoDB uses the Haplogrep 2 [233] quality scores when haplogrouping samples upon initial upload to the database. Additionally, there are plans for a user rating system, which enables users to rate samples after using them for a particular analysis. These two individual approaches to mark samples with inferior quality with a technical method and an additional user rating system should suffice for a population genetics database in our opinion.

4.2. Conceptual application design



The screenshot shows the web interface for the mitochondrial database (mitoDB). On the left is a dark sidebar with a menu containing 'Dashboard', 'Database', and 'Literature'. The main area is titled 'The mitochondrial database' and contains a table with columns: Reference_literature, Id, Haplogroup, Mito_type, Location, and Reference. A search bar at the top left of the table contains the text 'Pagani'. The 'Id' column has a value 'egy' selected. The table lists 18 entries, all from 'Pagani et al 2015', with various IDs, Haplogroups, and Mito_types.

Reference_literature	Id	Haplogroup	Mito_type	Location	Reference
Pagani	egy	L			
Pagani et al 2015	egypt.10AJ137	L2a3	fullMT		RSRS
Pagani et al 2015	egypt.11AJ139	L3f1b2a	fullMT		RSRS
Pagani et al 2015	egypt.12AJ1	L4b2b	fullMT		RSRS
Pagani et al 2015	egypt.12AJ136	L1c3b1a	fullMT		RSRS
Pagani et al 2015	egypt.12AJ68	L2a1+143	fullMT		RSRS
Pagani et al 2015	egypt.13AJ61	L2a1+143+@1...	fullMT		RSRS
Pagani et al 2015	egypt.14AJ61	L3d5	fullMT		RSRS
Pagani et al 2015	egypt.17AJ61	L1b1a	fullMT		RSRS
Pagani et al 2015	egypt.19AQ59	L1b1a2a	fullMT		RSRS
Pagani et al 2015	egypt.1AJ129	L3b2a	fullMT		RSRS
Pagani et al 2015	egypt.2AJ136	L2a1+143	fullMT		RSRS
Pagani et al 2015	egypt.2AJ61	L2a1+143+@1...	fullMT		RSRS
Pagani et al 2015	egypt.3AJ143	L3b2a	fullMT		RSRS
Pagani et al 2015	egypt.4AJ136	L3b1	fullMT		RSRS
Pagani et al 2015	egypt.4AL97	L3b1a4	fullMT		RSRS
Pagani et al 2015	egypt.5AJ137	L3f3	fullMT		RSRS
Pagani et al 2015	egypt.5AJ139	L3b1a	fullMT		RSRS

Figure 4.5: The general sample view as shown in the web UI of the mitoDB prototype. The UI already offers some basic filtering functionality and can be opened via a simple web browser. Data can be exported via drag and drop, without leaving the web browser for instance to Gnu R, Excel or other tools.

Specifically, since the context in which mitoDB will be used is less focused on individual diagnosis or identification than on population-scale analysis. In turns, this means that single sequencing errors will most likely not cause significant changes in, e.g., population clustering.

Standardization efforts in mitoDB

Crucial for fast interoperability of the developed applications are standards for data storage and curation. The previously mentioned Excel tables and other resources provided by members of the consortium were not consistently formatted in all cases. Within the mitoDB project, we, therefore, took measures to impose standardized data types on certain parts of the database, thus ensuring that all samples are homogeneously formatted in the database. One of the measures included following United Nations standards such as the M49 standard [228] for geographic or administrative information. For the classification of languages, we follow the definitions as imposed by Glottolog [60] and use ISO standards on all other data columns where this is possible and reasonable. For other cases, the future mitoDB application will integrate further official controlled vocabularies to ensure that data enters the database as consistently as possible.

To ease this process, a Python-based linting application was designed. This tool

can be used to automatically check submitted CSV/TSV files for consistency and check whether the data fulfills required minimum information criteria, which are yet to be finalized in a discussion within the consortium.

The current prototype of the mitoDB application provides access to a set of 2,512 mitochondrial genomes downloaded and annotated from the 1000 G project [1]. The current plans for mitoDB aim at extending this to incorporate all available mitochondrial genomes from GenBank and additionally upload private data collections from all contributing collaboration partners soon. Conservative estimates based on personal communication with collaborators assume a total number of 40,000-50,000 mitochondrial genomes that could be uploaded within the next months, making the mitoDB the largest ever assembled mitochondrial reference database if this holds true. This would also provide a reasonable test case for testing the responsiveness of the mitoDB , before publicly announcing the project in summer 2018.

Interaction with mitoDB

An important improvement over other applications in the general genomics community is the possibility to retrieve reference data from the mitoDB with ease. Other databases, such as the 1000G database, require complex retrieval tools to be installed on the user's workstation or do not provide efficient ways to get certain parts of a database with ease, as for example in GnomAD and ExAC [107]. Therefore, access to mitoDB is possible via two modes of action: First, there will be a direct interaction with the database from within mitoBench. This ensures that users working on a private project can integrate data from the online mitoDB seamlessly in mitoBench for easy analysis. The second planned method to access the mitoDB is via the web UI and thus with a web browser. Providing the database via a web UI (Figure 4.6) is commonly done for such databases, both to increase visibility of the overall project and to keep the engineered barrier as low as possible for new users, who typically have access to a modern web browser.

4.3 Conclusion & outlook

MitoBench has been conceptualized for exploring mitochondrial data in population genetics more efficiently. The application has been designed to aid population genetics in project organization, data analysis and visualization by providing desired methods. Our concept furthermore introduces several features to perform an explorative data analysis of mitochondrial datasets, which was previously not possible with other tools and methods available. The idea of mitoBench also introduces methods to visualize results and provides means to interact with the application in various ways. Therefore, mitoBench will be a valuable tool to investigate mitochondrial datasets in human population genetics, without a limitation to aDNA itself

4.3. Conclusion & outlook

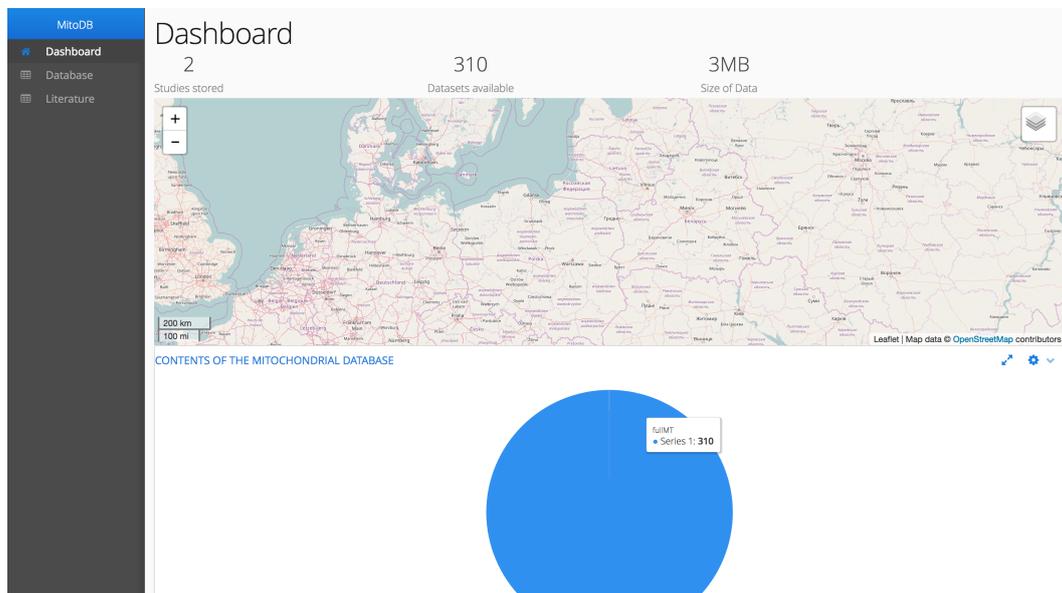


Figure 4.6: Current version of the mitoDB dashboard prototype. The main window displays the current contents of the database, accompanied by a map and preliminary overview charts. On the left side, the navigation bar offers some functionality to display the actual contents of the database in a browser view.

but as well covering projects on current datasets. Previous applications such as Arlequin [36] provided solely functions to run certain population genetics metrics to be calculated. In contrast to that, mitoBench aims at providing measures to combine, collect and visualize a variety of data and thus ensures that any calculations done on the investigated datasets are reproducible and consistent. This, combined with the collection of datasets in a central database application mitoDB aims at simplifying work on mitochondrial population genetics. Especially due to rising numbers of mtDNA genomes created in various projects, more efficient methods are a crucial requirement to keep pace with new data created.

MitoDB offers a standardized data set of human mitochondrial data with accompanying meta information. The database has been designed in a way to allow for the efficient storage of hundreds of thousands of samples. The current version provides direct access with drag and drop of data from the database directly in the mitoBench prototype, further improving the workflow of a human mitochondrial data analysis project when compared to Excel tables or other analysis methods. Both mitoBench and mitoDB are planned to interconnect directly with each other, enabling users to run their analysis without switching applications or requiring other data sources for many analysis questions.

As the mitoBench and mitoDB projects have just been conceptualized very recently, there exist only very basic prototypes of both applications. Within the scope of this thesis, the concept behind both applications was defined, and future

*Chapter 4. mitoBench & mitoDB:
Modern tools for mitochondrial genome analysis*

ideas for further features of both applications were introduced. Most of these will be implemented within the scope of Judith Neukamms Ph.D. thesis projects. The integration of some standard population genetics methods such as a principal component analysis (PCA) or direct F_{ST} calculation has already been achieved. Future directions in development furthermore include improving general visualization functionality, direct interaction with mitoDB via a more sophisticated REST API instead of direct SQL access for security reasons. The aim of both applications in the near upcoming future is, to provide users with an integrative and reproducible analysis workflow and framework for mitochondrial data in a variety of research questions.

Investigating north African population structure: Ancient DNA provides clues to ancient Egyptian history

Some of the text and figures in this chapter were adapted with minor modifications from our work previously published in Nature Communications [201].

5.1 Introduction

Due to its location at the junction of the African and Asian continent, Egypt is a well-suited country for studying population genetics in Northern Africa. Especially the fact that Egypt had substantial trade connections with other countries in the Levant, Sub-Saharan Africa and even with European countries, increases the attractiveness of studying ancient Egyptian populations in a genetic context. Studies to investigate the genetic composition of modern Egyptians have been performed lately [155] but are intrinsically hard to extrapolate to ancient time periods [203]. Of limited use are furthermore literary and archaeological resources, as some of these have shown to be influenced by the integration of foreign idols, lifestyle, and habits [20, 184]. In 1985, one of the first attempts at retrieving aDNA from ancient samples was performed on an ancient Egyptian liver sample of the high priest Nekht-Ankh [152]. This study demonstrated a novelty regarding experimental methodology and founded an entire field of active research, as described in Chapter 2. Per contra, the resulting data was later criticized mainly for being not authentic and most likely modern human contamination [153, 168]. In principle, ancient Egyptian mummies are a great study object, also for aDNA studies, since there are thousands of mummies kept at archaeological collections all over the globe and thus could serve as study object relatively easily. However, DNA preservation in Egyptian Mummies was primarily seen as improbable: The climatic conditions,

some chemicals, and oils used to embalm the Mummies are assumed to contribute to DNA degradation and ultimately destruction [53]. Other works using a proxy material such as ancient papyri from Egypt, further nurtured the overall skepticism about DNA preservation [125]. First results obtained from King Tutankhamen [62] were criticized mainly due to missing authenticity and contamination assessment methodology, too [117]. The most current work, published in 2013 [84] as well failed to provide stringent tests for authenticity using damage patterns [17] and detailed contamination tests. Within the scope of the Ancient Egyptian mummy project conducted at the University of Tuebingen, we presented the first data set of ancient Egyptian mummy genomes with a well-established methodology to test for authenticity by assessing DNA damage patterns [17, 218] and applying statistical contamination tests [181]. By systematically evaluating different tissue types, we were able to retrieve aDNA from 151 ancient specimens from Abusir El-Meleq, a small community near the Fayyum oasis in northern Egypt.

We studied the mitochondrial composition of the ancient inhabitants of Abusir El-Meleq, explicitly investigating whether effects of foreign rule or invasion could be found during Ptolemaic (332 — 30 BCE) or Roman rule (30 BCE — 395 BCE)[161]. The aim of the work presented in this chapter was the reconstruction of mitochondrial genomes of ancient Egyptian mummies from Abusir El-Meleq. The nature of NGS DNA sequencing data prohibits a manual analysis of sequencing reads. Specifically, as a total number of 151 samples were investigated, the EAGER pipeline has been applied to reconstruct the mitochondrial genomes of the studied individuals. Furthermore, an early prototype of the mitoBench application has been applied to analyze the reconstructed consensus sequences based on the NGS data that was obtained from the 151 ancient individuals. Besides that, this ancient Egyptian mummy project has been chosen to serve as an exemplary demonstration of the features that EAGER in combination with an early concept of mitoBench can provide to researchers in the field of aDNA. The processed results were used for performing several analyses such as a principal component (PCA) and multi-dimensional scaling (MDS) analysis. Note that the original study as published in Nature Communications also processed the autosomal genomes of three investigated samples, which are not further described in this chapter.

5.2 Challenges from a bioinformatics perspective

Apart from challenges in the laboratory environment, the analysis of ancient Egyptian mummy genome data posed several issues from a bioinformatics perspective. In general, these challenges can be split into two categories: The general analysis hurdles, that were resolved with the application of EAGER and additionally the

subsequent analysis difficulties that were addressed with the mitoBench prototype application.

During the scope of this project, EAGER was used to initially help in screening several hundred shallow sequencing libraries to estimate library complexity and contamination. This in general contained estimating library complexities with Preseq and generating informative library complexity charts, that were then subsequently used to determine which sequencing libraries are worth the investment of deeper as well as targeted sequencing. The inherent challenge here consists of screening hundreds of samples identically and generating reports to allow for a more sophisticated decision in the planning of forthcoming sequencing runs.

The second challenge from a bioinformatics perspective in this project was the combination of various data types (genetic, metadata, radiocarbon dating) in a suitable way. These challenges have been addressed with the application of an early prototype of mitoBench, that was used to combine the different meta information and the mtDNA consensus sequences semi-automatically. The collection of modern reference genomes for downstream population genetics analysis was performed manually by collecting these from previously collected datasets in CSV/TSV and XLS format. The main work required on the bioinformatics side in the project was then primarily to generate required input formats for population genetics methods such as Arlequin and visualize results in R.

The methods and results section follows the differentiation between sequence generation in the laboratory, processing, and authentication with EAGER, the sequence based downstream analysis and finally the frequency based downstream analysis of mitochondrial data using the early prototype of mitoBench.

5.3 Methods

5.3.1 Sequence generation

The experimental design was already used in several other publications [44, 201] and will only be briefly summarized here. All of the investigated samples were extracted in clean room facilities for aDNA work in Tübingen. Samples were first UV irradiated, and the surface of the examined material was removed to account for potential contamination. Afterwards, the DNA extraction was performed by grinding 50 mg material from either bone or teeth and 100 mg from (soft-) tissue respectively. Following a silica purification protocol, 20 μ l of the extract was converted into double-stranded Illumina libraries with a standard protocol [88]. Initially, untargeted whole genome sequencing libraries were created and sequenced using Illumina sequencing to serve as primary shallow screening libraries. Samples that showed promising endogenous DNA content in the screening phase were then enriched for human mitochondrial DNA with a bead capture protocol [124] (see Chapter 2). After purification and quantification with a MinElute PCR purification

kit (Qiagen, Hilden, Germany), the samples were sequenced on an Illumina HiSeq 2500 in paired-end sequencing mode. The autosomal DNA libraries analyzed in the publication in Nature Communications were sequenced on an Illumina NextSeq 500 sequencer in single-end sequencing mode.

5.3.2 Processing and authentication with EAGER

EAGER was utilized multiple times for the different tasks in this project. First, the initial shallow screening libraries were analyzed with EAGER to determine the endogenous DNA content of the investigated samples. Samples showing significant endogenous DNA content ($> 3\%$) were then taken for targeted capture. Resulting FastQ files from paired-end sequencing after mitochondrial capture were then also processed with EAGER. The pipeline was configured to trim off adapters using Clip&Merge, requiring an overlap of 10 nt of both paired reads for merging. Furthermore, a 25 nt minimum read length was enforced, removing any reads shorter than the specific 25 nt cutoff. EAGER was then configured to use the CircularMapper approach with BWA 0.7.12 and settings “-n 0.01 -l 1000” for mapping reads against the GrCH37 reference genome and the RSRS reference genome for human mitochondria. Duplicate reads were removed with DeDup v0.9.10 and QualiMap 2 was used to determine mapping statistics and collect other metrics during processing (Supplementary Table A.5).

As has been mentioned in the introduction of this Chapter, the *in silico* authentication of investigated samples was of essential importance during the project. We set up EAGER to use the MapDamage 2.0 [77] application to determine damage parameters and the schmutzi [181] method to detect potential contamination. We selected a cutoff at 3% detected contamination as found by schmutzi and manually verified that our haplogroups were not changing after the application of schmutzi by applying Haplogrep 2 [233] on both the unmodified consensus and the final consensus sequence obtained by running schmutzi. To further rule out contamination, we performed several filtering experiments with “log2fasta”. This method uses the likelihoods generated by schmutzi to generate a FASTA consensus sequence. On these consensus sequences, we used Haplogrep 2 to determine human mitochondrial haplogroups subsequently.

5.3.3 Sequence based analysis with mitoBench

After consensus sequence generation, the generated FASTA sequences of three ancient Egyptian populations ($n = 90$ samples) were imported in mitoBench. Additionally, a set of $n = 135$ modern Egyptian [99, 155] and $n = 120$ Ethiopian samples [155] were imported as FASTA consensus sequences into mitoBench, too. We converted the entire datasets within mitoBench to “ARP” format and compared the F_{ST} distances of the selected populations with each other using GNU R. We furthermore exported a multi-FASTA file and applied JModeltest 2 [28]

to determine the most appropriate substitution model for our analysis. This was found to be Tamura and Nei [221]) with a γ value of 0.26. The p-values for the calculated F_{ST} distances in Arlequin were corrected for multiple testing with *p.adjust* in Gnu R with the Benjamini Hochberg method [8]. As the intragroup differences were not significant, we merged all three ancient populations in an ancient metapopulation to perform F_{ST} calculation between modern and the ancient metapopulation.

As there were not many complete mitochondrial sequences for multiple populations available from the investigated area in Northern Africa and the Levant, we performed a multidimensional scaling (MDS) analysis of the hypervariable region (HVR-I) sequences in addition to the complete mitochondrial F_{ST} analysis performed on the 345 complete mitochondrial samples. The HVR-I sequences were extracted using a custom Python script and imported into the mitoBench prototype for converting them to the “ARP” format. We then determined optimal substitution parameters (gamma shape 0.26 and Tamura and Nei substitution pattern) using JModeltest 2 again. Using the linearized Slatkin’s F_{ST} values [211], we visualized the HVR-I F_{ST} MDS in a two dimensional plot generated with Gnu R and the *vegan* package. We used Gnu R to visualize the resulting F_{ST} distances on a geographical map.

5.3.4 Frequency based analysis with mitoBench

Using Haplogrep 2 on the exported FASTA file for all samples, we generated an HSD file, containing the mitochondrial haplogroups for all ancient and modern datasets. The resulting HSD file was reimported to mitoBench to annotate the imported datasets with accompanying haplogroups as determined by Haplogrep 2. Following up on that, we defined a set of haplogroups commonly found in the respective area and already used to investigate population changes in previous publications [99, 155]. The selected haplogroups were H, HV, I, J, K, L0-L4, M1, N, R, R0, T, T1, T2, U, W, X and “Other”. The latter consisted of all other haplogroups that were not explicitly placed in the defined haplogroups, following the Phylotree definition v17 [151]. With this information, we ran a summary statistic in mitoBench, counting all haplogroups within our dataset as defined before. We transformed the counts list to a frequency vector of all haplogroups and ran a principal component analysis (PCA) in GNU R 3.2.4 to evaluate relations between individual ancient and modern populations.

5.4 Results

All the initial processing results were processed using EAGER. We investigated, whether the different tissue types probed in the study differ with respect to obtained mean coverages, 3’ and 5’ DNA damage (Figure 5.1). As can be seen in

Figure 5.1A, the obtained coverages on all three tissue types are not significantly different from each other. However, the average damage on both 3' (Figure 5.1B) and 5' (Figure 5.1C) show typically higher DNA damage in mummified soft tissue than teeth or bone material. Overall, DNA damage patterns showed promising damage patterns of 5-49% with an average of 14% damage, indicative of authentic aDNA. After the primary analysis, we evaluated basic quality metrics of all 151

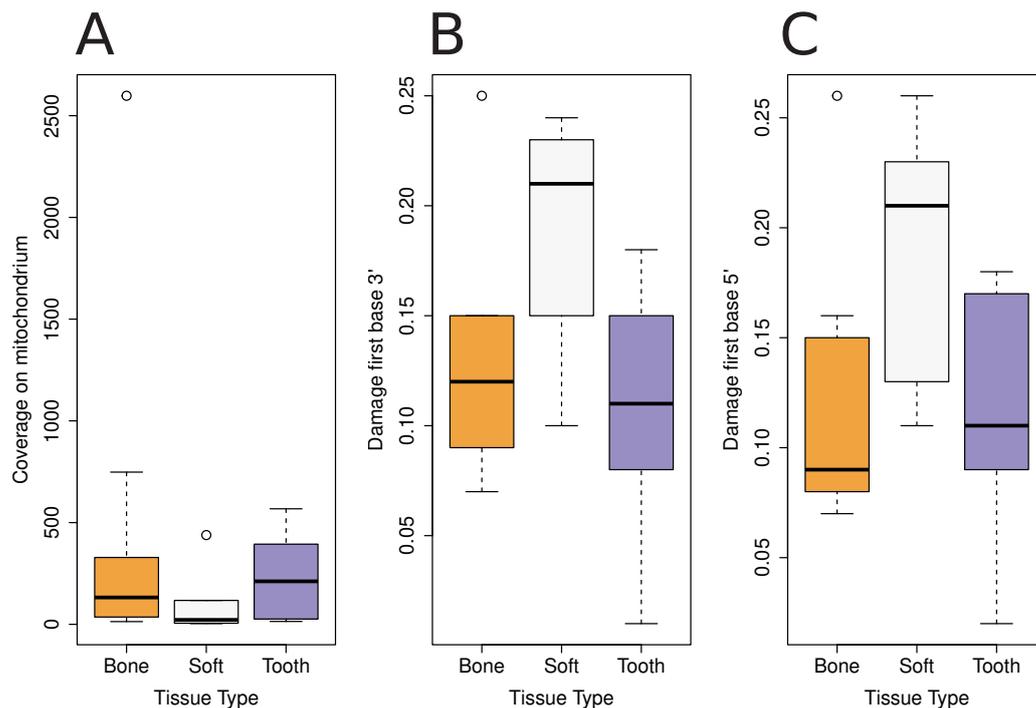


Figure 5.1: Preservation of endogenous DNA in three tissue types (bone, mummified soft tissue and teeth samples). Obtained coverages (A) were largely similar between tissue types. Average damage (B) at the 3' end of each read, also separated by tissue type, was typically higher in mummified soft tissue. Average damage (C) at the 5' end of each read, again separated by tissue type as in A and B, was also higher in mummified soft tissue.

sequenced mitochondrial capture samples. We defined quality requirements of 10-fold mitochondrial coverage at a minimum and a final contamination estimate of at most 3% to consider samples for any further analysis. That resulted in a total sample number of $n = 90$ mitochondrial samples. The 90 mitochondrial genomes were grouped into three categories, based on their radiocarbon dating information. The Pre-Ptolemaic (1388-332BCE) time period contained $n = 44$ samples, the Ptolemaic (332-30BCE) time period $n = 27$ and subsequently $n = 19$ samples were assigned to the Roman time period (30BCE-395CE). We then imported all consensus sequences in mitoBench to perform downstream population genetics analyses.

Sequence based analysis results

For our sequence based analysis results, we generated a stacked haplogroup bar chart to visualize differences between our ancient samples and modern reference populations. We observed very similar haplogroup profiles (Figure 5.2) between all three investigated ancient populations. In contrast to this, we found characteristic African haplogroups (L0-L4) in the modern reference populations that we could not find in all ancient populations. This corresponded with F_{ST} results obtained

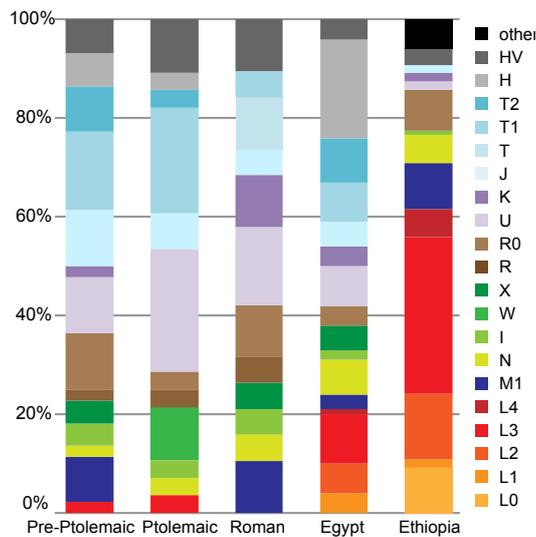


Figure 5.2: Stacked barchart of investigated haplogroups of ancient Egyptian mummy samples and two modern populations. Compared to the three ancient time sections (Pre-Ptolemaic, Ptolemaic, Roman), a drastic increase in characteristic African haplogroups (L0-L4, marked in red and orange tones) in the two modern (Egypt/Ethiopia) populations can be observed.

by running Arlequin on all three ancient populations and comparing them with modern Egyptian [99, 155] and Ethiopian [155] populations (Table 5.1). Furthermore, the increase of African mtDNA lineages L0-L4 to up to 20% in modern populations, as shown in Figure 5.2 matches with hypotheses previously reported by Pagani *et al.* [155]. The geographic visualization of obtained F_{ST} results also shows the highest similarity to populations to populations in Egypt, on the Saudi-Arabian peninsula and other populations from the Levant (Figure 5.4).

Corresponding with the F_{ST} analysis and the hypotheses stated by Pagani *et al.* [155], we found similar results (Figure 5.3), using a multiple dimension scaling (MDS) analysis. As all three ancient populations clustered closely together again, we merged all three populations into a single ancient meta population to increase statistical significance. Again, the MDS plot shows similar characteristics as the previous analyses. The Ancient Egyptians (AEGY) cluster closest to populations from the middle east (Saudi-Arabia, Kuwait, Oman). Ancient Roman

Table 5.1: F_{ST} results of genetic distance computation with Arlequin [36]. The investigated three ancient populations were pooled as the initial F_{ST} analysis suggested very close relatedness among the three ancient populations. The modern populations are from Cairo [155] and the El-Hayez oasis [99] as well as Ethiopia [155]. Shown are (top) the F_{ST} values with the corresponding p-values corrected for multiple testing using Gnu R and a Benjamini Hochberg correction for multiple testing.

F_{ST}	Ancient Egyptians	Egypt	Ethiopia
Ancient Egyptians	-	-	-
Egypt	0.01363	-	-
Ethiopia	0.10257	0.0565	-
p values			
Ancient Egyptians	-	-	-
Egypt	0.0001	-	-
Ethiopia	0.0001	0.0001	-

genomes from Ağlasun (TRO) do not cluster closely with AEGY (see Figure 5.3). We furthermore created a geographic map (Figure 5.4), mapping the calculated HVR-I F_{ST} values to geographical coordinates using Gnu R. The map shows a general trend that ancient Egyptians are most closely related with modern populations from Egypt, the Levant, and the Middle East.

Frequency based analysis results

We created a counts table in mitoBench to investigate changes in the haplogroup composition of the investigated populations. The resulting haplogroup frequencies were imported into R, and the result of the subsequent PCA was plotted (Figure 5.5). Similar to previous analyses on the sequence level, the maternal haplogroup frequencies of the investigated three ancient populations show a trend towards modern populations from the Levantine and the Middle East. The PCA did not give any indication of a closer relationship of the Egyptian samples from the Roman period with other roman period samples from Ağlasun, Turkey [150].

5.5 Discussion

In this chapter, we investigated the mitochondrial genomes of a total of 151 ancient Egyptians from Abusir El-Meleq, applying EAGER and mitoBench jointly in a single project. The methodology presented in this chapter is capable of dealing even with larger sample sizes than previous methods solely relying on manually converting and visualizing mitochondrial DNA samples. We found that EAGER can be used

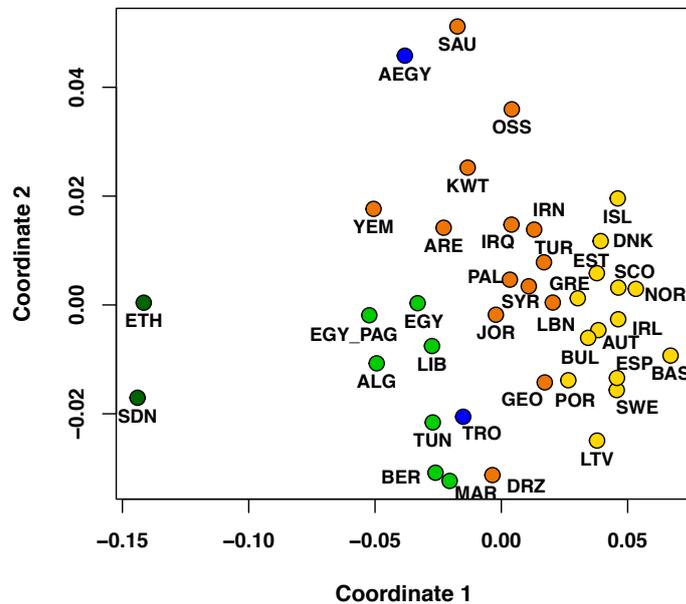


Figure 5.3: MDS of HVR-I data. Populations from Europe are marked in yellow, populations from the Levant and the Middle East in orange, North African populations are marked in green and Ancient populations are marked in blue.

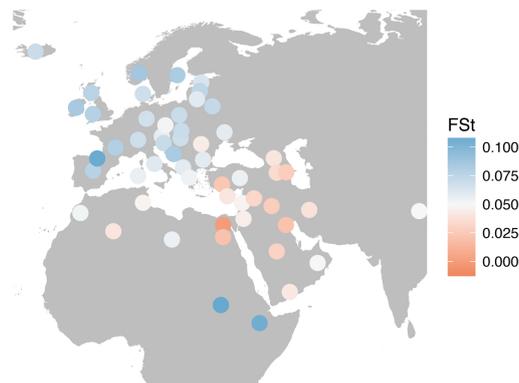


Figure 5.4: A geographic representation of calculated F_{ST} values compared to other populations.

to reconstruct mitochondrial genomes of ancient samples assess DNA contamination automatically using Schmutzi [181]. For instance, the initial 151 samples investigated within the scope of the Egyptian mummy project were screened using EAGER. Of all 151 samples, 61 samples were found to have either no endogenous DNA preserved or to be contaminated substantially, thus resulting in ($n = 90$) samples for the downstream analysis. In such settings, EAGER can provide the

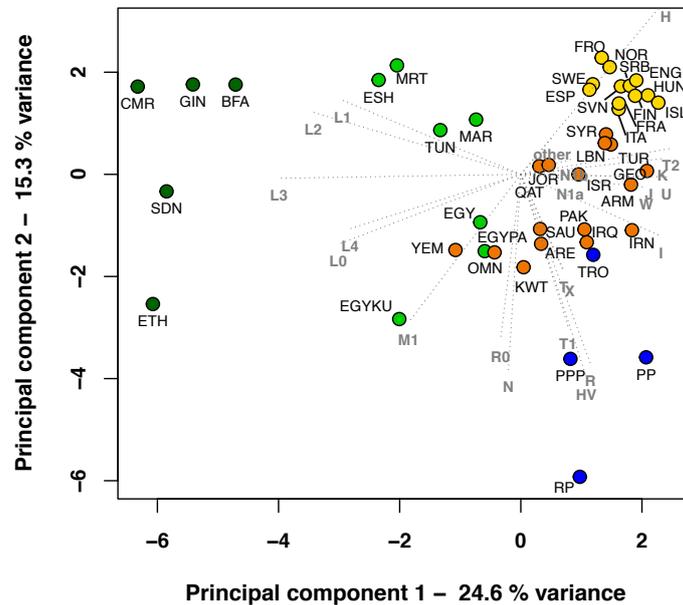


Figure 5.5: PCA plot of haplogroup frequencies. Yellow dots represent European populations, dark green dots African populations. North African populations are represented in light green, and Middle Eastern populations are marked in orange. All three ancient populations (PP=Ptolemaic Period, PPP=Pre-Ptolemaic Period, RP= Roman Period) and the Turkish Romans from Ağlasun [150] were plotted in dark blue. As can be seen, the ancient Egyptian populations show a general trend towards the Levantine and Middle Eastern populations. Country codes follow ISO-3 letter codes and are listed in Supplementary Table A.6.

framework for investigating large sample sizes of experimental screening attempts without requiring users to run individual analysis scripts manually. EAGER is also capable of processing high-density SNP capture experiments of autosomal DNA in a similar fashion, thus not requiring researchers to use several different pipelines or scripts to run their analysis.

Experimentally, we were able to demonstrate that aDNA of Egyptian mummies can be extracted even from material older than 2000 years, by applying enrichment techniques and subsequently using NGS methods. Even for samples where this was previously disputed [53], we were able to reconstruct entire mitochondrial genomes while ensuring that these were not contaminated with more than 3% and typically with much lower contamination values (Supplementary Table A.5). The value of this obtained information is especially interesting, as it can provide additional information other than literary or archaeological data. Previous attempts such as by Pagani *et al.* [155] and hypotheses raised by Henn *et al.* [64] relied on extrapolating information from modern mitochondrial genomes. While these can be useful for many use cases, the assumptions on mutation rates are still solely a proxy to actual events, making any aDNA retrieved from ancient specimens in-

trinsically much more valuable. Abusir El-Meleq as an archaeological site was a good place to investigate the genetic composition of ancient Egypt. There was a chartered variety of interactions with other sites next to the town, including the Fayum oasis and the Memphite provinces [37]. Ultimately, the site appears to have been a main center during the Roman period [26]. Furthermore, the vicinity of the Fayum oasis with substantial growth in population as a result of Greek immigration [26, 178] resulted in further trade and local interactions with other inhabitants of upper Egypt. Several sources furthermore state, that the immigrants formed social bonds and also intermarried with locals [4]. The German excavator Otto Rubensohn, from whom the samples investigated in this study originate from, also provided some written evidence, that foreign Greek influence was present in the area [192]. Unfortunately, lots of the contextual material on sample origins, provenance and excavation diaries were lost during the Second World War [234], making a direct link between investigated specimens and written sources cumbersome or impossible. Though, other sources describe the town to be one of the few sites in Egypt with an extensive chronological spread of specimen available [143]. Furthermore, the site seems to have been used for everyone in the society and not solely for wealthy inhabitants of the town [213]. A further hint that all inhabitants of the town were buried at the site can be seen in the practically decreasing costs for mummification due to the wide availability in the later time periods investigated [143].

Our analysis revealed direct connections between ancient Abusir El-Meleq inhabitants and modern populations from the Near East and the Levant. The previously raised hypotheses by Pagani *et al.* [155], estimating the average proportion of non-African ancestry in Egyptians with 80% were found to be consistent with aDNA results obtained in our study. The results were further unsurprising, given that ancient Egypt had long-standing sociological and trade connections with the Near East. Some of these connections date back to prehistory and include trade, immigration, invasion and other types of connection [217]. Especially during the later time periods, there was increased immigration of, e.g., Canaanite populations and the Hyksos people into Lower Egypt. The origins of these populations are commonly assumed to be in the middle bronze age levant [217], which is consistent with our findings.

We furthermore found a genetic continuity between our Pre-Ptolemaic, Ptolemaic and Roman time periods in Abusir El-Meleq, showing no larger impact of foreign rule in the town. An explanation of this could be, that the genetic impact of foreign invaders such as Greeks and Romans was much higher in the northwestern delta and the important Fayum oasis [178, 185] or in higher castes of the Egyptian population [185]. As there is written material available, describing intermarriage between northern and southern states such as Nubia [161, 213, 214], it is furthermore likely that the African maternal lineages were present in southern regions of the Egyptian empire. Additional studies on ancient human remains from other sites

Chapter 5. Investigating north African population structure

in Egypt are needed to test such geographical, social and chronological variation in more detail in the future.

In this project, we focused on the analysis of ancient mitochondrial genomes from Egypt. We demonstrated the general applicability of both the EAGER pipeline and mitoBench for a population genetics analysis of such data. Additionally, our mtDNA findings correlated strongly with findings on the autosomal level for three nuclear samples which were investigated as part of Stephan Schiffels work in our publication. This suggests, that mtDNA can still be successfully used to determine important population genetics patterns from aDNA data. The implemented prototype application mitoBench will be substantially extended as part of the Ph.D. work of Judith Neukamm (as explained in Chapter 4), incorporating more and more features required for the analysis of projects as investigated here. Thus, users will not need to consult different downstream analysis tools such as Arlequin or R for running statistical methods or visualizing their results soon. As we have seen before, the integrated usage of EAGER and mitoBench can decrease the time required for ancient genome analysis tremendously, while ensuring that the investigated samples are treated with state-of-the-art analysis methods. Furthermore, this reduces the potential for user errors substantially, making the analysis more stable than relying on single individual tools instead.

CHAPTER 6

Discussion

The number of challenges for bioinformatics is constantly rising due to NGS technologies and, consequently, the increasing amount of data. Moreover, studies in the context of aDNA and evolutionary biology are now incorporating increasingly more samples in individual projects to conduct population genetics studies, for example. The inherent challenges for bioinformatics are the development of methods to assist biologists or archaeologists to derive knowledge from obtained data. These require enabling data processing from a variety of NGS technologies in a standardized and well-established way. Furthermore, researchers require reproducible methods that can be used to reassess projects at any given point in time.

Aforementioned technological advances require bioinformatics tools to be developed in a modern way to be able to analyze data quickly and efficiently. A general flexibility and modularization of modern methods can ensure that the tools can be adapted to novel challenges or integrate methods that were newly published. Moreover, projects involving information from different fields of research require new methods to enable an integrative analysis of data, for example, mitochondrial data accompanied with metainformation from linguistics. Achieving all these requirements is a challenging process and can only be accomplished by leveraging modern technological design principles and efficient methodologies in software development.

The main contribution of my Ph.D. was the development of EAGER, a flexible, extendable and efficient software pipeline for the analysis of ancient genome data. EAGER aims at resolving many of the previously mentioned challenges and the standardization of these analysis questions. EAGER achieves:

- a state of the art analysis due to the integration of current methods and software tools

Chapter 6. Discussion

- a more efficient and sensitive analysis thanks to the addition of several applications and tools that were developed to improve results
- a higher usability of aDNA analysis methods by providing a user-friendly GUI and accompanying documentation
- an improved reproducibility by providing containerized environments with all dependencies of the pipeline

Furthermore, EAGER and its accompanying applications have been developed relying on current best-practices in software development to ensure the applications run properly on the targeted platforms. Most importantly, software tests and continuous integration services were added to ensure the ongoing development is improving the pipeline and its diverse applications. The major aim of the entire EAGER project was the creation of a one-stop solution aDNA analysis, which has been achieved. As such, EAGER supports a wide range of different analysis types, including support for paired-end and single-end data, various SNP capture protocols and mitochondrial capture data on humans. Furthermore, the pipeline can analyze bacterial data and incorporates modules to obtain consensus sequences for called variants of bacterial genomes. The module system exposes the processing ideas in the background to the end users by reflecting the basic modules on the processing side in the GUI. This results in a better comprehension of the pipeline for non-bioinformatics researchers. The GUI also provides a consistent user experience: All modules are easy to understand and hide most of the advanced functionality in the UI, only exposing the most important configuration possibilities to the user. Simultaneously, more advanced users can still configure modules in an advanced way, thus making the pipeline more flexible for non-standard research questions. The method has been proven to work well in some projects and all of the following projects have been analyzed successfully with EAGER: The data analysis on the George Bähr project [163], the Syphilis project [6], and the ancient Egyptian mummy project [201]. With a current total of 37 citations¹, there are several external users also utilizing the pipeline for data analysis in an aDNA context.

As the pipeline is widely applied now and offers several possibilities to add modules both for the processing part and in the GUI quite easily, the implementation of new methods is straightforward. Typically, this requires only about ten lines of code to, for example, create a new module in the CLI component of EAGER. These functions and the modularity of EAGER enable a future integration of novel methods, therefore ensuring that the pipeline can be updated easily.

The second part of this dissertation is motivated by an analysis project on ancient Egyptian mummies [201]. This resulted in the design and conceptualization of a software tool for the analysis of mtDNA data (mitoBench) and the accompanying

¹Google Scholar, 2018-02-20

6.1. *EAGER, an efficient ancient genome reconstruction pipeline*

database mitoDB. Both mitoBench and mitoDB aim at resolving several issues in mtDNA population genetics:

- an efficient handling of consensus sequences from mtDNA
- an automatic way to perform file conversions to various downstream population genetics formats without manual user interaction and thus reduced error-rates
- a single one-stop solution for integrative data analysis for mtDNA: data import, export, visualization and direct analysis
- a standardized database of reference panel data with metadata

In combination, both methods will allow users to perform their analyses with much lower risks for human errors: Wrong file names, the characteristic file conversion chaos, and export/import issues for the analysis of mtDNA data are resolved already. A more sophisticated web interface and the main database that incorporates more of the public data available for human mtDNA will be implemented in the near upcoming future within the scope of Judith Neukamms Ph.D. thesis.

In the following subsections, the individual contributions are discussed in detail, followed by a general outlook and conclusion.

6.1 **EAGER, an efficient ancient genome reconstruction pipeline**

As we have seen in Chapter 2, the analysis of aDNA poses several challenges for researchers in bioinformatics and computational biology. Foremost, the processes starting the DNA degradation at the death of an organism introduce lots of issues for a successful analysis both on an experimental and an in-silico level. The typical aDNA damage patterns combined with the fragmentation of the DNA fragments pose themselves severe problems to researchers. Furthermore, the situation is deteriorated even further by the low endogenous DNA content and thus the high contention of aDNA and modern DNA in samples. Methods that can deal with NGS data from modern applications are typically unable to cope with such types of data and thus are inapt to analyze aDNA with its specific requirements.

During this thesis, EAGER was introduced as a novel software solution for the automated and efficient analysis of aDNA data. As detailed in Chapter 3, EAGER can be used to process aDNA data by incorporating methods to perform basic QC, adapter clipping and read merging in a preprocessing phase. Following up on that, EAGER can perform read mapping, deduplication and generate a set of important quality metrics for mapped data. Ultimately, the pipeline is capable of performing state of the art genotyping, by applying best practices in an automated fashion.

Chapter 6. Discussion

One of the biggest achievements of EAGER is the possibility to configure all of the analysis with a simple GUI and thus hiding most of the complexity of the analysis processing from end users. Especially in a highly intradisciplinary context, such as the aDNA community, this can enable users without a bigger background in bioinformatics to configure and run the pipeline. The pipeline does not only feature a GUI but also integrates well-tested and commonly used applications and methods and integrated these in an automated fashion. Several improved applications have been developed during the scope of this thesis to even improve on current solutions for aDNA analysis, amongst which the DeDup and CircularMapper applications are the most prominent cases.

In particular, the problem of read deduplication in aDNA projects was referenced to in Chapter 3.7. Although there are several applications available, such as Picard MarkDuplicates [30] or the SAMTools `rmdup` [110] method, there are no published methods available to perform read deduplication for previously merged reads. As a matter of fact, during the scope of the EAGER project, the DeDup application has been developed that achieves improved deduplication performance in comparison to the previously introduced methods by taking both read ends of a merged read into account. This way, the DeDup method ensures that only true duplicates are removed, which improves the overall coverage of aDNA experiments.

A further issue that has been particularly addressed within the scope of EAGER was the problem of read mapping on circular genomes. Most mapping methods such as BWA [109] or Bowtie 2 [102] are unable to consistently map reads “around” the ends of a circular reference genome. Within EAGER, a novel tool that can handle read mapping on circular reference genomes by implementing an “extend & split” approach, as illustrated in Chapter 3.2.4, was integrated. The approach achieves significantly improved mapping performances at the ends of circular reference genomes as has been demonstrated in Chapter 3.2.4.

Apart from the improved performance, the pipeline demonstrates that many aDNA projects can benefit from automation. Users in aDNA projects today can create a configuration for their specific research analysis project and then subsequently run the created configuration file on their research workstation or cluster. Compared to previous solutions with simple bash scripts, this provides a much more cutting edge solution for analysis needs in an aDNA context. The pipeline is also in a state that future improvements can be added successively, due to current documentation and regular updates.

The applicability of EAGER was demonstrated in the George Bähr project within this thesis. The project was chosen, as it presents a real-world use case of the pipeline and several types of analysis were conducted within it, making it a perfect exemplary project to demonstrate the pipeline’s possibilities. Another paper on ancient Egyptian mummy genomes was furthermore used to demonstrate the capabilities of the pipeline in its current state. Motivated by the successful work that was already achieved using the current method, further developments in the context

6.2. *MitoBench & mitoDB: Modern methods for mitochondrial genome analysis*

of EAGER were started by several authors: Judith Neukamm started adapting the DamageProfiler application for general usage on aDNA and plans to incorporate mathematical models for single-stranded library preparation protocols. Alexander Seitz already implemented a successor tool to compile results from an EAGER run and, for example, run a phylogenetic tree reconstruction on a set of EAGER samples in his EAGER-Tools method². In Jena, the Max Planck Institute for the Science of Human History relies on EAGER for many daily analysis questions and currently analyzes up to hundreds of samples³ per month with the pipeline. To wrap up, the pipeline development steps taken during this thesis have already greatly shaped the way how aDNA samples are analyzed in bioinformatics today and will continue to do this in the upcoming future.

6.2 MitoBench & mitoDB: Modern methods for mitochondrial genome analysis

While EAGER resolves certain issues in aDNA analysis and enables a reproducible data analysis of aDNA data, there are still open issues with respect to data handling, storage and standardization in population genetics. From a bioinformatics perspective, data handling, storage and standardization will become increasingly more important in the next years. This is mostly due to increasing sample sizes wherever NGS technologies are applied. In addition to this, the number of tools requiring different file formats and thus conversion of input formats are creating a large potential for human-introduced errors. Furthermore, another problem is the integration of metadata that has been produced in a different context and should be co-analyzed with genetic information, for example, linguistic or archaeological data. The third type of issue in current population genetics is the acquisition of comparative data, which is nowadays mostly done by creating large collections of sample data in Excel that can be cumbersome to curate or even use in a project-based research context.

In this thesis, two new tools have been conceptually introduced to tackle all of the issues above. First, the concept of the mitoBench application aims at providing a single workbench application for the analysis of mtDNA. It introduces several tools for data import from several file formats and is capable of converting imported data and metadata to several output formats required for downstream analysis in population genetics. As such, mitoBench will also provide powerful visualization methods in the future, to enable an integrated exploratory data analysis without relying on other downstream tools. The big achievement the method already provides in the prototype state is that users can use the application to convert their data to various file formats and simultaneously generate an analysis project in the

²Unpublished work

³Personal communication, Stephen Clayton

application. This project can later be used to add data, convert these to other formats and add new information, e.g. after newly published data is available. Practically, the mitoBench will be designed in a way that enables the incorporation of data which has not been anticipated previously by providing import functions for generic data types. This ensures that a solid foundation for future work in mitoBench is created. Combined with the concept of the mitoDB application and providing access to a set of comparative datasets, the mitoBench method will most likely outperform any other population genetics solution available for the analysis of mtDNA. The consistent improvement over previously used Excel sheets will especially come into play when hundreds or even thousands of reference samples can be analyzed jointly in a single analysis project. Further extensions of mitoDB are already planned, incorporating more and larger reference datasets, our hope is that mitoDB will serve as a well-curated database for future applications of both tools.

In conclusion, mitoBench and mitoDB are two interesting concepts that will provide researchers with state of the art methods to run integrative and reproducible analysis tasks within the context of mitochondrial population genetics. Future extensions of the current prototypes will further improve the possibilities of both applications.

6.3 EAGER and mitoBench revealing ancient Egyptian population history

In this dissertation, it has been described how EAGER and mitoBench were applied on a study of 151 mtDNA genomes from ancient Egyptian mummies. The study of ancient Egyptian DNA has been the aim of several previous studies before, as the country provides an ideal setting for studying the population history of Northern Africa due to its prominent location at the joints of the African and Asian continents. Unfortunately, previous attempts to obtain genetic information from ancient Egyptian specimens have failed to provide authentication analyses as well as methods to ensure that generated data was indeed authentic and not human contamination introduced later. In our latest study, we studied the haplogroup composition of a set of 90 authenticated mtDNA genomes obtained from ancient Egyptian mummies unearthed in Abusir El-Meleq in 1905. We aimed to determine whether the haplogroup composition in ancient Egypt significantly changed compared to modern haplogroup distributions found in Egypt as of today. Furthermore, this project provided an ideal setting to evaluate both the EAGER pipeline for basic data processing and the mitoBench prototype application as downstream analysis method within the scope of a real-world analysis project. The application of EAGER was successful, and our obtained results showed damage patterns indicative of authentic ancient DNA, thus fulfilling the requirements for the authenticity

of the investigated samples. Moreover, the EAGER pipeline's integrated Schmutzi authentication method furthermore showed little to no contamination on the 90 mitochondrial samples investigated. This ensured us that the investigated samples were indeed authentic and of ancient origin and human contamination can be excluded. Using a prototype of mitoBench, we then generated haplogroup counts and compared the haplogroup profiles obtained on modern datasets from the investigated region with our ancient samples, ultimately finding similar haplogroup distributions. The only exception to this was that specific African haplogroups L0-L4 were entirely missing in our ancient datasets, indicating that the ancient Egyptian individuals from Abusir El-Meleq were more closely related to people in the Levante and the Near East than to African populations in Sub-Saharan Africa. A PCA analysis of our obtained frequency counts also suggested that the investigated individuals from Abusir El-Meleq share more ancestry with populations from the Near East than with Sub-Saharan people, for example from Ethiopia. Furthermore, our systematic sequence-based analysis of 345 mitochondrial samples (90 ancient, 155 modern Egyptians and 100 modern Ethiopians) revealed that ancient populations clustered closely in a MDS analysis, which was also confirmed by the determined haplogroup distribution. It has to be noted that the current version of mitoBench is unable to produce PCA and MDS plots, but future versions will incorporate the possibilities to calculate F_{ST} directly in the application. These will be able to perform a MDS analysis, as well as to calculate a PCA on the generated haplogroup frequency tables in the application without external dependencies.

The outlined study highlighted the first provenly authentic results of aDNA obtained from ancient Egyptian specimens, where other attempts previously failed in stringency and authentication. As such, it also provides the first reliable insight into the native North African population history of ancient Egypt. Future studies will provide more samples from different locations in Egypt and thus improve the current view of ancient Egyptian population history. All of these insights would, however, not have been possible without the applications of EAGER and mitoBench, which jointly enabled the analysis of the 90 mtDNA genomes of ancient Egyptian mummies.

6.4 Outlook

While EAGER has already integrated multiple methods and improvements over other competitor pipelines and thus features a cutting edge analysis pipeline, there are always possibilities to improve existing solutions. Some of the existing ideas to improve EAGER are considering to switch the pipeline execution component to standardized workflow languages such as Nextflow [31]. This would allow for a more extensible and smaller code base and furthermore add the ability to execute the pipeline on cloud providers such as Amazon AWS. A further aspect of such a switch would be that the integration of Nextflow would allow a community to

take up development efforts while the original application can remain unchanged for legacy reasons. Another idea to extend the pipeline is the inclusion of standard quality check metrics such as MultiQC⁴. Since the GUI of EAGER has not been renovated for some time, there would also be possibilities to improve the pipeline's interface to end users. One possibility would be to utilize JavaFX, similar to what has been planned for mitoBench, or completely switch to a web UI to run the users analysis and adapt this platform to their respective infrastructure more specifically. Some of the ideas mentioned here already led to the development of EAGER-Tools⁵ to collect results from an EAGER run and integrate the analysis of multiple samples commonly. One exemplary use case would be the reconstruction of a phylogenetic tree for a set of individual samples.

Similar to EAGER, there are some planned extensions and concepts for the mitoBench and mitoDB prototype applications. First of all, the current version of mitoBench requires the users to install other applications to calculate subsequent statistics (e.g., by running Arlequin [36] or visualize the results (e.g., with R). Future developments in this direction will provide functionality on its own that enables users to directly calculate several types of genetic distances, such as F_{ST} . Furthermore, the integration of dimensionality reduction methods and accompanying visualization methods, such as PCA, is planned for mitoBench as well. Moreover, novel visualization concepts are planned for mitoBench too, for example, a modified parallel coordinates [74] plot for displaying haplogroup variation between groups. Further planned additions are other methods to calculate and visualize haplotype sharing and perform a MDS analysis on calculated F_{ST} values within the application. Besides that, users typically request features for plotting F_{ST} values on geographical coordinates which is normally a cumbersome process requiring the interaction with (commercial) geo-information software (GIS) tools. A long-term idea might be the integration of mitoBench with measures to perform a time scale, e.g., enabling users to zoom in on a time slice of mitochondrial genomes and visually exploring changes in haplogroup composition through time in a specified area. Together with support for other downstream analysis tools and formats, Judith Neukamm plans to implement many of these functionalities within her Ph.D. thesis.

Concerning mitoDB, there are several ideas for future development efforts, too. First of all, there are plans to enable users to upload private data collections to mitoDB and then subsequently access these via mitoBench and mitoDB without sharing these with other users or individual groups. This would allow a more direct integration of mitoDB in analysis processes, as users would be more eager to upload data during generation and not just after publication which can take months or years. A further idea for future development concerning mitoDB is to provide an entire web-based dashboard for project-based work. The idea consists of the

⁴<https://multiqc.info>, last accessed November 10th 2017

⁵Alexander Seitz, currently unpublished

possibility to create a user, access all data in the database and create individual projects in a web application that can be directly accessed online. This would increase the usability of the software as users would not require any local software to be installed other than a browser. As most people are nowadays well suited to use web browsers, this would further enlarge the user base of mitoDB by a substantial amount. The integration of analysis methods and visualization techniques would furthermore allow us to deploy updated versions of the application directly to end users, without interfering with their daily work or requiring them to download or install the application themselves.

6.5 Conclusion

In the last years, research in the field of aDNA analysis moved from single sample analysis towards the study of dozens to hundreds of samples simultaneously. Due to this development and the rising availability of novel high-throughput technologies, genomic datasets are produced in large numbers at a previously unprecedented speed. During the last four years, EAGER has emerged from a small set of simple bash scripts and a small user interface to one of the leading pipelines for ancient genome reconstruction and processing. EAGER is now widely recognized as a notable academic computational application in this field. Complementing EAGER, the mitoBench and mitoDB applications will target the integration of results that are generated in standardized processing. We hope, that all three applications will aid the population genetics community with standardizing several portions of their analysis to tackle common analysis issues in population genetics on mtDNA. Simultaneously, EAGER and the mitoBench and mitoDB applications try to eradicate the issues of reproducibility [31], standardization of analysis [236] and automation to resolve some of the most prominent issues currently found in bioinformatics and thus more specifically in the field of aDNA research. In this context, both applications provide - particularly in collaboration - powerful and well-combined foundations for future advancement in the field of aDNA research and population genetics to keep up with newer analysis methods that can be easily integrated into both tools.

Chapter 6. Discussion

Bibliography

- [1] 1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, et al. "An integrated map of genetic variation from 1,092 human genomes". en. In: *Nature* 491.7422 (2012), pages 56–65.
- [2] M. E. Allentoft, M. Collins, D. Harker, J. Haile, et al. "The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils". en. In: *Proceedings. Biological sciences / The Royal Society* 279.1748 (2012), pages 4724–4733.
- [3] M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, et al. "Population genomics of Bronze Age Eurasia". en. In: *Nature* 522.7555 (2015), pages 167–172.
- [4] R. Alston. *Soldier and Society in Roman Egypt: A Social History*. en. Routledge, 2002.
- [5] S. Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2010.
- [6] N. Arora, V. J. Schuenemann, G. Jäger, A. Peltzer, et al. "Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster". en. In: *Nature microbiology* 2 (2016), page 16245.
- [7] D. M. Behar, M. van Oven, S. Rosset, M. Metspalu, et al. "A "Copernican" reassessment of the human mitochondrial DNA tree from its root". en. In: *American journal of human genetics* 90.4 (2012), pages 675–684.
- [8] Y. Benjamini and Y. Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B, Statistical methodology* 57.1 (1995), pages 289–300.
- [9] T. Bersaglieri, P. C. Sabeti, N. Patterson, T. Vanderploeg, et al. "Genetic signatures of strong recent positive selection at the lactase gene". en. In: *American journal of human genetics* 74.6 (2004), pages 1111–1120.

Bibliography

- [10] W. Bogdanowicz, M. Allen, W. Branicki, M. Lembring, et al. “Genetic identification of putative remains of the famous astronomer Nicolaus Copernicus”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.30 (2009), pages 12279–12282.
- [11] K. I. Bos, K. M. Harkins, A. Herbig, M. Coscolla, et al. “Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis”. en. In: *Nature* 514.7523 (2014), pages 494–497.
- [12] K. I. Bos, G. Jäger, V. J. Schuenemann, Å. J. Vågene, et al. “Parallel detection of ancient pathogens via array-based DNA capture”. en. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 370.1660 (2015), page 20130375.
- [13] 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”. en. In: *Nature* 478.7370 (2011), pages 506–510.
- [14] S. Brandini, P. Bergamaschi, M. F. Cerna, F. Gandini, et al. “The Paleo-Indian Entry into South America According to Mitogenomes”. en. In: *Molecular biology and evolution* 35.2 (2018), pages 299–311.
- [15] 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”. en. In: *Science* 342.6155 (2013), pages 257–261.
- [16] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, et al. “The potential and challenges of nanopore sequencing”. en. In: *Nature biotechnology* 26.10 (2008), pages 1146–1153.
- [17] A. W. Briggs, U. Stenzel, P. L. F. Johnson, R. E. Green, et al. “Patterns of damage in genomic DNA sequences from a Neandertal”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 104.37 (2007), pages 14616–14621.
- [18] A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, et al. “Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA”. en. In: *Nucleic acids research* 38.6 (2010).
- [19] P. Brotherton, P. Endicott, J. J. Sanchez, M. Beaumont, et al. “Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions”. en. In: *Nucleic acids research* 35.17 (2007), pages 5717–5728.
- [20] Y. Broux. *Double names and elite strategy in Roman Egypt*. Peeters, 2015.
- [21] C. Bycroft, C. Freeman, D. Petkova, G. Band, et al. “Genome-wide genetic data on 500,000 UK Biobank participants”. In: *bioRxiv* (2017), page 166298.
- [22] R. J. Cano, H. N. Poinar, N. J. Pieniasek, A. Acra, and G. O. Poinar Jr. “Amplification and sequencing of DNA from a 120-135-million-year-old weevil”. en. In: *Nature* 363.6429 (1993), pages 536–538.

- [23] M. Cariaso and G. Lennon. “SNPedia: a wiki supporting personal genome annotation, interpretation and analysis”. en. In: *Nucleic acids research* 40.DB issue (2012), pages D1308–12.
- [24] W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*. 2017.
- [25] M. Choi, U. I. Scholl, W. Ji, T. Liu, et al. “Genetic diagnosis by whole exome capture and massively parallel DNA sequencing”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.45 (2009), pages 19096–19101.
- [26] W. Clarysse and D. J. Thompson. *Counting the People in Hellenistic Egypt: Volume 2, Historical Studies*. en. Cambridge University Press, 2006.
- [27] T. Daley and A. D. Smith. “Predicting the molecular complexity of sequencing libraries”. en. In: *Nature methods* 10.4 (2013), pages 325–327.
- [28] D. Darriba, G. L. Taboada, R. Doallo, and D. Posada. “jModelTest 2: more models, new heuristics and parallel computing”. en. In: *Nature methods* 9.8 (2012), page 772.
- [29] C. Darwin. *On the origin of species*. John Murray, London, 1859.
- [30] 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”. en. In: *Nature genetics* 43.5 (2011), pages 491–498.
- [31] P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, et al. “Nextflow enables reproducible computational workflows”. en. In: *Nature biotechnology* 35.4 (2017), pages 316–319.
- [32] P. Di Tommaso, E. Palumbo, M. Chatzou, P. Prieto, et al. “The impact of Docker containers on the performance of genomic pipelines”. en. In: *PeerJ* 3 (2015).
- [33] J. Dissing, J. Binladen, A. Hansen, B. Sejrnsen, et al. “The last Viking King: a royal maternity case solved by ancient DNA analysis”. en. In: *Forensic science international* 166.1 (2007), pages 21–27.
- [34] H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, et al. “Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression”. en. In: *Human genetics* 123.2 (2008), pages 177–187.
- [35] N. S. Enattah, T. Sahi, E. Savilahti, J. D. Terwilliger, et al. “Identification of a variant associated with adult-type hypolactasia”. en. In: *Nature genetics* 30.2 (2002), pages 233–237.
- [36] L. Excoffier and H. E. L. Lischer. “Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows”. en. In: *Molecular ecology resources* 10.3 (2010), pages 564–567.

Bibliography

- [37] M. R. Falivene. “The Herakleopolite Nome”. In: *a Catalogue of the Toponyms. Atlanta* (1998).
- [38] N. R. Faria, E. C. Sabino, M. R. T. Nunes, L. C. J. Alcantara, et al. “Mobile real-time surveillance of Zika virus in Brazil”. en. In: *Genome medicine* 8.1 (2016), page 97.
- [39] L. Feuk, A. R. Carson, and S. W. Scherer. “Structural variation in the human genome”. en. In: *Nature reviews. Genetics* 7.2 (2006), pages 85–97.
- [40] H. Fischer. “Forschungen zu George Bähr und dem sächsischen Barock: I. und II. Teil”. PhD thesis. Dresden, 1967.
- [41] B. Fry. *Visualizing Data: Exploring and Explaining Data with the Processing Environment*. en. “O’Reilly Media, Inc.”, 2007.
- [42] Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, et al. “An early modern human from Romania with a recent Neanderthal ancestor”. en. In: *Nature* 524.7564 (2015), pages 216–219.
- [43] Q. Fu, H. Li, P. Moorjani, F. Jay, et al. “Genome sequence of a 45,000-year-old modern human from western Siberia”. en. In: *Nature* 514.7523 (2014), pages 445–449.
- [44] Q. Fu, M. Meyer, X. Gao, U. Stenzel, et al. “DNA analysis of an early modern human from Tianyuan Cave, China”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.6 (2013), pages 2223–2227.
- [45] Q. Fu, A. Mitnik, P. L. F. Johnson, K. Bos, et al. “A revised timescale for human evolution based on ancient mitochondrial genomes”. en. In: *Current biology: CB* 23.7 (2013), pages 553–559.
- [46] Q. Fu, C. Posth, M. Hajdinjak, M. Petr, et al. “The genetic history of Ice Age Europe”. en. In: *Nature* 534.7606 (2016), pages 200–205.
- [47] A. Fujimoto, R. Kimura, J. Ohashi, K. Omi, et al. “A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness”. en. In: *Human molecular genetics* 17.6 (2008), pages 835–843.
- [48] F. Gandini, A. Achilli, M. Pala, M. Bodner, et al. “Mapping human dispersals into the Horn of Africa from Arabian Ice Age refugia using mitogenomes”. en. In: *Scientific reports* 6 (2016), page 25472.
- [49] M.-T. Gansauge, T. Gerber, I. Glocke, P. Korlevic, et al. “Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase”. en. In: *Nucleic acids research* 45.10 (2017).
- [50] F. García-Alcalde, K. Okonechnikov, J. Carbonell, L. M. Cruz, et al. “Qualimap: evaluating next-generation sequencing alignment data”. en. In: *Bioinformatics* 28.20 (2012), pages 2678–2679.

- [51] J. Gardy, N. J. Loman, and A. Rambaut. “Real-time digital pathogen surveillance—the time is now”. In: *Genome biology* 16.1 (2015), page 155.
- [52] S. Gerlach. *George Bähr: Der Erbauer der Dresdner Frauenkirche: ein Zeitbild.* de. Böhlau Verlag Köln Weimar, 2005.
- [53] 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)”. en. In: *American journal of physical anthropology* 128.1 (2005), pages 110–114.
- [54] S. Goodwin, J. D. McPherson, and W. R. McCombie. “Coming of age: ten years of next-generation sequencing technologies”. en. In: *Nature reviews. Genetics* 17.6 (2016), pages 333–351.
- [55] S. Gopalakrishnan, J. A. Samaniego Castruita, M.-H. S. Sinding, L. F. K. Kuderna, et al. “The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics”. en. In: *BMC genomics* 18.1 (2017), page 495.
- [56] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, et al. “A draft sequence of the Neandertal genome”. en. In: *Science* 328.5979 (2010), pages 710–722.
- [57] R. E. Green, A.-S. Malaspina, J. Krause, A. W. Briggs, et al. “A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing”. en. In: *Cell* 134.3 (2008), pages 416–426.
- [58] W. Haak, I. Lazaridis, N. Patterson, N. Rohland, et al. “Massive migration from the steppe was a source for Indo-European languages in Europe”. en. In: *Nature* 522.7555 (2015), pages 207–211.
- [59] F. Hahne and R. Ivanek. “Visualizing Genomic Data Using Gviz and Bioconductor”. en. In: *Methods in molecular biology* 1418 (2016), pages 335–351.
- [60] H. Hammarström, R. Forkel, and M. Haspelmath. *Glottolog 3.0.* Max Planck Institute for the Science of Human History. Jena, 2017.
- [61] S. Hartmans, J. A. M. de Bont, and E. Stackebrandt. “The Genus *Mycobacterium* Nonmedical”. en. In: *The Prokaryotes*. Edited by M. D. P. Dr., S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt. Springer, New York, NY, 2006, pages 889–918.
- [62] Z. Hawass, Y. Z. Gad, S. Ismail, R. Khairat, et al. “Ancestry and pathology in King Tutankhamun’s family”. en. In: *JAMA: the journal of the American Medical Association* 303.7 (2010), pages 638–647.
- [63] J. B. Hays and B. H. Zimm. “Flexibility and stiffness in nicked DNA”. en. In: *Journal of molecular biology* 48.2 (1970), pages 297–317.

Bibliography

- [64] B. M. Henn, L. R. Botigué, S. Gravel, W. Wang, et al. “Genomic ancestry of North Africans supports back-to-Africa migrations”. en. In: *PLoS genetics* 8.1 (2012), e1002397.
- [65] A. Herbig. “Computational Methods for the Identification and Characterization of Non-Coding RNAs in Bacteria”. PhD thesis. Eberhard Karls Universität Tübingen, 2015.
- [66] M. Herper. *Illumina Promises To Sequence Human Genome For \$100 – But Not Quite Yet*. <https://www.forbes.com/sites/matthewherper/2017/01/09/illumina-promises-to-sequence-human-genome-for-100-but-not-quite-yet/>. Accessed: 2017-9-26. 2017.
- [67] 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”. en. In: *Nature* 312.5991 (1984), pages 282–284.
- [68] E. Hodges, Z. Xuan, V. Balija, M. Kramer, et al. “Genome-wide in situ exon capture for selective resequencing”. en. In: *Nature genetics* 39.12 (2007), pages 1522–1527.
- [69] M. Hofreiter, V Jaenicke, D Serre, A. von Haeseler, and S. Pääbo. “DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA”. en. In: *Nucleic acids research* 29.23 (2001), pages 4793–4799.
- [70] K. E. Holsinger and B. S. Weir. “Genetics in geographically structured populations: defining, estimating and interpreting F(ST)”. en. In: *Nature reviews. Genetics* 10.9 (2009), pages 639–650.
- [71] S.-Y. N. Huang, S. Ghosh, and Y. Pommier. “Topoisomerase I alone is sufficient to produce short DNA deletions and can also reverse nicks at ribonucleotide sites”. en. In: *The Journal of biological chemistry* 290.22 (2015), pages 14068–14076.
- [72] R. R. Hudson, M Slatkin, and W. P. Maddison. “Estimation of levels of gene flow from DNA sequence data”. en. In: *Genetics* 132.2 (1992), pages 583–589.
- [73] E. Huerta-Sánchez, X. Jin, Asan, Z. Bianba, et al. “Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA”. en. In: *Nature* 512.7513 (2014), pages 194–197.
- [74] A. Inselberg. “The plane with parallel coordinates”. en. In: *The Visual computer* 1.2 (1985), pages 69–91.
- [75] K. Ishiya and S. Ueda. “MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing”. en. In: *PeerJ* 5 (2017).
- [76] G. Jäger. “Advanced Visual Analytics Approaches for the Integrative Study of Genomic and Transcriptomic Data”. en. PhD thesis. Eberhard Karls Universität Tübingen, 2016.

- [77] H. Jónsson, A. Ginolhac, M. Schubert, P. L. F. Johnson, and L. Orlando. “mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters”. en. In: *Bioinformatics* 29.13 (2013), pages 1682–1684.
- [78] H. Jónsson, P. Sulem, B. Kehr, S. Kristmundsdottir, et al. “Whole genome characterization of sequence diversity of 15,220 Icelanders”. en. In: *Scientific data* 4 (2017), page 170115.
- [79] K. Kawaguchi, A. Bayer, and R. T. Croy. *Jenkins-an extensible open source continuous integration server*. <http://jenkins.io>. Accessed: 2017-9-26. 2017.
- [80] 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”. en. In: *Nature communications* 3 (2012), page 698.
- [81] B. M. Kemp, C. Monroe, K. G. Judd, E. Reams, and C. Grier. “Evaluation of methods that subdue the effects of polymerase chain reaction inhibitors in the study of ancient and degraded DNA”. In: *Journal of archaeological science* 42.Supplement C (2014), pages 373–380.
- [82] B. M. Kemp, C. Monroe, and D. G. Smith. “Repeat silica extraction: a simple technique for the removal of PCR inhibitors from DNA extracts”. In: *Journal of archaeological science* 33.12 (2006), pages 1680–1689.
- [83] B. M. Kemp and D. G. Smith. “Use of bleach to eliminate contaminating DNA from the surface of bones and teeth”. en. In: *Forensic science international* 154.1 (2005), pages 53–61.
- [84] R. Khairat, M. Ball, C.-C. H. Chang, R. Bianucci, et al. “First insights into the metagenome of Egyptian mummies using next-generation sequencing”. en. In: *Journal of applied genetics* 54.3 (2013), pages 309–325.
- [85] R. Kimura, T. Yamaguchi, M. Takeda, O. Kondo, et al. “A common variation in EDAR is a genetic determinant of shovel-shaped incisors”. en. In: *American journal of human genetics* 85.4 (2009), pages 528–535.
- [86] T. E. King, G. G. Fortes, P. Balaesque, M. G. Thomas, et al. “Identification of the remains of King Richard III”. en. In: *Nature communications* 5 (2014), page 5631.
- [87] M. Kircher. “Analysis of High-Throughput Ancient DNA Sequencing Data”. en. In: *Ancient DNA*. Edited by B. Shapiro and M. Hofreiter. Volume 840. Methods in Molecular Biology. Humana Press, 2012, pages 197–228.
- [88] M. Kircher, S. Sawyer, and M. Meyer. “Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform”. en. In: *Nucleic acids research* 40.1 (2012).

Bibliography

- [89] L. Kistler, R. Ware, O. Smith, M. Collins, and R. G. Allaby. “A new model for ancient DNA decay based on paleogenomic meta-analysis”. en. In: *Nucleic acids research* 45.11 (2017), pages 6310–6320.
- [90] M. Knapp and M. Hofreiter. “Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives”. en. In: *Genes* 1.2 (2010), pages 227–243.
- [91] Kolossos. *Front view of Dresden Frauenkirche*. <https://commons.wikimedia.org/wiki/File:Dresden-Frauenkirche-night.jpg>. Accessed: 2017-12-2.
- [92] P. Korlević, T. Gerber, M.-T. Gansauge, M. Hajdinjak, et al. “Reducing microbial and human contamination in DNA extractions from ancient bones and teeth”. en. In: *BioTechniques* 59.2 (2015), pages 87–93.
- [93] T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. “ANGSD: Analysis of Next Generation Sequencing Data”. en. In: *BMC bioinformatics* 15.1 (2014), page 356.
- [94] A. Koschmieder, K. Zimmermann, S. Trissl, T. Stoltmann, and U. Leser. “Tools for managing and analyzing microarray data”. en. In: *Briefings in bioinformatics* 13.1 (2012), pages 46–60.
- [95] J. Krause, A. W. Briggs, M. Kircher, T. Maricic, et al. “A complete mtDNA genome of an early modern human from Kostenki, Russia”. en. In: *Current biology: CB* 20.3 (2010), pages 231–236.
- [96] J. Krause, C. Lalueza-Fox, L. Orlando, W. Enard, et al. “The derived FOXP2 variant of modern humans was shared with Neandertals”. en. In: *Current biology: CB* 17.21 (2007), pages 1908–1912.
- [97] J. Krause and S. Pääbo. “Genetic Time Travel”. en. In: *Genetics* 203.1 (2016), pages 9–12.
- [98] S. Krüger, F. Battke, A. Sprecher, M. Munz, et al. “Rare Variants in Neurodegeneration Associated Genes Revealed by Targeted Panel Sequencing in a German ALS Cohort”. en. In: *Frontiers in molecular neuroscience* 9 (2016), page 92.
- [99] M Kujanova, L Pereira, Fernandes, J. B. Pereira, and V Cerný. “Near Eastern Neolithic genetic input in a small oasis of the Egyptian Western Desert”. In: *American Journal of Physical Anthropology* 140.2 (2009), pages 336–346.
- [100] G. M. Kurtzer, V. Sochat, and M. W. Bauer. “Singularity: Scientific containers for mobility of compute”. en. In: *PloS one* 12.5 (2017).
- [101] E. S. Lander, L. M. Linton, B Birren, C Nusbaum, et al. “Initial sequencing and analysis of the human genome”. en. In: *Nature* 409.6822 (2001), pages 860–921.

- [102] B. Langmead and S. L. Salzberg. “Fast gapped-read alignment with Bowtie 2”. en. In: *Nature methods* 9.4 (2012), pages 357–359.
- [103] M. H. D. Larmuseau, B. Bekaert, M. Baumers, T. Wenseleers, et al. “Bio-historical materials and contemporary privacy concerns—the forensic case of King Albert I”. en. In: *Forensic science international. Genetics* 24 (2016), pages 202–210.
- [104] I. Lazaridis, A. Mittnik, N. Patterson, S. Mallick, et al. “Genetic origins of the Minoans and Mycenaeans”. en. In: *Nature* 548.7666 (2017), pages 214–218.
- [105] I. Lazaridis, D. Nadel, G. Rollefson, D. C. Merrett, et al. “Genomic insights into the origin of farming in the ancient Near East”. en. In: *Nature* 536.7617 (2016), pages 419–424.
- [106] I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, et al. “Ancient human genomes suggest three ancestral populations for present-day Europeans”. en. In: *Nature* 513.7518 (2014), pages 409–413.
- [107] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, et al. “Analysis of protein-coding genetic variation in 60,706 humans”. en. In: *Nature* 536.7616 (2016), pages 285–291.
- [108] H. Li. “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”. In: (2013). arXiv: 1303.3997 [q-bio.GN].
- [109] H. Li and R. Durbin. “Fast and accurate short read alignment with Burrows-Wheeler transform”. en. In: *Bioinformatics* 25.14 (2009), pages 1754–1760.
- [110] H. Li, B. Handsaker, A. Wysoker, T. Fennell, et al. “The Sequence Alignment/Map format and SAMtools”. en. In: *Bioinformatics* 25.16 (2009), pages 2078–2079.
- [111] R. Li, H. Zhu, J. Ruan, W. Qian, et al. “De novo assembly of human genomes with massively parallel short read sequencing”. en. In: *Genome research* 20.2 (2010), pages 265–272.
- [112] V. Link, A. Kousathanas, K. Veeramah, C. Sell, et al. “ATLAS: Analysis Tools for Low-depth and Ancient Samples”. en. 2017.
- [113] M. Lipson, A. Szécsényi-Nagy, S. Mallick, A. Pósa, et al. “Parallel palaeogenomic transects reveal complex genetic history of early European farmers”. en. In: *Nature* 551.7680 (2017), pages 368–372.
- [114] M. List, P. Ebert, and F. Albrecht. “Ten Simple Rules for Developing Usable Software in Computational Biology”. en. In: *PLoS computational biology* 13.1 (2017). Edited by S. Markel.

Bibliography

- [115] M. List, M. P. Elnegaard, S. Schmidt, H. Christiansen, et al. "Efficient Management of High-Throughput Screening Libraries with SAVANAH". en. In: *SLAS discovery : advancing life sciences R & D* 22.2 (2017), pages 196–202.
- [116] 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". en. In: *Journal of molecular evolution* 39.2 (1994), pages 174–190.
- [117] E. D. Lorenzen and E. Willerslev. "King Tutankhamun's family and demise". en. In: *JAMA: the journal of the American Medical Association* 303.24 (2010), page 2471.
- [118] M. T. Lott, J. N. Leipzig, O. Derbeneva, H. M. Xie, et al. "mtDNA Variation and Analysis Using Mitomap and Mitomaster". en. In: *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]* 44 (2013), pages 1.23.1–26.
- [119] M. Z. Ludwig. "Functional evolution of noncoding DNA". en. In: *Current opinion in genetics & development* 12.6 (2002), pages 634–639.
- [120] G. Lunter and M. Goodson. "Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads". en. In: *Genome research* 21.6 (2011), pages 936–939.
- [121] H. Magirius. *Die Dresdner Frauenkirche von George Bähr: Entstehung und Bedeutung*. Deutscher Verlag für Kunstwissenschaft, 2005.
- [122] A.-S. Malaspinas, M. C. Westaway, C. Muller, V. C. Sousa, et al. "A genomic history of Aboriginal Australia". en. In: *Nature* 538.7624 (2016), pages 207–214.
- [123] L. Maretty, J. M. Jensen, B. Petersen, J. A. Sibbesen, et al. "Sequencing and de novo assembly of 150 genomes from Denmark as a population reference". en. In: *Nature* 548.7665 (2017), pages 87–91.
- [124] T. Maricic, M. Whitten, and S. Pääbo. "Multiplexed DNA sequence capture of mitochondrial genomes using PCR products". en. In: *PloS one* 5.11 (2010). Edited by R. C. Fleischer.
- [125] I. Marota, C. Basile, M. Ubaldi, and F. Rollo. "DNA decay rate in papyri and human remains from Egyptian archaeological sites". en. In: *American journal of physical anthropology* 117.4 (2002), pages 310–318.
- [126] C. D. Matheson, T. E. Marion, S. Hayter, N. Esau, et al. "Technical note: removal of metal ion inhibition encountered during DNA extraction and amplification of copper-preserved archaeological bone using size exclusion chromatography". en. In: *American journal of physical anthropology* 140.2 (2009), pages 384–391.

- [127] I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, et al. “The genomic history of southeastern Europe”. en. In: *Nature* (2018).
- [128] I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, et al. “Genome-wide patterns of selection in 230 ancient Eurasians”. en. In: *Nature* 528.7583 (2015), pages 499–503.
- [129] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, et al. “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”. en. In: *Genome research* 20.9 (2010), pages 1297–1303.
- [130] D. Merkel. “Docker: Lightweight Linux Containers for Consistent Development and Deployment”. In: *Linux Journal* 2014.239 (2014), page 2.
- [131] M. L. Metzker. “Sequencing technologies - the next generation”. en. In: *Nature reviews. Genetics* 11.1 (2010), pages 31–46.
- [132] M. Meyer, J.-L. Arsuaga, C. de Filippo, S. Nagel, et al. “Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins”. en. In: *Nature* 531.7595 (2016), pages 504–507.
- [133] M. Meyer, Q. Fu, A. Aximu-Petri, I. Glocke, et al. “A mitochondrial genome sequence of a hominin from Sima de los Huesos”. en. In: *Nature* 505.7483 (2014), pages 403–406.
- [134] M. Meyer and M. Kircher. “Illumina sequencing library preparation for highly multiplexed target capture and sequencing”. en. In: *Cold Spring Harbor protocols* 2010.6 (2010).
- [135] A. Mittnik, C.-C. Wang, J. Svoboda, and J. Krause. “A Molecular Approach to the Sexing of the Triple Burial at the Upper Paleolithic Site of Dolní Věstonice”. en. In: *PloS one* 11.10 (2016). Edited by F. Calafell, e0163019.
- [136] W. Möllering. *George Bähr, ein protestantischer Kirchenbaumeister des Barock*. Leipzig: Frommhold & Wendler, 1933.
- [137] D. Moreira. “Efficient removal of PCR inhibitors using agarose-embedded DNA preparations”. en. In: *Nucleic acids research* 26.13 (1998), pages 3309–3310.
- [138] N. M. Myres, S. Rootsi, A. A. Lin, M. Järve, et al. “A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe”. en. In: *European journal of human genetics: EJHG* 19.1 (2011), pages 95–101.
- [139] R. A. Neher and T. Bedford. “nextflu: real-time tracking of seasonal influenza virus evolution in humans”. en. In: *Bioinformatics* 31.21 (2015), pages 3546–3548.

Bibliography

- [140] R. A. Neher, T. Bedford, R. S. Daniels, C. A. Russell, and B. I. Shraiman. "Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses". en. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.12 (2016), E1701–9.
- [141] J. Neukamm. "Testing framework and improved algorithms for ancient genome reconstruction". en. Master's thesis. Universität Tübingen, 2015.
- [142] J. Neukamm and A. Peltzer. *DamageProfiler*. Unpublished. 2017.
- [143] P. T. Nicholson and I. Shaw. *Ancient Egyptian Materials and Technology*. en. Cambridge University Press, 2000.
- [144] J. P. Noonan, M. Hofreiter, D. Smith, J. R. Priest, et al. "Genomic sequencing of Pleistocene cave bears". en. In: *Science* 309.5734 (2005), pages 597–599.
- [145] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, et al. "Genes mirror geography within Europe". en. In: *Nature* 456.7218 (2008), pages 98–101.
- [146] K. Okonechnikov, A. Conesa, and F. García-Alcalde. "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data". en. In: *Bioinformatics* 32.2 (2016), pages 292–294.
- [147] I. Olalde, S. Brace, M. E. Allentoft, I. Armit, et al. "The Beaker phenomenon and the genomic transformation of northwest Europe". en. In: *Nature* (2018).
- [148] A. Olivieri, C. Sidore, A. Achilli, A. Angius, et al. "Mitogenome Diversity in Sardinians: A Genetic Window onto an Island's Past". en. In: *Molecular biology and evolution* 34.5 (2017), pages 1230–1239.
- [149] L. Orlando, A. Ginolhac, G. Zhang, D. Froese, et al. "Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse". en. In: *Nature* 499.7456 (2013), pages 74–78.
- [150] C. Ottoni, R. Rasteiro, R. Willet, J. Claeys, et al. "Comparing maternal genetic variation across two millennia reveals the demographic history of an ancient human population in southwest Turkey". en. In: *Royal Society open science* 3.2 (2016), page 150250.
- [151] M. van Oven. "PhyloTree Build 17: Growing the human mitochondrial DNA tree". In: *Forensic Science International: Genetics Supplement Series* 5. Supplement C (2015), e392–e394.
- [152] S Pääbo. "Molecular cloning of Ancient Egyptian mummy DNA". en. In: *Nature* 314.6012 (1985), pages 644–645.
- [153] S Pääbo. "Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification". en. In: *Proceedings of the National Academy of Sciences of the United States of America* 86.6 (1989), pages 1939–1943.

- [154] L. Pagani, D. J. Lawson, E. Jagoda, A. Mörseburg, et al. “Genomic analyses inform on migration events during the peopling of Eurasia”. en. In: *Nature* 538.7624 (2016), pages 238–242.
- [155] 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”. en. In: *American journal of human genetics* 96.6 (2015), pages 986–991.
- [156] W. Parson, C. Berger, T. Sängler, and S. Lutz-Bonengel. “Molecular genetic analysis on the remains of the Dark Countess: Revisiting the French Royal family”. en. In: *Forensic science international. Genetics* 19 (2015), pages 252–254.
- [157] W. Parson and A. Dür. “EMPOP—a forensic mtDNA database”. en. In: *Forensic science international. Genetics* 1.2 (2007), pages 88–92.
- [158] Patrícia Pečnerová. *The Hype Cycle of Ancient DNA*. <http://www.molecularecologist.com/2017/04/the-hype-cycle-of-ancient-dna/>. Accessed: 2017-10-3. 2017.
- [159] N. Patterson, A. L. Price, and D. Reich. “Population structure and eigenanalysis”. en. In: *PLoS genetics* 2.12 (2006).
- [160] P. Pečnerová, D. Díez-Del-Molino, N. Dussex, T. Feuerborn, et al. “Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth”. en. In: *Current biology: CB* 27.22 (2017), pages 3505–3510.
- [161] W. P. van Pelt. “Revising Egypto-Nubian Relations in New Kingdom Lower Nubia: From Egyptianization to Cultural Entanglement”. In: *Cambridge Archaeological Journal* 23.3 (2013), pages 523–550.
- [162] A. Peltzer, G. Jäger, A. Herbig, A. Seitz, et al. “EAGER: efficient ancient genome reconstruction”. en. In: *Genome biology* 17.1 (2016), page 60.
- [163] A. Peltzer, A. Mittnik, C.-C. Wang, T. Begg, et al. “Inferring genetic origins and phenotypic traits of George Bähr, the architect of the Dresden Frauenkirche”. en. In: *Scientific reports* 8.1 (2018), page 2115.
- [164] E. Pennisi. “Genomics. ENCODE project writes eulogy for junk DNA”. en. In: *Science* 337.6099 (2012), pages 1159, 1161.
- [165] E. Pilli, C. L. Fox, C. Capelli, M. Lari, et al. “Ancient DNA and forensics genetics: The case of Francesco Petrarca”. en. In: *Forensic Science International: Genetics Supplement Series* 1.1 (2008), pages 469–470.
- [166] R. Pinhasi, D. Fernandes, K. Sirak, M. Novak, et al. “Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone”. en. In: *PLoS one* 10.6 (2015).

Bibliography

- [167] M. Pinto, V. Borges, M. Antelo, M. Pinheiro, et al. "Genome-scale analysis of the non-cultivable *Treponema pallidum* reveals extensive within-patient genetic variation". en. In: *Nature microbiology* 2 (2016), page 16190.
- [168] H. N. Poinar and A Cooper. "Ancient DNA: do it right or not at all". In: *Science* 5482.1139 (2000), page 416.
- [169] 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". en. In: *Nature communications* 8 (2017), page 16046.
- [170] K. Prüfer, U. Stenzel, M. Hofreiter, S. Pääbo, et al. "Computational challenges in the analysis of ancient DNA". en. In: *Genome biology* 11.5 (2010).
- [171] J. Quick, N. J. Loman, S. Duraffour, J. T. Simpson, et al. "Real-time, portable genome sequencing for Ebola surveillance". en. In: *Nature* 530.7589 (2016), pages 228–232.
- [172] A. R. Quinlan and I. M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". en. In: *Bioinformatics* 26.6 (2010), pages 841–842.
- [173] F. Racimo, G. Renaud, and M. Slatkin. "Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans". en. In: *PLoS genetics* 12.11 (2016).
- [174] M Ragan-Kelley, F Perez, B Granger, T Kluver, et al. "The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication". In: *AGU Fall Meeting Abstracts*. Volume 1. 2014, page 7.
- [175] M. Raghavan, M. DeGiorgio, A. Albrechtsen, I. Moltke, et al. "The genetic prehistory of the New World Arctic". en. In: *Science* 345.6200 (2014), page 1255832.
- [176] M. Rasmussen, X. Guo, Y. Wang, K. E. Lohmueller, et al. "An Aboriginal Australian genome reveals separate human dispersals into Asia". en. In: *Science* 334.6052 (2011), pages 94–98.
- [177] M. Rasmussen, Y. Li, S. Lindgreen, J. S. Pedersen, et al. "Ancient human genome sequence of an extinct Palaeo-Eskimo". en. In: *Nature* 463.7282 (2010), pages 757–762.
- [178] D. W. Rathbone. "Villages, land and population in Graeco-Roman Egypt". In: *The Cambridge Classical Journal* 36 (1990), pages 103–142.
- [179] D. Reich, R. E. Green, M. Kircher, J. Krause, et al. "Genetic history of an archaic hominin group from Denisova Cave in Siberia". en. In: *Nature* 468.7327 (2010), pages 1053–1060.

- [180] G. Renaud, M. Kircher, U. Stenzel, and J. Kelso. “freelbis: an efficient basecaller with calibrated quality scores for Illumina sequencers”. en. In: *Bioinformatics* 29.9 (2013), pages 1208–1209.
- [181] G. Renaud, V. Slon, A. T. Duggan, and J. Kelso. “Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA”. en. In: *Genome biology* 16.1 (2015), page 224.
- [182] G. Renaud, U. Stenzel, T. Maricic, V. Wiebe, and J. Kelso. “deML: robust demultiplexing of Illumina sequences using a likelihood-based approach”. en. In: *Bioinformatics* 31.5 (2015), pages 770–772.
- [183] A. Rhoads and K. F. Au. “PacBio Sequencing and Its Applications”. en. In: *Genomics, proteomics & bioinformatics* 13.5 (2015), pages 278–289.
- [184] C. Riggs. *The Beautiful Burial in Roman Egypt: Art, Identity, and Funerary Religion*. en. OUP Oxford, 2005.
- [185] C. Riggs. *The Oxford Handbook of Roman Egypt*. en. OUP Oxford, 2012.
- [186] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, et al. “Integrative genomics viewer”. en. In: *Nature biotechnology* 29.1 (2011), pages 24–26.
- [187] E. I. Rogaev, A. P. Grigorenko, Y. K. Moliaka, G. Faskhutdinova, et al. “Genomic identification in the historical case of the Nicholas II royal family”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.13 (2009), pages 5258–5263.
- [188] N. Rohland, E. Harney, S. Mallick, S. Nordenfelt, and D. Reich. “Partial uracil-DNA-glycosylase treatment for screening of ancient DNA”. en. In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 370.1660 (2015), page 20130624.
- [189] J.-M. Rouillard. *MYcroarray - MYbaits Target Enrichment Kit: Capture Probes for Targeted NGS Sequencing*. <http://www.mycroarray.com/mybaits/MYbaits+sequence+capture+target+enrichment+kit.html>. Accessed: 2017-10-29.
- [190] F. Rousset. “genepop’007: a complete re-implementation of the genepop software for Windows and Linux”. In: *Molecular ecology resources* 8.1 (2008), pages 103–106.
- [191] J. Rowley, I. Toskin, and F. Ndowa. “Global incidence and prevalence of selected curable sexually transmitted infections: 2008”. In: *Global incidence and prevalence of selected curable sexually transmitted infections: 2008*. 2012, pages 20–20.
- [192] O Rubensohn and F Knatz. “Bericht über die Ausgrabungen bei Abusir el Mäläq im Jahre 1903”. In: *Zeitschrift für Ägyptische Sprache* 41.JG (1905), pages 1–21.

Bibliography

- [193] F. Sanger and A. R. Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". en. In: *Journal of molecular biology* 94.3 (1975), pages 441–448.
- [194] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". en. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pages 5463–5467.
- [195] S. Sawyer, J. Krause, K. Guschanski, V. Savolainen, and S. Pääbo. "Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA". en. In: *PloS one* 7.3 (2012).
- [196] C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, et al. "Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago". en. In: *Science* 358.6363 (2017), pages 652–655.
- [197] C Schrader, A Schielke, L Ellerbroek, and R Johne. "PCR inhibitors - occurrence, properties and removal". en. In: *Journal of applied microbiology* 113.5 (2012), pages 1014–1026.
- [198] M. Schubert, L. Ermini, C. Der Sarkissian, H. Jónsson, et al. "Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX". en. In: *Nature protocols* 9.5 (2014), pages 1056–1082.
- [199] M. Schubert, S. Lindgreen, and L. Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". In: *BMC research notes* 9.1 (2016), page 88.
- [200] V. J. Schuenemann and J. Krause. *Einführung in die Paläogenetik*. 2017.
- [201] 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 (2017), page 15694.
- [202] V. J. Schuenemann, P. Singh, T. A. Mendum, B. Krause-Kyora, et al. "Genome-wide comparison of medieval and modern *Mycobacterium leprae*". en. In: *Science* 341.6142 (2013), pages 179–183.
- [203] D. Shriner and S. O. Y. Keita. "Migration Route Out of Africa Unresolved by 225 Egyptian and Ethiopian Whole Genome Sequences". In: *Frontiers in genetics* 7 (2016), page 98.
- [204] E. E. Sigsgaard, I. B. Nielsen, S. S. Bach, E. D. Lorenzen, et al. "Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA". In: *Nature ecology & evolution* 1.1 (2016), page 4.
- [205] K. A. Sirak, D. M. Fernandes, O. Cheronet, M. Novak, et al. "A minimally-invasive method for sampling human petrous bones from the cranial base for ancient DNA analysis". In: *BioTechniques* 62.6 (2017), pages 283–289.

- [206] P. Skoglund, H. Malmström, A. Omrak, M. Raghavan, et al. “Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers”. In: *Science* 344.6185 (2014), pages 747–750.
- [207] P. Skoglund and I. Mathieson. “Ancient genomics: a new view into human prehistory and evolution”. 2017.
- [208] 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 of the United States of America* 111.6 (2014), pages 2229–2234.
- [209] P. Skoglund, C. Posth, K. Sirak, M. Spriggs, et al. “Genomic insights into the peopling of the Southwest Pacific”. In: *Nature* 538.7626 (2016), pages 510–513.
- [210] P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, et al. “Reconstructing Prehistoric African Population Structure”. In: *Cell* 171.1 (2017), pages 59–71.
- [211] M Slatkin. “A measure of population subdivision based on microsatellite allele frequencies”. In: *Genetics* 139.1 (1995), pages 457–462.
- [212] M. Slatkin and F. Racimo. “Ancient DNA and human history”. en. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.23 (2016), pages 6380–6387.
- [213] S. Smith. “Pharaohs, Feasts, and Foreigners”. In: *The archaeology and politics of food and feasting in early states and empires* (2003), pages 39–64.
- [214] S. T. Smith. “Ethnicity, Egypt”. In: *The Encyclopedia of Ancient History*. John Wiley & Sons, Inc., 2013.
- [215] T. F. Smith and M. S. Waterman. “Identification of common molecular subsequences”. en. In: *Journal of molecular biology* 147.1 (1981), pages 195–197.
- [216] M. Soejima and Y. Koda. “Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2”. en. In: *International journal of legal medicine* 121.1 (2007), pages 36–39.
- [217] M. L. Steiner and A. E. Killebrew. *The Oxford Handbook of the Archaeology of the Levant: C. 8000-332 BCE*. en. OUP Oxford, 2014.
- [218] M. Stoneking and J. Krause. “Learning about human population history from ancient and modern genomes”. en. In: *Nature reviews. Genetics* 12.9 (2011), pages 603–614.

Bibliography

- [219] R. A. Sturm, D. L. Duffy, Z. Z. Zhao, F. P. N. Leite, et al. “A single SNP in an evolutionary conserved region within intron 86 of the *HERC2* gene determines human blue-brown eye color”. en. In: *American journal of human genetics* 82.2 (2008), pages 424–431.
- [220] P. Sulem, D. F. Gudbjartsson, S. N. Stacey, A. Helgason, et al. “Genetic determinants of hair, eye and skin pigmentation in Europeans”. en. In: *Nature genetics* 39.12 (2007), pages 1443–1452.
- [221] K Tamura and M Nei. “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees”. en. In: *Molecular biology and evolution* 10.3 (1993), pages 512–526.
- [222] The FastX community & HannonLab. *The FastX Toolkit*. 2014.
- [223] *The Online Ancient Genome Repository (OAGR)*. <https://www.oagr.org.au/help/>. Accessed: 2017-11-4.
- [224] The TestFX Community. *TestFX: Testing GUI applications in Java*.
- [225] The Travis CI Community. *Travis CI: A continuous integration service*.
- [226] *The Vaadin web framework*. <https://vaadin.com/>. Accessed: 2017-11-5.
- [227] J. T. Troelsen. “Adult-type hypolactasia and regulation of lactase expression”. en. In: *Biochimica et biophysica acta* 1723.1-3 (2005), pages 19–32.
- [228] United Nations Statistics Division. *UNSD — Methodology*. <https://unstats.un.org/unsd/methodology/m49/>. Accessed: 2017-11-11.
- [229] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, et al. “The sequence of the human genome”. en. In: *Science* 291.5507 (2001), pages 1304–1351.
- [230] J. Vivian, A. A. Rao, F. A. Nothhaft, C. Ketchum, et al. “Toil enables reproducible, open source, big biomedical data analyses”. en. In: *Nature biotechnology* 35.4 (2017), pages 314–316.
- [231] S. Walsh, L. Chaitanya, L. Clarisse, L. Wirken, et al. “Developmental validation of the HirisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage”. en. In: *Forensic science international. Genetics* 9 (2014), pages 150–161.
- [232] Y Wang, C. B. Harvey, E. J. Hollox, A. D. Phillips, et al. “The genetically programmed down-regulation of lactase in children”. en. In: *Gastroenterology* 114.6 (1998), pages 1230–1236.
- [233] H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, et al. “Haplo-Grep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing”. en. In: *Nucleic acids research* 44.W1 (2016), W58–63.
- [234] B. Welte. *Zeitzeugen Aus Dem Wüstensand. Die Altägyptischen Mumien-schädel Aus Abusir El-Meleq*. VML Verlag Marie Leidorf, 2016.

Bibliography

- [235] S. R. Woodward, N. J. Weyand, and M Bunnell. “DNA sequence from Cretaceous period bone fragments”. en. In: *Science* 266.5188 (1994), pages 1229–1232.
- [236] M. Ziemann, Y. Eren, and A. El-Osta. “Gene name errors are widespread in the scientific literature”. en. In: *Genome biology* 17.1 (2016), page 177.

Bibliography

APPENDIX A

Supplementary Information

Table A.1: Whole genome shotgun screening, mitochondrial capture and 390K SNP capture results on libraries created from a skeletal sample of George Bähr. Mapped reads are stated after deduplication. Coverage is on target, for example the 390K capture positions or the mitochondria only.

<i>Sample</i>	<i>Mapped Reads</i>	<i>Mean coverage</i>	<i>Cluster Factor</i>	<i>MT Contamination estimate</i>	<i>DNA Damage 3'/5'</i>
WGS	134,962	0.0027x	1.005	-	7.5% / 7.5%
MT-Capture	97,539	395.34x	2.212	1-2%	7.4% / 7.4%
390K	25,415,564	29.19x	-	-	-

Table A.2: Detailed coverage comparison of 110 *Treponema pallidum* samples, after applying SAMtools rmdup and DeDup to the same reads, respectively.

Sample	Coverage (X)		Δ (X)
	Samtools rmdup	DeDup	
SRR2996733	1.17	1.17	0
SRR3054906	7.98	8.05	0.07
ARG1	5.22	5.94	0.72
ARG2	81.91	157.48	75.57
ARG3	3.97	4.36	0.39
ARG4	2.98	3.2	0.22
SRR2996728	31.15	32.41	1.26
BAL3	38.46	52.11	13.65
BAL73	18.93	20.8	1.87
BAL9	1.9	1.93	0.03
C22	1.96	1.99	0.03
C23	0.44	1.15	0.71
C24	0.11	0.15	0.04
C25	0.14	0.15	0.01
C26	2.46	2.5	0.04
C27	55.81	74.79	18.98
C28	0.65	0.66	0.01
SRR2996729	139.63	168.09	28.46
C30	0.33	0.42	0.09
C31	0.43	0.51	0.08
C32	0.18	0.19	0.01
C33	4.39	4.88	0.49
C34	1.24	2.71	1.47
CDC	2.86	3.03	0.17
GHA1	22.49	30.35	7.86
GRA1	0.69	0.69	0
GRA2	42.16	103.1	60.94
HAIB	2.62	2.68	0.06
IND1	86.84	249.19	162.35
IND2	1.41	1.44	0.03
IRAB	0.96	0.97	0.01
SRR2996730	190.89	245.13	54.24
N13	4.65	4.79	0.14
N14	5.31	5.47	0.16
N15	12.37	12.91	0.54

Table continues

Appendix A. Supplementary Information

Sample	Samtools rmdup	DeDup	Δ (X)
N16	1.51	1.54	0.03
N17	35.97	41.26	5.29
N18	0.12	0.13	0.01
N19	4.96	5.37	0.41
N20	22.91	24.37	1.46
N21	0.39	0.45	0.06
NIC1	275.98	1948.79	1672.81
NIC2	55.68	125.09	69.41
SRR2996731	356.86	687	330.14
SRR3571774	41.55	45.14	3.59
SRR3584962	25.75	26.84	1.09
SRR3584965	48.15	54.7	6.55
SRR3571775	84.76	96.02	11.26
SRR3571776	8.48	8.76	0.28
SRR3571778	133.75	59.37	-74.38
SRR3571783	184.87	230.03	45.16
SRR3571785	198.28	253.32	55.04
SRR3571786	35.54	40.86	5.32
SRR3571794	10.06	11.22	1.16
SRR3584837	20.09	20.7	0.61
SRR3584838	6.81	10.49	3.68
SRR3584839	29.07	33.86	4.79
SRR3584840	400.57	790.28	389.71
SRR3584842	111.34	133.18	21.84
SRR3584843	336.34	639.62	303.28
SRR3584844	90.29	126.13	35.84
SRR3584845	246.29	356.23	109.94
SRR3584871	206.2	323.37	117.17
SRR3584875	178.19	250.85	72.66
SRR3584879	32.97	35.98	3.01
SRR3584882	197.92	314.86	116.94
SRR3584884	30.45	31.47	1.02
SRR3584886	23.07	25.22	2.15
SRR3584841	51.44	56.93	5.49
SRR2996732	344.31	614.43	270.12
S1	19.58	20.89	1.31
S10	0.59	1.05	0.46
S11	0.09	0.09	0
S12	0.18	0.22	0.04

Table continues

Sample	Samtools rmdup	DeDup	Δ (X)
S13	5.06	5.45	0.39
S14	0.38	0.78	0.4
S15	63.58	83.7	20.12
S16	42.23	53.65	11.42
S18	0.41	0.5	0.09
S19	1.3	1.46	0.16
S2	3.19	3.77	0.58
S3	0.19	0.22	0.03
S4	6.19	11.01	4.82
S5	0.24	0.29	0.05
S6	8.45	11.52	3.07
S7	0.87	1.63	0.76
S8	17.63	19.06	1.43
S9	0.57	0.87	0.3
SAM1	254.18	1418.39	1164.21
SEA86	9.16	9.67	0.51
SRR2996724	302.25	457.7	155.45
SRR2996725	296.25	447.75	151.5
SRR2996726	59.91	63.95	4.04
SRR2996727	12.59	12.81	0.22
SPA1	0.26	0.27	0.01
UK1	0.2	0.21	0.01
UK2	0.2	0.45	0.25
UK3	0.38	0.4	0.02
UK5	0.25	0.29	0.04
UK6	0.36	0.43	0.07
UW1	27.67	33.54	5.87
UW2	7.1	7.75	0.65

Table A.3: Mean coverage on the first and last 50 bases of 52 ancient Egyptian Mitochondrial captures. Compared here are the mean coverages on the first and last 50 bases of each sample, obtained by applying BWA or CircularMapper for read mapping.

Sample	Mean coverage (X)		Δ (X)
	BWA	CircularMapper	
JK2866	177.14	230.25	53.11
JK2867	0.78	1.71	0.93
JK2868	0.75	10.06	9.31

Table continues

Appendix A. Supplementary Information

Sample	BWA	CircularMapper	Δ (X)
JK2869	6.18	14.59	8.41
JK2870	23.65	35.37	11.72
JK2871	3.33	7.65	4.32
JK2872	29.47	46.88	17.41
JK2873	27.33	47.8	20.47
JK2874	7.25	21.12	13.87
JK2875	5.39	41.78	36.39
JK2876	33.25	43.41	10.16
JK2877	0.14	0.63	0.49
JK2878	51.24	76.63	25.39
JK2879	329.88	418.29	88.41
JK2880	22.88	34.14	11.26
JK2881	2.24	2.49	0.25
JK2882	0.8	2.8	2
JK2883	212.65	256.69	44.04
JK2884	4.06	11.86	7.8
JK2886	10.78	16.86	6.08
JK2887	3.67	9.45	5.78
JK2888	281.04	331.16	50.12
JK2889	117.63	153.57	35.94
JK2890	482.94	521.63	38.69
JK2891	0.18	16	15.82
JK2893	120.18	168.57	48.39
JK2894	8.67	25.9	17.23
JK2895	66.39	81.12	14.73
JK2896	623.35	705.08	81.73
JK2897	0.02	0.02	0
JK2898	4.47	7.82	3.35
JK2900	17.84	45.88	28.04
JK2901	0.37	2.39	2.02
JK2902	77.47	106.24	28.77
JK2903	1.39	2.98	1.59
JK2904	659.41	722.31	62.9
JK2907	19.16	31.06	11.9
JK2908	0.61	2.8	2.19
JK2910	0.53	1.53	1
JK2911	1417.73	1478.06	60.33
JK2912	2.08	4.24	2.16
JK2916	5.88	5.88	0

Table continues

Sample	BWA	CircularMapper	Δ (X)
JK2917	7.59	13.86	6.27
JK2918	11.69	18.49	6.8
JK2919	114.73	151.24	36.51
JK2920	22.9	29.76	6.86
JK2922	12.55	18.65	6.1
JK2923	23.1	34.67	11.57
JK2924	7.02	9.51	2.49
JK2951	16.41	32.94	16.53
JK2952	15.14	27.12	11.98
JK2953	99.8	114.31	14.51

Table A.4: Pipeline evaluation criteria table. The category of each evaluation is given, together with a short description of the respective evaluation, following standards as stated by Koschmieder *et al.* [94].

Category	Criteria	Description
Basics	Project homepage	The URL of the project homepage
Basics	Organisation	The University that developed the tool
Basics	People Involved	The people involved in development of the tool
Basics	Brief description	A brief description of the tool
Basics	Year of Publication	The year of the tool publication
Basics	Number of references (Google)	Reference number on Google Scholar
Basics	Analysis	Supports analysis of aDNA data
Basics	Commercial	Academic or commercial licence
System properties	Type of installation	What type of installation - Server based / local installation / HPC installation / cloud installation
System properties	Source code available	If the source code of the application open sourced?
System properties	How to obtain	Under which licence and from where can the source code be downloaded
System properties	Data storage	How is data stored
System properties	Operating system	Which operating system is required? On which does the pipeline work?
System properties	Software requirements	Which software is required to operate the tool?
System properties	Maintained	Is the tool still maintained? Last year of maintenance?
Standards	FastQ/SAM/BAM	Support for FastQ/SAM/BAM formats?
Standards	VCF	Support for VCF formats?
Pre-processing	Pre-processing methods	Which tools are supported for pre-processing?
Pre-processing	Quality control methods	Which tools are supported for quality control?
Analysis	Mappers	Which mapping tools are supported?
Analysis	PCR Duplicate removal	Which duplicate removal tools are supported?
Analysis	Workflow support	Is there support for workflow languages in the pipeline? Can workflows be used in the pipeline?
Analysis	DNA Damage metrics	Does the pipeline produce DNA damage metrics?
Analysis	DNA contamination assessment	Does the pipeline assess contamination for investigated samples?
Analysis	DNA mapping metrics	Does the pipeline produce basic mapping metrics, e.g. endogenous DNA?
Analysis	Other analysis support	Other supported analysis tools in the pipeline?
Analysis	Parallelized? (Process/Sequential)	Parallelized tools (processes) and pipeline execution?
Software testing	Tested	Is the pipeline tested using continuous integration? If yes, is there some basic workflows that are tested?
Reproducibility	Container support	Does the pipeline support containerization and/or is available in scientific workflow containers?
Output	Result	What is the main type of output from the tool?
Output	Export formats	Which export formats are supported?
Output	Additional output features	Which additional output features are available?

Table A.5: Analysis results of 90 mitochondrial genomes of Ancient Egyptian mummies, created with EAGER. Haplogroups have been determined using Haplogrep 2[233], C14 dates were manually added as meta-information.

ID	Mapped Reads	Mean Cov.	Damage 3'/5'	Init. Cont. mean/low/high	Fin. Cont. mean/low/high	Haplogroup	cal. C14
JK2127	386,315	1252.85	0.13/0.14	0/0/0.005	0.01/0/0.02	W6	BC 358-208
JK2128	22,297	77.99	0.09/0.09	0/0/0.005	0.01/0/0.02	HV21	BC 185-107
JK2130	104,148	310.99	0.12/0.13	0/0/0.005	0.01/0/0.02	M1a1	AD 91-212
JK2131	644,835	2430.26	0.1/0.1	0.015/0.01/0.02	0.01/0/0.02	U3b	BC 749-517
JK2132	7,546	22.66	0.16/0.16	0/0/0.005	0.01/0/0.02	T	AD 83-208
JK2133	845,037	3587.08	0.07/0.08	0/0/0.005	0.01/0/0.02	X	BC 750-525
JK2134	204,658	666.96	0.08/0.08	0/0/0.005	0.01/0/0.02	J1d	BC 776-569
JK2135	5,058	15.97	0.14/0.14	0/0/0.01	0.01/0/0.02	M1a2a	BC 992-923
JK2136	109,427	434.90	0.1/0.1	0/0/0.005	0.01/0/0.02	R0a2	BC 405-394
JK2137	17,071	51.60	0.1/0.11	0/0/0.005	0.01/0/0.02	J2a2b	BC 164-60
JK2139	161,049	438.50	0.1/0.11	0/0/0.005	0.01/0/0.02	K1a	AD 54- 124
JK2141	45,820	117.47	0.22/0.22	0/0/0.005	0.01/0/0.02	J2a2e	BC 358- 204
JK2142	4,640	10.95	0.24/0.26	0/0/0.005	0.01/0/0.02	U6a	BC 382- 234
JK2143	11,989	32.44	0.2/0.2	0/0/0.005	0.01/0/0.02	T1a7	BC 801- 777
JK2150	6,002	18.09	0.4/0.41	0.065/0.045/0.085	0.02/0.01/0.03	K1a4	BC 759-551
JK2153	9,694	25.34	0.49/0.47	0/0/0.005	0.02/0.01/0.03	R0a1a	BC 43-AD 15
JK2155	3,959	13.99	0.17/0.16	0.19/0.15/0.23	0.01/0/0.02	T	AD 386-426
JK2158	9,533	33.77	0.14/0.12	0.12/0.08/0.16	0.01/0/0.02	X1c	AD 261-382
JK2165	230,088	688.88	0.19/0.2	0/0/0.005	0.01/0/0.02	W3a1	BC 364-211

Table continues

ID	Mapped Reads	Mean Cov.	Damage 3'/5'	Init. Cont. mean/low/high	Fin. Cont. mean/low/high	Haplogroup	cal. C14
JK2169	9,992	33.14	0.16/0.16	0.27/0.245/0.295	0.01/0/0.02	W8	BC 355-204
JK2866	125,736	436.37	0.09/0.09	0/0/0.015	0.01/0/0.02	R0a2	BC 395-263
JK2870	14,049	43.86	0.14/0.13	0/0/0.005	0.01/0/0.02	R0a	BC 899-841
JK2872	44,497	132.90	0.13/0.13	0/0/0.005	0.01/0/0.02	HV1a2a	AD 81-132
JK2873	36,522	103.92	0.16/0.17	0/0/0.005	0.01/0/0.02	T2	BC 804-792
JK2874	8,350	34.63	0.08/0.06	0/0/0.03	0.01/0/0.02	U	BC 151-48
JK2875	11,446	40.83	0.13/0.11	0.42/0.395/0.445	0.01/0/0.02	N	AD 340-395
JK2876	27,579	73.92	0.23/0.22	0/0/0.005	0.01/0/0.02	T1a8a	BC 151-46
JK2878	52,799	184.78	0.08/0.08	0/0/0.01	0.01/0/0.02	T1a7	BC 344-126
JK2880	25,424	68.16	0.23/0.25	0/0/0.005	0.01/0/0.02	T1a2	BC 770-567
JK2881	7,283	19.82	0.29/0.28	0/0/0.005	0.01/0/0.02	T2c1	BC 367-212
JK2884	4,281	13.68	0.2/0.21	0/0/0.005	0.01/0/0.02	T1a5	BC 158-54
JK2885	3,642	11.74	0.15/0.15	0/0/0.065	0.01/0/0.02	R2'JT	BC 1304-1136
JK2886	7,730	23.66	0.18/0.17	0/0/0.02	0.01/0/0.02	T1a7	BC 398-373
JK2887	5,105	15.86	0.22/0.21	0/0/0.005	0.02/0.01/0.03	J2a1a1	BC 1388-1311
JK2888	172,400	567.20	0.08/0.09	0.055/0.04/0.07	0.01/0/0.02	U6a2	BC 97-2
JK2889	69,433	209.26	0.14/0.15	0/0/0.005	0.01/0/0.02	U7	BC 797-674
JK2890	325,501	1001.31	0.15/0.16	0/0/0.005	0.01/0/0.02	I	BC 794-671
JK2893	81,910	278.72	0.08/0.08	0/0/0.005	0.01/0/0.02	H5	BC 797-771
JK2895	81,553	264.11	0.18/0.18	0/0/0.005	0.01/0/0.02	K 16T	AD 25-111
JK2899	481,234	1884.30	0.09/0.09	0/0/0.005	0.01/0/0.02	T1a7	BC 795-674
JK2900	11,759	45.08	0.12/0.12	0.095/0.06/0.13	0.01/0/0.02	HV	BC 804-786
JK2902	69,310	217.31	0.12/0.12	0/0/0.005	0.01/0/0.02	I	BC 902-842

Table continues

ID	Mapped Reads	Mean Cov.	Damage 3'/5'	Init. Cont. mean/low/high	Fin. Cont. mean/low/high	Haplogroup	cal. C14
JK2903	2,157	10.95	0.05/0.05	0/0/0.95	0.01/0/0.02	U5a	BC 87-AD 2
JK2904	300,905	1111.80	0.11/0.11	0/0/0.005	0.01/0/0.02	R0a1a	BC 362-210
JK2907	16,112	51.45	0.09/0.08	0.29/0.26/0.32	0.01/0/0.02	HV1a'b'c	AD 26-84
JK2911	982,165	4284.30	0.16/0.16	0/0/0.005	0.01/0/0.02	M1a1	BC 769-560
JK2913	11,389	32.59	0.18/0.19	0/0/0.025	0.01/0/0.02	X1	BC 895-834
JK2914	10,583	35.81	0.11/0.11	0.11/0.07/0.145	0.01/0/0.02	T2	BC 510-408
JK2916	5,413	17.40	0.23/0.25	0.07/0.03/0.105	0.01/0/0.02	R0	BC 1111-998
JK2918	16,605	43.46	0.16/0.16	0.325/0.30/0.345	0.01/0/0.02	J2a2e	AD 84-129
JK2919	81,545	264.54	0.12/0.12	0/0/0.005	0.01/0/0.02	J2a2c	BC 790-671
JK2920	18,276	48.55	0.17/0.17	0/0/0.005	0.02/0.01/0.03	U8b1a1	BC 758-552
JK2921	74,392	236.38	0.09/0.1	0/0/0.005	0.01/0/0.02	R0a1	AD 35-120
JK2922	4,684	14.91	0.14/0.11	0.44/0.365/0.515	0.01/0/0.02	R	BC 352-200
JK2923	22,059	56.31	0.15/0.16	0.065/0.045/0.085	0.01/0/0.02	U8b1a1	BC 753-544
JK2925	52,593	168.75	0.09/0.1	0/0/0.005	0.01/0/0.02	U7	AD 5-54
JK2950	102,240	346.33	0.12/0.12	0.005/0/0.02	0.01/0/0.02	H6b	BC 357-206
JK2951	20,680	62.94	0.13/0.13	0/0/0.005	0.01/0/0.02	U8b1b1	BC 344-169
JK2952	25,218	76.59	0.16/0.15	0.095/0.075/0.115	0.01/0/0.02	J2a2c	BC 790-603
JK2953	71,710	194.12	0.1/0.09	0/0/0.005	0.01/0/0.02	M1a1	BC 37-AD 48
JK2955	6,274	20.88	0.11/0.13	0.305/0.265/0.345	0.01/0/0.02	L3	BC 391-260
JK2956	25,847	75.83	0.12/0.13	0/0/0.005	0.01/0/0.02	U1a1a3	BC 823-785
JK2957	60,706	196.30	0.11/0.11	0/0/0.005	0.01/0/0.02	J2a2c	BC 788-595
JK2958	54,544	152.83	0.12/0.12	0/0/0.005	0.01/0/0.02	I	AD 27-83
JK2960	3,306	13.28	0.25/0.26	0.315/0.275/0.355	0.01/0/0.02	N1'5	BC 44-AD 16

Table continues

ID	Mapped Reads	Mean Cov.	Damage 3'/5'	Init. Cont. mean/low/high	Fin. Cont. mean/low/high	Haplogroup	cal. C14
JK2961	58,369	178.78	0.1/0.11	0/0/0.01	0.01/0/0.02	T1a7	BC 87-AD 1
JK2963	11,722	35.51	0.15/0.13	0.225/0.195/0.255	0.01/0/0.02	M1a1i	BC 1211-1126
JK2965	192,980	591.19	0.13/0.12	0/0/0.01	0.01/0/0.02	T2c1c	BC 979-914
JK2966	31,904	93.32	0.12/0.12	0/0/0.01	0.01/0/0.02	T1a7	BC 384-235
JK2970	109,087	309.24	0.18/0.18	0/0/0.005	0.01/0/0.02	U1a1	BC 357-206
JK2972	15,567	41.66	0.24/0.26	0/0/0.005	0.01/0/0.02	T1a5	BC 156-53
JK2973	235,490	773.76	0.15/0.15	0/0/0.005	0.01/0/0.02	U6a3	BC 347-168
JK2974	18,664	56.64	0.11/0.1	0.125/0.095/0.155	0.01/0/0.02	H	BC 889-803
JK2975	5,674	17.17	0.23/0.23	0/0/0.005	0.01/0/0.02	R	BC 43-AD 45
JK2977	27,649	73.43	0.29/0.29	0/0/0.005	0.02/0.01/0.03	T2e	BC 389-235
JK2978	21,330	69.51	0.1/0.1	0.095/0.065/0.125	0.01/0/0.02	N1a1a2	BC 975-905
JK2979	44,646	138.95	0.13/0.13	0.06/0.04/0.08	0.01/0/0.02	HV1a2a	BC 369-211
JK2980	86,605	245.63	0.12/0.12	0/0/0.005	0.01/0/0.02	I	BC 357-204
JK2981	75,044	207.00	0.26/0.25	0/0/0.005	0.01/0/0.02	M1a1e	BC 399-376
JK2985	429,782	1334.22	0.11/0.11	0/0/0.005	0.01/0/0.02	HV1a'b'c	BC 352-195
JK2986	7,715	22.66	0.11/0.12	0.08/0.035/0.125	0.01/0/0.02	HV	BC 508-406
JK2987	481,796	1911.05	0.08/0.07	0/0/0.005	0.01/0/0.02	HV1a'b'c	BC 342-117
JK2879	174,268	766.08	0.08/0.08	0.125/0.115/0.135	0.11/0.1/0.12	U3b	BC 45-AD 4
JK2883	146,310	516.92	0.09/0.09	0.17/0.16/0.18	0.11/0.1/0.12	T1a	BC 799-781
JK2896	312,867	1268.43	0.09/0.09	0.095/0.09/0.1	0.11/0.1/0.12	HV1b2	BC 394-239
JK2959	483,910	2264.69	0.09/0.09	0.045/0.035/0.055	0.06/0.05/0.07	T1a	BC 44-AD 16
JK2962	481,208	1855.14	0.08/0.08	0.04/0.025/0.055	0.01/0.01/0.02	H13c1	BC 756-545
JK2982	131,095	394.03	0.18/0.18	0/0/0.005	0.01/0.01/0.02	T1a5	BC 92-1

Table continues

ID	Mapped Reads	Mean Cov.	Damage 3'/5'	Init. Cont. mean/low/high	Fin. Cont. mean/low/high	Haplogroup	cal. C14
JK2984	30,651	94.50	0.11/0.11	0/0/0.005	0.01/0.01/0.02	U7	AD 32-122
1622BM	11,620	35.70	0.15/0.15	0.14/0.11/0.17	0.01/0.01/0.02	R0a2f	BC 806- 784

Table A.7: List of publications for modern mtDNA reference populations.

ISO-3 Country Code	Reference publication
ARE	Rowold et al. 2007; Alshamali et al. 2008
BFA	Pereira et al. 2010;Cherny et al. 2006
CMR	Destro-Bisol et al. 2004;Coia et al 2005;Quintana-Murci et al. 2008
EGY	Krings et al. 1999;Rowold et al. 2007;Saunier et al. 2009 ¹
EGYKU	Kujanova et al. 2013
EGYPA	Pagani et al 2015
ENG	Garcia et al. 2011;Piercy et al. 1993;Tonks et al. unpub. ¹
ESP	Corte-Real et al. 1996;Pinto et al. 1996;Prieto et all 2011 ¹
ETH	Kivisild et al. 2004;Non et al. 2011;Poloni et al. 2009;Pagani et al. 2015
FIN	Hedmann et al. 2007;Finnil et al. 2001;Kittles et al. 1999 ¹
FRA	Garcia et al. 2011;Richard et al. 2007;Richards et al. 2000 ¹
FRO	Als et al. 2006
GEO	Reidla unpub.;Quintana-Murci et al. 2004;Comas et al. 2000 ¹
HUN	Szécsényi-Nagy et al. 2013 Szécsényi-Nagy et al. 2015
IRN	Richards et al. 2000;Metspalu et al. 2004;Comas et al. 2004 ¹
IRQ	Richards et al. 2000;Al-Zahery et al 2003;Al-Zahery et al 2011
ISL	Sajantila et al. 1995;Richards et al. 1996;Helgason et al. 2003 ¹
ISR	Di Rienzo and Wilson 1991;Richards et al. 2000;Amar et al. 2007 ¹
ITA	Guimaraes et al. 2009;Vai et al. 2015
JOR	Rowold et al. 2007;Gonzales et al. 2008
KWT	Scheible et al. 2011
LBN	Haber et al. 2011;Shlush et al. 2008
MAR	Rando et al. 1998;Turchi et al. 2009;Plaza et al. 2003 ¹
MRT	Gonzales et al. 2006;Rando et al. 1998
NOR	Helgason et al. 2001;Richards et al. 2000;Opdal et al. 1998 ¹
OMN	Rowold et al. 2007
PAK	Quintana-Murci et al. 2004;Rakha et al. 2011;Coudaux et al. 2003 ¹
QAT	Rowold et al. 2007
SAU	Abu-Amero et al. 2008;Di Rienzo and Wilson 1991
SDN	Krings et al. 1999
SRB	Harvey et al. unpub.;Zgonjanin et al. 2010
SVN	Malyarchuk et al. 2003;Metspalu unpub.;Zupanic Pajnic et al. 2004
SWE	Tillmar et al. 2010;Sajantila et al. 1995;Tonks et al. unpub.; ¹
SYR	Richards et al. 2000;Vernesi et al. 2001
TRO	Otoni et al. 2016
TUN	Plaza et al. 2003;Turchi et al. 2009;Cherni et al. 2009; ¹
TUR	Calafell et al. 1996;Comas et al. 1996;Kivisild et al. 2002 ¹
YEM	Kivisild et al. 2004;Non et al. 2011;Rowold et al. 2007; ¹

¹For a complete list, please check Schuenemann *et al.* [201]

Appendix A. Supplementary Information

Table A.6: ISO-3 country codes for all investigated populations in the ancient Egyptian mummy project.

Country	ISO-3 Code
ARE	United Arab Emirates
BFA	Burkina Faso
CMR	Cameroon
GIN	Guinea
EGY	Egypt
EGYKU	Egypt Kujanova [99]
EGYPA	Egypt Pagani [155]
ENG	England
ESP	Spain
ETH	Ethiopia
FIN	Finland
FRA	France
FRO	Faroe Islands
GEO	Georgia
HUN	Hungary
IRN	Iran
IRQ	Iraq
ISL	Iceland
ISR	Israel
ITA	Italy
JOR	Jordania
KWT	Kuwait
LBN	Lebanon
MAR	Morocco
MRT	Mauritania
NOR	Norway
OMN	Oman
PAK	Pakistan
PPP	Pre-Ptolemaic Period [201]
PP	Ptolemaic Period [201]
QAT	Qatar
SAU	Saudia Arabia
SDN	Sudan& South Sudan
SRB	Serbia
SVN	Slovenia
SWE	Sweden
SYR	Syria
TRO	Turkish Romans [150]
TUN	Tunisia
TUR	Turkey
YEM	Yemen

APPENDIX B

Publications

B.1 Articles

- Jäger G, **Peltzer A**, and Nieselt K. *inPHAP: Interactive visualization of genotype and phased haplotype data*. **BMC Bioinformatics**, Jul 2014, 15:200
- **Peltzer A**, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K. *EAGER: efficient ancient genome reconstruction* **GenomeBiology** 2016, 17:60
- Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwängler A, Haak W, Meyer M, Mittnik A, Nickel B, **Peltzer A**, Rohland N, Slon V, Talamo S, Lazaridis I, Lipson M, Mathieson I, Schiffels S, Skoglund P, Derevianko A, Drozdov NM, Slavinsky V, Tsybankov A, Cremonesi R, Mallegni F, Gély B, Vacca E, González Morales M R, Straus L G, Neugebauer-Maresch C, Teschler-Nicola M, Constantin S, Teodora Moldovan O, Benazzi S, Persani M, Coppola D, Lari M, Ricci S, Ronchitelli A, Valentin F, Thevenet C, Wehrberger K, Grigorescu D, Rougier H, Crevecoeur I, Flas D, Semal P, Mannino M A, Cupillard C, Bocherens H, Conard N J, Harvati K, Moiseyev V, Drucker D G, Svoboda J, Richard M P, Caramelli D, Pinhasi R, Kelso J, Patterson N, Krause J, Pääbo S and Reich D. *The genetic history of Ice Age Europe* **Nature** 534, 200-205, June 2016
- Arora N, Schuenemann VJ, Jäger G, **Peltzer A**, Seitz A, Herbig A, Strouhal M, Grillová L, Sánchez-Busó L, Kühnert D, Bos KI, Rivero Davis L, Mikalová L, Bruisten S, Komericki P, French P, Grant PR, Pando MA, Gallo Vaulet L, Fermepin MR, Martinez A, Centurion Lara A, Giacani L, Norris SJ, Šmajš D,

Appendix B. Publications

Bosshard PP, González-Candelas F, Nieselt K, Krause J, Bagheri HC. *Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster. Nature Microbiology* 2016, 2:16245

- Schünemann VJ & **Peltzer A**¹, Welte B, van Pelt WP, Molak M, Wang CC, Furtwängler A, Urban C, Reiter E, Nieselt K, Teßmann B, Francken M, Harvati K, Haak W, Schiffels S and Krause J. *Ancient Egyptian mummy genomes suggest an increase of Sub-Saharan African ancestry in post-Roman periods. Nature Communications* 8:15694
- Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, Furtwängler A, **Peltzer A**, Posth C, Vasilakis A, MCGeorge PJP, Konsolaki-Yannopoulou E, Korres G, Martlew H, Michalodimitrakis M, Özsait M, Öz-sait N, Papathanasiou A, Richards M, Roodenbrg S A, Tzedakis Y, Arnott R, Fernandes D M, Hughey J R, Lotakis D M, Navas P A, Maniatis Y, Stamatoyannopoulos J A, Stewardson K, Stockhammer P, Pinhasi R, Reich D, Krause J and Stamatoyannopoulos G. *Genetic origins of the Minoans and Mycenaeans Nature* 548, 214-218, August 2017
- Skoglund P, Mittnik A, Sirak K, Hajdinjak M, Rohland N, Mallick S, Salie T, Heinze A, Meyer M, **Peltzer A**, Ferry M, Harney E, Michel M Stewardson K Cerezo-Roman J, Chiumia C, Crowther A, Gomani-Chindebvu E, Helm R, Horton M, Morris A, Parkington J, Prendergast ME, Ramesar R, Shipton C, Thompson J, Ribesasa R, Hayes V, Pääbo S, Patterson N, Boivin N, Krause J and Reich D. *Reconstruction Prehistoric African Population Structure Cell* 171, 1-13, September 2017
- **Peltzer A**, Mittnik A, Wang CC, Begg T, Posth C, Nieselt K and Krause J. *Inferring genetic origins and phenotypic traits of George Bähr, the architect of the Dresden Frauenkirche Scientific Reports* 8, 2115, January 2018.
- Neukamm J, **Peltzer A**¹, Achilli A, Balanovsky O, Boder M, Macholdt E, Olivieri A, Pala M, Parson W, Richards MB, Schönherr S, Stoneking M, Torroni A, van Oven M, Weißensteiner H, Zaporozhchenko V, Krause J, Nieselt K & Haak W. *MitoBench & MitoDB: Modern methods for the analysis of human mitochondrial data in preparation*

B.2 Posters, presentations & workshops

- **Alexander Peltzer**, Günter Jäger, Kay Nieselt and Johannes Krause. *EAGER: Efficient ancient genome reconstruction. Presentation* at the 6th International Symposium on Biomolecular Archaeology (ISBA) 2014 in Basel/Switzerland, August 27, 2014

¹The first two authors contributed equally to this study.

B.2. Posters, presentations & workshops

- **Alexander Peltzer**, Günter Jäger, Kay Nieselt and Johannes Krause. *EAGER: Efficient ancient genome reconstruction*. **Presentation** at the annual StEvE meeting, Students of Evolution & Ecology Graduate School Tübingen/Germany, October 10, 2014
- **Alexander Peltzer**, Günter Jäger, Kay Nieselt and Johannes Krause. *EAGER: Efficient ancient genome reconstruction*. **Invited Talk** at the Christian-Albrechts-Universität zu Kiel/Germany, November 3, 2014
- **Alexander Peltzer**, Verena J Schünemann. *Applications of Bioinformatics methods on ancient archaeological datasets*. **Presentation** at the Urgeschichtliches Museum Blaubeuren, July 5, 2015
- **Alexander Peltzer**, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause and Kay Nieselt. *EAGER: Efficient Ancient Genome Reconstruction* **Poster** at the ISMB 2015 Conference in Dublin/Ireland, July 6, 2015
- **Alexander Peltzer** *Bioinformatics in a box - Dockerizing complex bioinformatics applications for improved reproducibility*. **Invited Talk** at the ISC Cloud & Big Data Conference in Frankfurt am Main/Germany, July 10, 2015
- **Alexander Peltzer**, Stephan Schiffels, Stephen Clayton. *Genome Analysis: Contamination Assessment, Population Genetics and Data Analysis on Genome-wide data* **Workshop** at the Max-Planck-Institute for the Science of Human History in Jena/Germany, May 5, 2016
- **Alexander Peltzer**, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause and Kay Nieselt. *EAGER: Efficient Ancient Genome Reconstruction* **Poster** at the SMBE 2016 Conference in Gold Coast/Australia, July 8, 2016
- **Alexander Peltzer**, Judith Neukamm, Johannes Krause, Kay Nieselt and Wolfgang Haak. *MitoBench: An interactive visual workbench for population genetics on mitochondrial DNA* **Talk** at the UKAS 2017 Conference in London/UK 2017, April 5, 2017
- **Alexander Peltzer**, Verena J Schuenemann, Kay Nieselt, Wolfgang Haak, Stephan Schiffels and Johannes Krause. *Ancient egyptian mummy genomes suggest an increase of sub-saharan African ancestry in post-Roman periods* **Poster** at the UKAS 2017 Conference in London/UK, April 5, 2017
- **Alexander Peltzer**, Judith Neukamm, Alessandro Achilli, Oleg Balanovsky, Martin Bodner, Enrico Macholdt, Anna Olivieri, Maria Pala, Walther Parson, Martin B Richards, Sebastian Schönherr, Mark Stoneking, Antonio Torroni, Mannis van Oven, Hansi Weißensteiner, Valery Zaporozhchenko, Kay Nieselt

Appendix B. Publications

and Wolfgang Haak. *MitoBench & MitoDB: Novel interactive methods for population genetics on mitochondrial DNA* **Poster** at the Mitochondrial Genomics and Evolution (MGE) meeting 2017 in Ein Gedi/Israel, September 3-7, 2017

- **Alexander Peltzer**, Gabriel Renaud, Judith Neukamm. *Modern computational methods in ancient DNA analysis* **Workshop** at the German Conference on Bioinformatics (GCB) in Tübingen/Germany, September 18, 2017

APPENDIX C

Academic teaching experience

C.1 Supervised lectures and course

SS 2014

- Practical course: *Software Engineering* for Bachelor students
- Tutorial: *Grundlagen der Bioinformatik* for Bachelor students

WS 2014/15

- Tutorial: *Advanced Transcriptomics* for Master students
- Practical course: *Advanced Transcriptomics* for Master students

SS 2015

- Practical course: *Software Engineering* for Bachelor students
- Lecture: Paleogenetics I (Guest lecture) for Master students

WS 2015/16

- Tutorial: *Bioinformatics I* for Master students

SS 2016

- Seminar: *Absolventenseminar Bioinformatik* for Bachelor / Master students

Appendix C. Academic teaching experience

WS 2016/17

- Practical course: *Advanced Transcriptomics* for Master students

SS 2017

- Lecture: Paleogenetics I (Guest lecture) for Master students

C.2 Supervised Bachelor/Master theses

2014

- Christopher Jürges. *Toolbox zur Erstellung von Contamination Reports innerhalb der EAGER Pipeline*. April 2014, Bachelorthesis
- Maximilian Hanussek. *Implementierung einer Toolbox zur automatisierten Reportgenerierung innerhalb der EAGER Pipeline*. August 2014, Bachelorthesis

2015

- Judith Neukamm. *Improved algorithms and pipelines for ancient genome reconstruction*. December 2015, Masterthesis

2016

- Max-Emil Schön. *Metabarcoding workflows for Biodiversity assessment of belowground fungal communities*. June 2016, Masterthesis (Co-Supervision)

2017

- Veronika Böttcher. *Phylogenetic reconstruction of genomes from ancient DNA* August 2017, Masterthesis (Co-Supervision)