

UNIVERSITY OF COPENHAGEN

UNIVERSITY OF YORK



UNIVERSITY
of York

PhD Thesis | Jonas Niemann

**Elucidating the past using ancient genomes
and metagenomes**

Supervisor: Professor M. Thomas P. Gilbert

Co-supervisor: Dr. Nathan Wales

Submitted on: December 2019

Elucidating the past
using ancient genomes
and metagenomes

Jonas Niemann
Doctor of Philosophy

University of York
Archaeology

December 2019

Host Institute: University of Copenhagen
The GLOBE Institute
Section for Evolutionary Genomics

Partner Institute: University of York
Department of Archaeology
BioArCh

Funding: European Union's Horizon 2020 research and innovation programme,
grant agreement no. 676154 (ArchSci2020)

Author: Jonas Niemann

Title: Elucidating the past using ancient genomes and metagenomes

Supervisor: Professor M. Thomas P. Gilbert

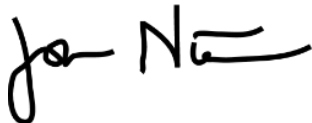
Co-supervisor: Dr. Nathan Wales

Preface

The current thesis represents the work of three years and was prepared at the Evolutionary Genomics group, The GLOBE Institute, University of Copenhagen, and BioArCh, Department of Archaeology, University of York. This work was supervised by Professor Tom Gilbert, Section of EvoGenomics, and co-supervised by Dr. Nathan Wales, BioArCh, University of York. The work was funded by the European Union's Horizon 2020 research and innovation programme, grant agreement no. 676154 (ArchSci2020).

This dissertation comprises of five sections: a general introduction, followed by three research chapters and the thesis conclusions. The first research chapter is a manuscript that discusses the population history of the extinct Honshū wolves based on the nuclear DNA obtained from one Honshū wolf museum specimen. The second research chapter investigates the potential of using fluid-preserved seabirds as a substrate for metagenomics studies. The last research chapter is a published paper that analyses the human, microbial, and non-human eukaryotic DNA recovered from a chewed birch bark pitch.

Four appendix chapters that I was involved in are included at the end of this thesis, but should not be considered part of the body of work submitted for examination.

A handwritten signature in black ink, appearing to read 'Jonas Niemann'.

Jonas Niemann, December 2019

Acknowledgements

First and foremost I want to express my gratitude for the excellent guidance of my supervisors Tom Gilbert and Nathan Wales.

I am deeply grateful to Tom Gilbert for hiring me as a PhD student and giving me the opportunity to work with really exciting datasets. Thanks to you, the past three years have been so transformative for me, and I was truly inspired by your problem-solving skills and thought-provoking ideas. I truly appreciated your encouragement and the consistently immediate response when I needed your help.

I also want to express my deepest gratitude to Nathan Wales, who was also absolutely essential for the last year of my PhD. Thank you so much for the countless insightful discussions, your support, and the immense help you provided for putting this dissertation together.

Being part of the Marie Skłodowska-Curie Innovative Training Network and European Joint Doctorate ArchSci2020 was very interesting, and I really enjoyed sharing this experience with my 14 fellow ArchSci2020 PhD students.

There were multiple “ArchScis” in Copenhagen that made life so much easier and entertaining. I am very grateful to Anne-Marijn for all the candy and board game sessions we had, Eden for being the best next-door neighbour and housemate one could wish for, Maiken for always being so supportive and being a lifesaver so many times in the last few years, Tatiana for all the funny and stimulating conversations we had, Theis for the great company that made the snus- and caffeine-fuelled chewing gum analyses more enjoyable, and Xenia for the great time with desserts, games, and movies on Vestamager.

I was also really lucky that I could share my York secondment with my fellow ArchScis AK, Alison, and Manon. AK, thank you so much for all the support and kindness throughout the past year. Alison, I really miss our chats and awesome cinema-and-beers evenings. Manon, the last few months were insane for both of us, and I was so happy to have you as my friend by my side during that time.

It was great to regularly meet the ArchScis that were based at other institutes when I was in York and Copenhagen - Ari, Jack, Maddie, Mariana, and Özge. I always looked forward to chatting with you at one of our courses and can't wait to see you again in Groningen.

There are many people at the Evogenomics group in Copenhagen that I would like to thank for the amazing first two years of my PhD: Abby, Anna, Anne Marie, Antton, Ashot, Åshild, Christian, Christina Lehmkuhl, Christina Lynggaard, Emily, Eva, Fabiana, Fatima, Filipe, George, Hannes, Inger, Jazmín, Jonathan, Katharina, Kristine, Lara, Liam, Lis, Marcela, Marisa, Marta, Martin, Matthew, Meaghan, Mick, Mikkel Sinding, Mikkel Skovrind, Miyako, Morten Limborg, Morten Tange Olsen, Nathan, Ostaizka, Physilia, Ricardo, Sama, Sarah, Sarai, Shanlin, Shyam, Thanassis, Vanessa, Victor and many more. I really enjoyed and appreciated the fantastic work environment, great friendships, and entertaining conversations over lunches and dinners. You really made me feel at home in the group and in Copenhagen.

I would especially like to thank the past and present bioinformaticians in Tom's group - Filipe, Jazmín, Lis, Sama, and Shyam - for the very insightful discussions and support when I felt stuck. Shyam in particular has been a true lifesaver at times and taught me so much about population genomics.

I would also like to thank the entire BioArCh group in York for the great last year of my PhD, with a special shout-out to the past and present aDNA group members AK, Aurelie, Aurore, Eleanor, Eve, Katharina, Krista, Matthew, Nathan, and Richard. I will miss our spontaneous pub meetings, movie nights, Indian take-aways, and all the new insights I gained over discussions with all of you.

I was also very fortunate to spend some months at the IBE Comparative Genomics lab in Barcelona. I am deeply grateful to Tomas Marques-Bonet for giving me the opportunity to work in his amazing group. I would like to use this occasion to thank Aitor, Claudia, Esther, Irene, Jessica, Laura, Luis, Lukas, Manolo, Marc, Marina, Martin, Paula, Raquel, and So Jung for making me feel so welcome in the group from the very first day and being great friends ever since.

I am very grateful to Maiken and AK for translating the thesis summary to Danish and would like to express my deepest gratitude to Nathan, Tom, and AK for giving me valuable feedback during the thesis preparation.

I also want to thank my friends Marianna, Elsa, Sophia, Oleguer, Linda, Riccardo, Mathias, Jane, Felipe, and Ikue for many great memories of the last years and keeping me sane.

Most of all I would like to thank my family for their unconditional love and support over the years. I am deeply grateful to my parents Werner and Helga, who have motivated and encouraged me throughout my career and without whose support this dissertation would not have been possible. I would also like to thank my sister Hannah and her fiancé Dominik for sheltering me in the last month of my PhD and keeping me happy and caffeinated, my brother Tobias and his partner Eva for all their support and great past adventures, and my brother Daniel, sister-in-law Steffi, and niece Emma for providing blissful moments in stressful periods.

English summary

The recognition that DNA from long dead organisms can be extracted and sequenced from a multitude of substrates has revolutionised the field of bioarchaeology. Apart from yielding profound discoveries into the biology, migrations, and admixture of past populations based on the host DNA, some artefacts are now recognised as repositories for dietary and host-associated microbial DNA and thus hold vital clues to the health status and lifestyle of the individual. This thesis is composed of three studies on three quite distinct substrates – historic hide, fluid-preserved museum specimens, and an ancient birch “chewing gum” – where I applied population genomic and metagenomic analyses to infer the population history and microbiome compositions of past organisms. After briefly introducing the broad themes of this dissertation in Chapter 1, Chapter 2 of this dissertation explores the population history of the Honshū wolves, a poorly-understood grey wolf subspecies that was endemic to the Japanese archipelago and went extinct at the beginning of the 20th century. The nuclear genome from the museum hide of one specimen was sequenced at an average depth of coverage of 3.8×, and I discovered that Honshū wolves were likely the relict of a Pleistocene Siberian wolf population that was up to now believed to have gone extinct about 10,000 years ago. Chapter 3 and Chapter 4 discuss the metagenomic potential of two novel substrates. In Chapter 3 we sequenced gut samples of six historic fluid-preserved birds with the aim of capturing the host-associated microbial profile. While I was able to characterise the gut microbiome of one specimen, further research is necessary to improve the feasibility of performing metagenomic analyses on fluid-preserved samples. Finally, in Chapter 4 I analysed the DNA extracted from a 5,700 year old chewed birch bark pitch. We obtained a complete human genome at an average depth of coverage of 2.3× and found that the female who chewed the birch pitch genetically closely resembles Western hunter-gatherers. The birch “chewing gum” also proved to be a rich source of microbial and non-human eukaryotic DNA, and I was able to recover the genomes of bacterial taxa that are closely associated with the oral microbiome as well as DNA from mallard, hazelnut, and birch that are likely derived from a recent meal and the birch pitch material itself.

In conclusion, this dissertation sheds light on wolf and human evolution as well as introduces two novel substrates with potential for future metagenomic analyses. These projects demonstrate that researchers must continue exploring whether unusual archaeological and historic substrates contain genetic material that can be used to resolve long standing questions, thereby unlocking new opportunities to understand the history of our world.

Dansk resumé

Erkendelsen af at DNA fra for længst døde organismer kan blive ekstraheret og sekventeret fra adskillige typer af materialer har revolutioneret bioarkæologi som videnskabeligt feltet. Udover at have givet dybdegående opdagelser indenfor biologi, migration, og genetisk opblanding af fortidige populationer baseret på endogent DNA, har nogle fortidige objekter nu også vist sig at indeholde DNA fra fødeindtag samt værtsbaseret mikrobielt DNA, som derved kan give fundamental indsigt i det pågældende individs helbred og livsstil. Denne PhD afhandling består af tre studier som bygger på tre vidt forskellige typer af materialer - historiske skind, sprit/formalin-konserverede museumsobjekter, og et gammel birkebark "tyggegummi" - for hvilke jeg har anvendt både populationsgenomiske og metagenomiske analyser til at tolke populationshistorie samt den mikrobielle sammensætning i fortidige organismer.

Efter en kort introduktion af de brede temaer for denne afhandling i kapitel 1, vil kapitel 2 udforske de populationshistoriske aspekter af Honshū ulvene, en endnu forholdsvist ukendt underart af gråulve, som var endemisk for det japanske øhav og uddøde i begyndelsen af det 20. århundrede. Kernegenomet fra et individ, repræsenteret af et skind fra et museum, opnåede en gennemsnitlig dækningsdybde på $3,8\times$, og jeg fandt frem til at Honshū ulvene sandsynligvis er et levn fra en Pleistocæn sibirisk ulvepopulation, som indtil nu menedes at være uddød for ca. 10.000 år siden. I kapitel 3 og kapitel 4 diskuteres det metagenomiske potentiale for to nye typer af materialer. Kapitel 3 omhandler sekventeringen af tarmprøver fra seks historiske sprit/formalin-konserverede fugle med det formål at fastslå den værtsbaserede mikrobielle profil. Jeg var i stand til at karakterisere tarmfloraen for én prøve, men yderligere forskning vil være nødvendig for at forbedre muligheden for at udføre metagenomiske analyser af sprit/formalin-konserverede prøver. Endelig omhandler kapitel 4 analysen af DNA ekstraheret fra en 5.700 år gammel tygget birkebegklump. Vi fik udtrukket et komplet menneskegenom med en gennemsnitlig dækningsdybde på $2,3\times$ og fandt ud af at kvinden, som havde tygget på begklumpen er genetisk beslægtet med vesteuropæiske jæger-samlere. Birkebark "tyggegummiet" viste sig også at være en rig kilde til mikrobielt og ikke-menneskeligt eukaryotisk DNA, og jeg var i stand til at gendanne taxonomiske grupper af bakterier, der er tæt associeret med det orale mikrobiom, samt DNA fra gråand, hasselnød og birk, som sandsynligvis stammer fra henholdsvis et nyligt måltid og fra birkebegklumpen selv. Denne PhD afhandling belyser aspekter af henholdsvis ulvens og menneskets evolution, og introducerer desuden to nye materialer til brug for fremtidige metagenomiske analyser.

Author Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Table of contents

| | |
|--|------|
| Preface..... | I |
| Acknowledgements..... | III |
| English summary..... | VI |
| Dansk resumé..... | VII |
| Author declaration..... | VIII |
| | |
| Chapter 1 | 1 |
| Introduction | |
| 1.1 History of Palaeogenomics | 3 |
| 1.1.1 Challenges of aDNA | 3 |
| 1.1.2 Noteworthy archaeological applications of aDNA | 5 |
| 1.1.3 Hides and skins | 6 |
| 1.1.4 Fluid-preserved specimens | 6 |
| 1.1.5 Birch bark pitch | 7 |
| 1.2 Metagenomic analyses | 8 |
| 1.3.1 Ancient and historical metagenomes | 9 |
| 1.3.2 Caveats of ancient metagenomic methods | 10 |
| 1.3 Thesis structure | 14 |
| | |
| PhD objectives and contributions | 22 |
| | |
| Chapter 2 | 23 |
| Complete genome of historic Honshū wolf reveals Pleistocene heritage | |
| | |
| Chapter 3 | 68 |
| Unsealing the jars - characterizing gut microbial DNA preservation in fluid-preserved museum specimens | |
| | |
| Chapter 4 | 106 |
| A 5700 year-old human genome and oral microbiome from chewed birch pitch | |
| | |
| Chapter 5 | 155 |
| Conclusions | |
| | |
| Appendix | 163 |

Chapter 1

Introduction

Preface

This dissertation explores a wide range of palaeogenomic topics and analytical approaches, and therefore it is necessary to provide a brief account of how these projects developed and build on one another. At the beginning of my PhD I set out to study the genomes of two extinct species – Maclear’s rat (*Rattus macleari*) and great auk (*Pinguinus impennis*) – with the aim to explore the effect of population collapse on the genome and the viability of the controversial concept of de-extinction. While these two projects are still in progress, ultimately they had to be sidelined due to extensive delays with the data generation at our collaborator’s facility in China. Thus so as to keep moving forward, I elected to apply my newly acquired skills in palaeogenomics to analyse several other interesting related datasets that were available for study. The first of these was a population genomic analysis of the now extinct Japanese wolf. During my analyses on the palaeogenomic dataset, I also appreciated the wealth of microbial information that can be present in such materials. Thus I also became interested in understanding the diversity of the microbial communities that resided in past organisms. This “metagenomic” approach is one of the most rapidly developing themes in ancient DNA research, so I readily accepted the opportunity to undertake the study of historic bird gut microbiomes and Late Mesolithic/Early Neolithic chewing gum.

In summary, this dissertation revolves around the analysis of aDNA retrieved from three non-bone substrates - hides, fluid-preserved specimens, and birch bark pitch - and the opportunities and limitations that come with each substrate.

My chapters are not chronologically ordered by the age of the samples, but are listed in order of the skills I acquired and reflect therefore my growth as a bioinformatician during my time as a PhD student.

Below I introduce the main themes of my dissertation with general background information on the development of palaeogenomics. Given the heterogeneous nature of the dissertation topics, the discussion swiftly shifts between themes, with the understanding that the research chapters delves into more detail in the respective introductory sections.

1.1 History of palaeogenomics

Until the 1980s researchers mostly relied on morphometric data of fossils and mummified remains to study the biology of extinct specimens. This changed dramatically with the recognition that DNA could persist long after the death of an organism, otherwise known as ancient DNA (aDNA). The first group that was able to recover DNA from long-dead organisms was Russel Higuchi and colleagues, who succeeded in sequencing two fragments with an overall length of 229 base pairs extracted from the dried muscle tissues from a quagga (*Equus quagga quagga*) museum sample (Higuchi et al. 1984), an equid subspecies that went extinct in 1883. Soon after, Pääbo and colleagues published an article on the first ancient human DNA sequence extracted from an Egyptian mummy in 1985 (Pääbo 1985). The first study on ancient plant DNA was published in 1988 by Rollo and colleagues, who extracted DNA from Peruvian maize ears that were dated to approximately 1,000 BP (Rollo et al. 1988).

Around 2005, the ground-breaking technological advances of next-generation sequencing (NGS) or high-throughput sequencing (HTS) allowed the sequencing of DNA on a much larger scale, facilitating the generation of whole genome data.

This also had a profound impact on aDNA research, and in 2008 the first mammalian whole genome, a woolly mammoth, was sequenced (Miller et al. 2008). Further studies on the woolly mammoth genome revealed insights into the population decline (Palkopoulou et al. 2015) and genomic erosion (Rogers and Slatkin 2017). aDNA also enabled population genetic analyses on other extinct species such as the thylacine (White, Mitchell, and Austin 2018), passenger pigeon (Guiry et al. 2020), and moa (Allentoft and Rawlence 2012), uncovering information on their relatedness to extant taxa and their demographic history that is inaccessible with morphological approaches. In the last decade there has been an explosion of published ancient whole genomes, with a single study in October 2019 publishing 524 ancient human genomes (Narasimhan et al. 2019).

1.1.1 Challenges of aDNA

The generation of short DNA fragments from the first aDNA sequencing efforts meant that analyses were often limited to the construction of simple phylogenies. The sequencing of entire ancient genomes, however, enabled an in-depth look into the genetic makeup of

individuals and populations that lived thousands of years ago, allowing, for example, the detection of past admixture, migration routes, population expansions and declines, rise of adaptive traits and deleterious mutations, and the prediction of phenotypic features of long-extinct organisms.

The generation of short DNA fragments from the first aDNA sequencing efforts meant that analyses were often limited to the construction of simple phylogenies. The sequencing of entire ancient genomes, however, enabled an in-depth look into the genetic makeup of individuals and populations that lived thousands of years ago, allowing, for example, the detection of past admixture (Sánchez-Quinto and Lalueza-Fox 2015), migration routes (Furholt 2018), population expansions and declines (Palkopoulou et al. 2015), rise of adaptive traits and deleterious mutations (Fry et al. 2020), and the prediction of phenotypic features of long-extinct organisms (Roca et al. 2009). However, the analysis of ancient DNA is not without its challenges. Once a cell dies and the DNA repair mechanism is disrupted, the DNA strands form cross-links, undergo chemical alterations, and start to disintegrate into smaller fragments (Mitchell, Willerslev, and Hansen 2005). Over time, only traces of very short, degraded DNA molecules remain. In a typical aDNA sequencing run, endogenous DNA comprises the minority of reads, while the remainder is derived from contaminants such as bacteria and fungi colonizing the substrate post-mortem, or originates from the humans handling the sample (Poinar et al. 2006). As a consequence, sequencing the genome of an ancient organism is much more costly than generating one for a modern specimen, as the low proportion of endogenous sequences and the short read length requires a substantial number of sequencing data in order to obtain a genome with a passable depth of coverage (Hansen et al. 2017). Furthermore, aDNA studies also rely on reference genomes of closely related species, as the short length of aDNA fragments prevents the de-novo assembly of ancient genomes (Millar et al. 2008). This can be highly problematic if the closest extant relative is highly divergent from the ancient organism. Another problem is the risk of short, damaged sequences aligning to multiple regions of the genome or even to the incorrect reference genome. Since some aDNA studies in the 1990s (Cano, Poinar, and Poinar 1992; Poinar, Cano, and Poinar 1993; An et al. 1995) turned out to be based on contaminant rather than authentic aDNA (Austin et al. 1997; Hebsgaard, Phillips, and Willerslev 2005), strict guidelines had to be introduced to minimize the risk of modern contamination. These include extracting the DNA in designated laboratories under sterile conditions (Yang and Watt 2005; Fulton 2012) and implementing protocols that maximise the endogenous DNA yield (Boessenkool et al. 2012; Sandoval-Velasco et al. 2017). Nowadays, there are dozens of these aDNA laboratory facilities worldwide dedicated to the extraction of

aDNA from a large range of substrates, such as bones (Brown et al. 2016), hair (Gilbert et al. 2007), tanned soft tissues (O’Sullivan et al. 2016), teeth (Wadsworth et al. 2017) and dental plaque (dental calculus) (Warinner, Speller, and Collins 2015), desiccated plant material (Hagenblad et al. 2017), and archaeological artefacts (von Holstein et al. 2014).

1.1.2 Noteworthy archaeological applications of aDNA

The ability to extract aDNA from hundreds of thousands of years old specimens had a deep impact on archaeological sciences. Probably the best known example is the generation of the first draft Neanderthal genome in 2010 (Green et al. 2010) and the ensuing recognition that interbreeding between Neanderthals and modern humans occurred as recently as 47,000-65,000 years ago (Sankararaman et al. 2012), resulting in a Neanderthal contribution of about 1.8-2.6% to all contemporary non-African human populations (Prüfer et al. 2017).

The discovery of another archaic human called Denisovan was only made possible through the sequencing the aDNA extracted from a finger bone (Krause et al. 2010; Reich et al. 2010), as solely bone fragments and teeth of this hominin group have been found (Sawyer et al. 2015; Slon et al. 2017; Chen et al. 2019). It could be shown that several Asian and especially Oceanian modern human populations also have Denisovan ancestry (Sankararaman et al. 2016) and there is furthermore evidence for interbreeding between Neanderthals and Denisovans (Brown et al. 2016), unfolding a highly complex picture of human history that would be unattainable without the analysis of aDNA.

Further aDNA studies also shed light, for instance, on the population turnover in Neolithic Europe (Brace et al. 2019) and the initial peopling of the Americas (Moreno-Mayar et al. 2018), complementing earlier archaeological findings. In addition to aDNA research on humans, palaeogenomics studies of mammals like dogs (Ameen et al. 2019), pigs (Ottoni et al. 2013), and goats (Daly et al. 2018), as well as ancient crops (Ramos-Madrigal et al. 2016; Russell et al. 2016; Ramos-Madrigal et al. 2019), have shaped our knowledge of these archaeologically-relevant species. Nowadays, archaeologists are highly aware of the true value of aDNA studies, and organic material from archaeological sites is often handled with the prospect of potential future biomolecular analyses. The majority of aDNA studies are based on bone or tooth samples. In my three research chapters, I explore the potential of three alternative substrates—hides, fluid-preserved specimens, and birch bark pitch—which I will introduce below.

1.1.3 Hides and skins

Since Russel Higuchi and colleagues were able to extract aDNA from a skin sample of the extinct quagga (Higuchi et al. 1984), hides and skins have been among the most widely used substrates for historical specimens. Due to the fast decomposition of skin after death of the organism, this substrate only preserves under extreme conditions such as low temperatures, anoxia, and dryness and is therefore rarely found in pre-historical context (Brandt et al. 2014). Zoological hides collections however present an important DNA repository of specimens from the last centuries and offer an opportunity to study organisms from historical or even extinct populations. While tooth or bone material, particularly the petrous bone, commonly yields less degraded and more endogenous DNA than skin, the removal of skin patches from museum specimens is less intrusive than the drilling of bones as it is possible to take skin samples while preserving the overall morphology of the specimen, for example by targeting inconspicuous body sites such as toe or paw pads (Burrell et al. 2015).

1.1.4 Fluid-preserved specimens

My first application of metagenomics to historic specimens was using bird intestines preserved in fluid jars. The storage of organic material in preservative fluids has been documented since Babylonian times (Ransome 2004), but only since the mid-17th century have zoological and botanical specimens been fluid-preserved in alcohol for the purpose of almost perfectly maintaining their appearance centuries after their collection (Reid 1994).

Previous attempts to sequence DNA from historical fluid-preserved specimens have produced mixed results. While some studies (Persing et al. 1990; Stuart et al. 2006) were able to successfully retrieve DNA from specimens that had been preserved for up to 100 years, others failed to recover endogenous DNA from specimens that were only preserved for a few weeks (Seutin, White, and Boag 1991). It has been hypothesised that the cause for the unsuccessful attempts is the use of DNA degrading fixative and preservative agents such as formaldehyde. Formaldehyde, or formalin when in an aqueous solution, is prevalent in fluid-preservation due to its exceptional fixation properties, but also causes cross-linking of DNA strands and its usage therefore hinders or even prevents DNA extraction (Fang, Wan, and Fujihara 2002). Nevertheless, the sheer wealth of fluid-preserved specimens presents a great opportunity for studying the microbiomes of historical organisms, such as recovering the

microbial profile of species that have since gone extinct or detecting general changes of the microbiome over time with a time series analysis.

1.1.5 Birch bark pitch

The other microbiome-related substrate examined in this dissertation is ancient birch bark (Fig. 3). Birch pitch/tar/mastic is a viscoelastic material manufactured by heating birch bark in a process called dry distillation. In prehistory this material was probably mostly used as an adhesive, for example to haft axe- or arrowheads (Kozowyk et al. 2017). The oldest examples of birch pitch date back to the Middle Pleistocene and were recovered from Neanderthal sites in Germany (Koller, Baumer, and Mania 2001; Pawlik and Thissen 2011) and Italy (Mazza et al. 2006). Several more lumps of birch pitch dating back to the Mesolithic and Neolithic have also been found in Scandinavia (Hernek and Nordqvist 1995; Regnell et al. 1995), Germany (Schlichtherle and Wahlster 1986; Alexandersen 1989), and Switzerland (Schoch, Kroll, and Pasternak 1995).

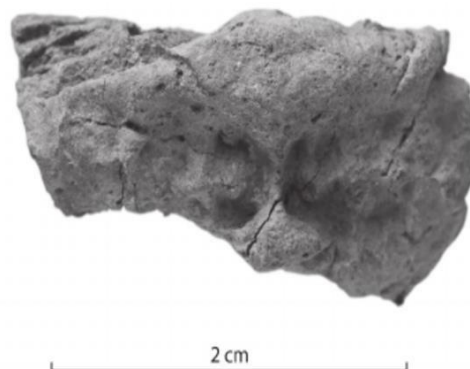


Fig. 1: Birch pitch with visible tooth marks from Raahe, Finland. Taken from Salminen, 2013

Characteristic for these finds are the tooth marks that are often imprinted on the lumps of birch pitch. While it is still unclear why these pieces of birch pitch were chewed, with proposed explanations ranging from dental hygiene to enhanced malleability (Aveling and Heron 1999), it resulted in cells from the oral cavity becoming entrapped in the pitch. Kashuba and colleagues were the first to publicly report on the viability of recovering DNA from chewed birch pitch. They succeeded in obtaining genome-wide data from three of the eight sampled specimens from a site called Huseby Klev that was dated to 10,040-9,610 BP (Kashuba et al. 2019).

1.2 Metagenomic analyses

In my second and third research chapter I expanded the focus of my work from the genome of the organisms themselves to their associated microbial components through computational metagenomic analyses. The term metagenomics was coined by Joshua Lederberg and refers to the study of all available DNA sequences present in environmental samples (Lederberg and McCray 2001). While metabarcoding, which exclusively uses phylogenetic markers such as the prokaryotic 16S rRNA to identify source organisms, is sometimes included under the umbrella term of metagenomics, I specifically refer to shotgun metagenomics, i.e. the untargeted sequencing of all DNA fragments in the sample, when I use the term metagenomics in this introduction.

A common procedure for metagenomic studies begins with the extraction of the DNA present in the sample of interest and converting it to an NGS library. Subsequently this library can be sequenced on an NGS platform, yielding reads that require bioinformatic processing to reveal the information within them.

There are two main approaches to analyze metagenomic data. The first is read-based metagenomics, in which each sequence is assigned to a taxon from a database on the basis of sequence identity. Common strategies to accomplish this are aligning sequence k-mers to a k-mer database, such as KrakenUniq (F. P. Breitwieser, Baker, and Salzberg 2018), aligning the sequences against a clade-specific marker gene database, for example MetaPhlAn (Truong et al. 2015), or aligning all sequences against entire reference genomes, with programs such as MALT (Vågene et al. 2018) or MGMapper (Petersen et al. 2017).

The second, more computationally challenging approach; to analyze metagenomic data is assembly-based metagenomics, in which overlapping sequences are first merged into contigs (Florian P. Breitwieser, Lu, and Salzberg 2017). The original sequences are subsequently aligned to the resulting contigs, which are then grouped into clusters depending on the average depth of alignment and sequence similarity of the alignments. While some of these clusters can then be assigned to known taxa, the true strength of the assembly-based approach is that even taxa which have not been sequenced yet can be detected.

1.2.1 Ancient and historical metagenomes

The ability to identify microbes in ancient and historical samples from DNA traces has added a new dimension to the field of bioarchaeology. While the analysis of endogenous host aDNA enables a look at the genetic makeup of the specimen, the identification of ancient microbes and non-host eukaryotic DNA can furthermore inform on the health and diet of long-dead organisms (Adler et al. 2013; Harbeck et al. 2013). Among others, metagenomic aDNA has been successfully extracted from dental calculus, coprolites, and sediment, which I will briefly discuss below.

Dental calculus is built-up plaque that has mineralised along the gumline (Schroeder 1969; Hardy et al. 2009). During calcification, food remains and microbes from the oral cavity can become entrapped between the layers of calculus. As dental plaque can rapidly mineralize and is relatively resistant to exogenous bacteria (Mann et al. 2018), it provides an excellent source for dietary and oral microbial aDNA (Adler et al. 2013; Weyrich, Dobney, and Cooper 2015).

Coprolites on the other hand are desiccated or fossilised palaeofaeces and have been long appreciated in the field of bioarchaeology as they allow the identification of parasites and food remnants using macroscopic and microscopic inspection. Aside from confirming these findings (Hofreiter et al. 2000) and detecting taxa that could not be detected by visual examination (Wood et al. 2016), metagenomic analyses of coprolites further enable the characterisation of the distal gut microbiome (Tito et al. 2012). The gut microbiome is closely linked to the health of the individual and can therefore provide vital clues to the health status of the individual (Ghaisas, Maher, and Kanthasamy 2016).

Ancient DNA can also be recovered from environmental DNA (eDNA), such as lake and marine sediment, offering a glimpse into the fauna, flora, and microbial diversity of past ecosystems. Sedimentary aDNA (sedaDNA) can preserve relatively well due to anoxia, low temperature, and lack of irradiation (Armbrecht et al. 2019). Indeed, with an estimated age of 400,000 years before present, one of the oldest authenticated aDNA comes from permafrost sediment samples (Willerslev et al. 2003). Besides describing the biodiversity of ancient environments, sedaDNA can also be utilized for a more targeted approach such as in 2016, when Graham and colleagues estimated the extinction time point of woolly mammoths on St. Paul Island by tracking the presence of mammoth DNA across a series of lake sediment cores (Graham et al. 2016).

While dental calculus, coprolites, and sediment are the standard substrates for aDNA metagenomic studies, there could be many other uninvestigated substrates out there of interest for biologists and archaeologists. In chapter three and four I explore the potential of two substrates that have not been used in ancient or historical metagenomic studies before.

1.2.2 Caveats of ancient metagenomic methods

Independent of the substrate are there several challenges for the analysis of ancient and historical metagenomes. De-novo assemblies, i.e the merging of sequences without the use of reference genomes, are not always possible for highly degraded DNA sequences, as most commonly used assembly tools, such as MetaVelvet (Namiki et al. 2012), MEGAHIT (Li et al. 2015), and IDBA-UD (Peng et al. 2012) require relatively long sequences to accurately create contigs.

As the endogenous DNA content of most ancient and historical substrates tends to be low (Poinar et al. 2006; Der Sarkissian et al. 2014), the DNA of bacteria and fungi that colonized the substrate after deposition is often found in high abundance in aDNA data and it can be a challenging task to distinguish between post-mortem contaminants and endogenous taxa (Warinner et al. 2017).

For reference-guided metagenomics approaches, there are major database limitations that need to be considered for the interpretation of the results. While current estimates of the total number of microbial species are in the order of 10^{11} to 10^{12} (Locey and Lennon 2016), as of December 2019 there are just 487,286 microbial reference genomes available on the GenBank sequence database (Benson et al. 2015), and only 17,924 of these are complete reference genomes (GenBank release 235.0, December 2019).

The genomes of microbial taxa that inhabit human tissues, for instance from the Human Microbiome Project (HMP) (Gevers et al. 2012), dominate the reference genome databases, while taxa from highly complex communities such as soil and oceans are severely underrepresented (Quince et al. 2017). The DNA sequences of a species that is not present in the utilised database can therefore be misassigned to a species with high sequence homology, which is especially problematic if the assigned taxon is pathogenic and thus of greater importance for the interpretation of the health status of the individual. Horizontal gene transfer and genome rearrangements present further problems that can cause false positive and false negative assignments (Warinner, Speller, and Collins 2015; Warinner et al. 2017).

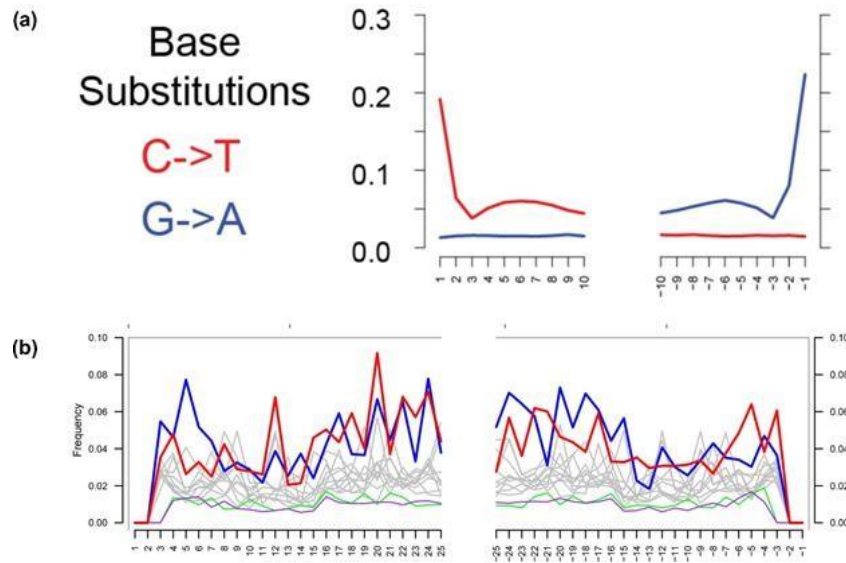


Fig. 2: a) Characteristic deamination pattern of ancient DNA. b) Lack of C-to-T and G-to-A substitutions at the end of the sequences, disproving the presence of ancient DNA damage (Figure taken from Eisenhofer, Cooper, and Weyrich 2017)

One current challenge for metagenomic research is that many reference genomes of eukaryotic taxa are contaminated with the DNA of other organisms (Delmont and Eren 2016; Fierst and Murdock 2017). This happens because eukaryotic genomes tend to be much larger and more repetitive than microbial genomes, making the genome assembly exceptionally challenging and vulnerable to errors such as incorporating the DNA from other organisms present in the DNA extracts. Ultimately, this leads to a significant risk of false positive assignments to eukaryotic taxa in metagenomics studies (Laurence, Hatzis, and Brash 2014; Lu and Salzberg 2018).

Given the multitude of challenges for metagenomics in modern and ancient samples, rigorous authentication of the assignments is crucial in ancient metagenomics. In order to establish that a particular assigned species is not a modern contaminant, mapDamage (Jónsson et al. 2013) is often used to quantify the DNA damage of the aligned sequences (Fig. 4). A further tool called PMDtools (Skoglund et al. 2014) can then be used to filter out sequences that do not exhibit any DNA damage and are therefore potentially derived from contaminants.

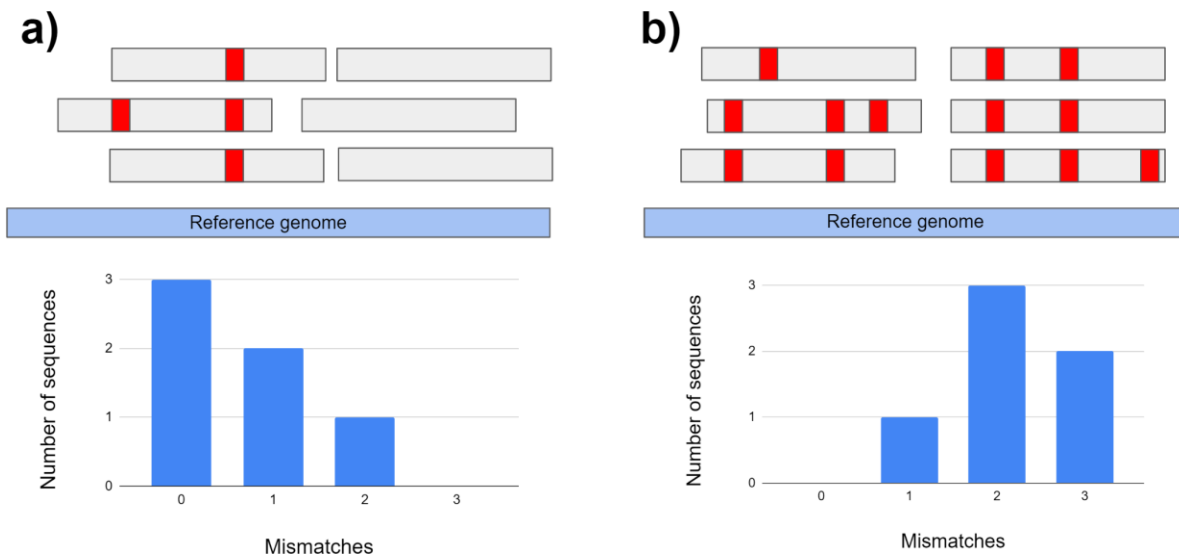


Fig. 3. Schematic illustration of edit distance distribution. Sequences (grey bars) with mismatches (red bars) are visualised in a histogram (bottom). A declining edit distance distribution is expected when the sequences are aligned to the correct reference genome (a). Aligning sequences to the incorrect reference genome shifts the distribution to the right (b). Created by author

Another commonly used statistic to validate a metagenomic assignment is the edit distance distribution, which I implemented extensively in the analysis of historic bird microbiomes and the chewed birch bark pitch. The edit distance encapsulates the number of mismatches of the aligned sequences (Fig. 5). In effect, aligning the DNA sequences to the correct reference genome should result in few or no mismatches per sequence, thus a large proportion of sequences with many mismatches indicates a poor match between the DNA sequences and the reference genome. While it should be cautioned that DNA damage and changes of the genome over time can skew the distribution towards a higher proportion of mismatches per sequence, a declining edit distance distribution indicates that the reference genome belongs to the correct or a closely related species (Key et al. 2017; Huebler et al. 2019).

a



b



Fig. 4. Evenness of coverage should be observed in order to authenticate an assignment (a). Stacking of aligned sequences and large gaps are indicators for a false positive assignment (b). Created by author.

Finally, alignments of true positive assignments should also exhibit a homogenous distribution of aligned sequences across the genome (Lindner et al. 2013; Warinner et al. 2017). If the majority of sequences are aligned to only a few segments of the genome, the DNA sequences were likely misassigned to a species with homologous genomic regions (Fig. 6). While it is feasible to assess the evenness of coverage of the alignment by creating a coverage plot of the whole genome for prokaryotic taxa, the nuclear genomes of eukaryotic species are often too large and the depth of coverage too low to evaluate the distribution of aligned sequences to visualise the genome coverage in that way. Instead, the mitochondrial genome can be used as a proxy, as the copies of the mitochondrial genome exceed that of the nuclear genome several-fold, or the coverage histogram of the nuclear genome can be inspected for an overabundance of sites with a high coverage.

1.3 Thesis structure and objectives

This dissertation is divided into three main research papers, followed by a final summary chapter that reviews the key findings of each.

Chapter two consists of research on the extinct Honshū wolf. In the project I analysed its nuclear genome in order to study the evolutionary history of Japanese wolves and their relationship with past and extant grey wolves and dogs.. The distinct morphology of Honshū wolves and previous mitochondrial studies suggest that they were only distantly related to modern wolves, but until now their ancestry has been unclear.

Chapter three explores the feasibility of recovering the microbiome from fluid-preserved specimens. While the wet collections of natural history museums have been used for morphometric as well as genetic analyses in the past, the profiling of historical microbiomes from ethanol-preserved specimens could provide vital clues to the health and lifestyle of past populations of innumerable species.

Finally, research chapter four encompasses the palaeogenomic and ancient metagenomic aspects of chapter two and three, where I analysed the human, microbial, and dietary DNA recovered from an approximately 5,700-year-old chewed birch bark pitch from Denmark with the aim to investigate the DNA preservation of the different sources in this novel substrate as well as shed light on the population history of the human from Late Mesolithic/Early Neolithic Denmark. birch bark pitch. On top of our discoveries on a human from Late Mesolithic/Early Neolithic Denmark, we were also able to recover the DNA of a multitude of oral microbes and eukaryotic taxa.

References

- Adler, Christina J., Keith Dobney, Laura S. Weyrich, John Kaidonis, Alan W. Walker, Wolfgang Haak, Corey J. A. Bradshaw, et al. 2013. "Sequencing Ancient Calcified Dental Plaque Shows Changes in Oral Microbiota with Dietary Shifts of the Neolithic and Industrial Revolutions." *Nature Genetics* 45 (4): 450–55, 455e1.
- Alexandersen, V. 1989. "Bipuren in Bronzezeitlichen Klumpen von Birkenrindenpech Aus Spjaid." *Acta Archaeologica* 60: 219–23.
- Allentoft, Morten E., and Nicolas J. Rawlence. 2012. "Moa's Ark or Volant Ghosts of Gondwana? Insights from Nineteen Years of Ancient DNA Research on the Extinct Moa (Aves: Dinornithiformes) of New Zealand." *Annals of Anatomy = Anatomischer Anzeiger: Official Organ of the Anatomische Gesellschaft* 194 (1): 36–51.
- Ameen, Carly, Tatiana R. Feuerborn, Sarah K. Brown, Anna Linderholm, Arden Hulme-Beaman, Ophélie Lebrasseur, Mikkel-Holger S. Sinding, et al. 2019. "Specialized Sledge Dogs Accompanied Inuit Dispersal across the North American Arctic." *Proceedings. Biological Sciences / The Royal Society* 286 (1916): 20191929.
- An, Chengcai, Yi Li, Yuxian Zhu, and Xing Shen. 1995. "Molecular Cloning and Sequencing the 18S rDNA From Specialized Dinosaur Egg Fossil Found in Xixia Henan, China." *Chinese Science Abstracts Series B 4 Part B* (14): 51.
- Armbrrecht, Linda H., Marco J. L. Coolen, Franck Lejzerowicz, Simon C. George, Karita Negandhi, Yohey Suzuki, Jennifer Young, et al. 2019. "Ancient DNA from Marine Sediments: Precautions and Considerations for Seafloor Coring, Sample Handling and Data Generation." *Earth-Science Reviews* 196 (September): 102887.
- Austin, Jeremy J., Andrew J. Ross, Andrew B. Smith, Richard A. Fortey, and Richard H. Thomas. 1997. "Problems of Reproducibility – Does Geologically Ancient DNA Survive in Amber–preserved Insects?" *Proceedings of the Royal Society of London. Series B: Biological Sciences*. <https://doi.org/10.1098/rspb.1997.0067>.
- Aveling, E. M., and C. Heron. 1999. "Chewing Tar in the Early Holocene: An Archaeological and Ethnographic Evaluation." *Antiquity* 73 (281): 579–84.
- Benson, Dennis A., Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2015. "GenBank." *Nucleic Acids Research* 43 (Database issue): D30–35.
- Boessenkool, Sanne, Laura S. Epp, James Haile, Eva Bellemain, Mary Edwards, Eric Coissac, Eske Willerslev, and Christian Brochmann. 2012. "Blocking Human Contaminant DNA during PCR Allows Amplification of Rare Mammal Species from Sedimentary Ancient DNA." *Molecular Ecology* 21 (8): 1806–15.
- Brace, Selina, Yoan Diekmann, Thomas J. Booth, Lucy van Dorp, Zuzana Faltyskova, Nadin Rohland, Swapan Mallick, et al. 2019. "Ancient Genomes Indicate Population Replacement in Early Neolithic Britain." *Nature Ecology & Evolution*. <https://doi.org/10.1038/s41559-019-0871-9>.
- Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg. 2017. "A Review of Methods and Databases for Metagenomic Classification and Assembly." *Briefings in Bioinformatics*. <https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbx120/4210288>.
- Breitwieser, F. P., D. N. Baker, and S. L. Salzberg. 2018. "KrakenUniq: Confident and Fast Metagenomics Classification Using Unique K-Mer Counts." *Genome Biology* 19 (1): 198.
- Brown, Samantha, Thomas Higham, Viviane Slon, Svante Pääbo, Matthias Meyer, Katerina Douka, Fiona Brock, et al. 2016. "Identification of a New Hominin Bone from Denisova Cave, Siberia Using Collagen Fingerprinting and Mitochondrial DNA Analysis." *Scientific Reports* 6 (March): 23559.

- Cano, Raúl J., Hendrik Poinar, and George O. Poinar Jr. 1992. "Isolation and Partial Characterization of DNA from the Bee *Proplebeia Dominicana* (Apidae: Hymenoptera) in 25-40 Million Year Old Amber." *Medical Science Research* 20 (7): 249–51.
- Chen, Fahu, Frido Welker, Chuan-Chou Shen, Shara E. Bailey, Inga Bergmann, Simon Davis, Huan Xia, et al. 2019. "A Late Middle Pleistocene Denisovan Mandible from the Tibetan Plateau." *Nature* 569 (7756): 409–12.
- Daly, Kevin G., Pierpaolo Maisano Delser, Victoria E. Mullin, Amelie Scheu, Valeria Mattiangeli, Matthew D. Teasdale, Andrew J. Hare, et al. 2018. "Ancient Goat Genomes Reveal Mosaic Domestication in the Fertile Crescent." *Science* 361 (6397): 85–88.
- Delmont, Tom O., and A. Murat Eren. 2016. "Identifying Contamination with Advanced Visualization and Analysis Practices: Metagenomic Approaches for Eukaryotic Genome Assemblies." *PeerJ* 4 (March): e1839.
- Der Sarkissian, C., L. Ermini, H. Jónsson, A. N. Alekseev, E. Crubezy, B. Shapiro, and L. Orlando. 2014. "Shotgun Microbial Profiling of Fossil Remains." *Molecular Ecology* 23 (7): 1780–98.
- Eisenhofer, Raphael, Alan Cooper, and Laura S. Weyrich. 2017. "Reply to Santiago-Rodriguez et Al.: Proper Authentication of Ancient DNA Is Essential." *FEMS Microbiology Ecology*. <https://doi.org/10.1093/femsec/fix042>.
- Fang, Sheng-Guo, Qiu-Hong Wan, and Noboru Fujihara. 2002. "Formalin Removal from Archival Tissue by Critical Point Drying." *BioTechniques* 33 (3): 604, 606, 608–10.
- Fierst, Janna L., and Duncan A. Murdock. 2017. "Decontaminating Eukaryotic Genome Assemblies with Machine Learning." *BMC Bioinformatics* 18 (1): 533.
- Fry, Erin, Sun K. Kim, Sravanthi Chigurapti, Katelyn M. Mika, Aakrosh Ratan, Alexander Dammermann, Brian J. Mitchell, Webb Miller, and Vincent J. Lynch. 2020. "Functional Architecture of Deleterious Genetic Variants in the Genome of a Wrangel Island Mammoth." *Genome Biology and Evolution*. <https://doi.org/10.1093/gbe/evz279>.
- Fulton, Tara L. 2012. "Setting up an Ancient DNA Laboratory." *Methods in Molecular Biology* 840: 1–11.
- Furholt, Martin. 2018. "Massive Migrations? The Impact of Recent aDNA Studies on Our View of Third Millennium Europe." *European Journal of Archaeology* 21 (2): 159–91.
- Gevers, Dirk, Rob Knight, Joseph F. Petrosino, Katherine Huang, Amy L. McGuire, Bruce W. Birren, Karen E. Nelson, Owen White, Barbara A. Methé, and Curtis Huttenhower. 2012. "The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome." *PLoS Biology* 10 (8): e1001377.
- Ghaisas, Shivani, Joshua Maher, and Anumantha Kanthasamy. 2016. "Gut Microbiome in Health and Disease: Linking the Microbiome-Gut-Brain Axis and Environmental Factors in the Pathogenesis of Systemic and Neurodegenerative Diseases." *Pharmacology & Therapeutics* 158 (February): 52–62.
- Gilbert, M. Thomas P., Lynn P. Tomsho, Snjezana Rendulic, Michael Packard, Daniela I. Drautz, Andrei Sher, Alexei Tikhonov, et al. 2007. "Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts." *Science* 317 (5846): 1927–30.
- Graham, Russell W., Soumaya Belmecheri, Kyungcheol Choy, Brendan J. Culleton, Lauren J. Davies, Duane Froese, Peter D. Heintzman, et al. 2016. "Timing and Causes of Mid-Holocene Mammoth Extinction on St. Paul Island, Alaska." *Proceedings of the National Academy of Sciences of the United States of America* 113 (33): 9310–14.
- Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22.
- Guiry, Eric J., Trevor J. Orchard, Thomas C. A. Royle, Christina Cheung, and Dongya Y. Yang. 2020. "Dietary Plasticity and the Extinction of the Passenger Pigeon (*Ectopistes Migratorius*)." *Quaternary Science Reviews* 233 (April): 106225.

- Hagenblad, Jenny, Jacob Morales, Matti W. Leino, and Amelia C. Rodríguez-Rodríguez. 2017. "Farmer Fidelity in the Canary Islands Revealed by Ancient DNA from Prehistoric Seeds." *Journal of Archaeological Science* 78 (February): 78–87.
- Hansen, Henrik B., Peter B. Damgaard, Ashot Margaryan, Jesper Stenderup, Niels Lynnerup, Eske Willerslev, and Morten E. Allentoft. 2017. "Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum." *PloS One* 12 (1): e0170940.
- Harbeck, Michaela, Lisa Seifert, Stephanie Hänsch, David M. Wagner, Dawn Birdsell, Katy L. Parise, Ingrid Wiechmann, et al. 2013. "Yersinia Pestis DNA from Skeletal Remains from the 6th Century AD Reveals Insights into Justinianic Plague." *PLoS Pathogens*. <https://doi.org/10.1371/journal.ppat.1003349>.
- Hardy, Karen, Tony Blakeney, Les Copeland, Jennifer Kirkham, Richard Wrangham, and Matthew Collins. 2009. "Starch Granules, Dental Calculus and New Perspectives on Ancient Diet." *Journal of Archaeological Science* 36 (2): 248–55.
- Hebsgaard, Martin B., Matthew J. Phillips, and Eske Willerslev. 2005. "Geologically Ancient DNA: Fact or Artefact?" *Trends in Microbiology* 13 (5): 212–20.
- Hernek, Robert, and Bengt Nordqvist. 1995. "Världens äldsta Tuggummi." *Ett Urval Spännande Arkeologiska Fynd Och Upptäckter Som Gjordes Vid Huseby Klev, Och Andra Platser, Inför Väg 178*.
- Higuchi, R., B. Bowman, M. Freiberger, O. A. Ryder, and A. C. Wilson. 1984. "DNA Sequences from the Quagga, an Extinct Member of the Horse Family." *Nature* 312 (5991): 282–84.
- Hofreiter, M., H. N. Poinar, W. G. Spaulding, K. Bauer, P. S. Martin, G. Possnert, and S. Pääbo. 2000. "A Molecular Analysis of Ground Sloth Diet through the Last Glaciation." *Molecular Ecology* 9 (12): 1975–84.
- Holstein, Isabella C. C. von, Steven P. Ashby, Nienke L. van Doorn, Stacie M. Sachs, Michael Buckley, Meirav Meiri, Ian Barnes, Anne Brundle, and Matthew J. Collins. 2014. "Searching for Scandinavians in Pre-Viking Scotland: Molecular Fingerprinting of Early Medieval Combs." *Journal of Archaeological Science* 41 (January): 1–6.
- Huebler, Ron, Felix M. M. Key, Christina Warinner, Kirsten I. Bos, Johannes Krause, and Alexander Herbig. 2019. "HOPS: Automated Detection and Authentication of Pathogen DNA in Archaeological Remains." *bioRxiv*. <https://doi.org/10.1101/534198>.
- Jónsson, Hákon, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, and Ludovic Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics* 29 (13): 1682–84.
- Kashuba, Natalija, Emrah Kırdök, Hege Damlien, Mikael A. Manninen, Bengt Nordqvist, Per Persson, and Anders Götherström. 2019. "Ancient DNA from Mastics Solidifies Connection between Material Culture and Genetics of Mesolithic Hunter–gatherers in Scandinavia." *Communications Biology* 2 (1): 185.
- Key, Felix M., Cosimo Posth, Johannes Krause, Alexander Herbig, and Kirsten I. Bos. 2017. "Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication." *Trends in Genetics: TIG* 33 (8): 508–20.
- Koller, Johann, Ursula Baumer, and Dietrich Mania. 2001. "High-Tech in the Middle Palaeolithic: Neandertal-Manufactured Pitch Identified." *European Journal of Archaeology* 4 (3): 385–97.
- Kozowyk, P. R. B., M. Soressi, D. Pomstra, and G. H. J. Langejans. 2017. "Experimental Methods for the Palaeolithic Dry Distillation of Birch Bark: Implications for the Origin and Development of Neandertal Adhesive Technology." *Scientific Reports* 7 (1): 8033.
- Krause, Johannes, Qiaomei Fu, Jeffrey M. Good, Bence Viola, Michael V. Shunkov, Anatoli P. Derevianko, and Svante Pääbo. 2010. "The Complete Mitochondrial DNA Genome of an Unknown Hominin from Southern Siberia." *Nature* 464 (7290): 894–97.
- Laurence, Martin, Christos Hatzis, and Douglas E. Brash. 2014. "Common Contaminants in next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes." *PloS One* 9 (5): e97876.

- Lederberg, Joshua, and Alexa T. McCray. 2001. "Ome SweetOmics--A Genealogical Treasury of Words." *Scientist* 15 (7): 8–8.
- Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76.
- Lindner, Martin S., Maximilian Kollock, Franziska Zickmann, and Bernhard Y. Renard. 2013. "Analyzing Genome Coverage Profiles with Applications to Quality Control in Metagenomics." *Bioinformatics* 29 (10): 1260–67.
- Locey, Kenneth J., and Jay T. Lennon. 2016. "Scaling Laws Predict Global Microbial Diversity." *Proceedings of the National Academy of Sciences of the United States of America* 113 (21): 5970–75.
- Lu, Jennifer, and Steven L. Salzberg. 2018. "Removing Contaminants from Databases of Draft Genomes." *PLOS Computational Biology*. <https://doi.org/10.1371/journal.pcbi.1006277>.
- Mann, Allison E., Susanna Sabin, Kirsten Ziesemer, Åshild J. Vågane, Hannes Schroeder, Andrew T. Ozga, Krithivasan Sankaranarayanan, et al. 2018. "Differential Preservation of Endogenous Human and Microbial DNA in Dental Calculus and Dentin." *Scientific Reports* 8 (1): 9822.
- Mazza, Paul Peter Anthony, Fabio Martini, Benedetto Sala, Maurizio Magi, Maria Perla Colombini, Gianna Giachi, Francesco Landucci, Cristina Lemorini, Francesca Modugno, and Erika Ribechini. 2006. "A New Palaeolithic Discovery: Tar-Hafted Stone Tools in a European Mid-Pleistocene Bone-Bearing Bed." *Journal of Archaeological Science* 33 (9): 1310–18.
- Millar, Craig D., Leon Huynen, Sankar Subramanian, Elmira Mohandesan, and David M. Lambert. 2008. "New Developments in Ancient Genomics." *Trends in Ecology & Evolution* 23 (7): 386–93.
- Miller, Webb, Daniela I. Drautz, Aakrosh Ratan, Barbara Pusey, Ji Qi, Arthur M. Lesk, Lynn P. Tomsho, et al. 2008. "Sequencing the Nuclear Genome of the Extinct Woolly Mammoth." *Nature* 456 (7220): 387–90.
- Mitchell, David, Eske Willerslev, and Anders Hansen. 2005. "Damage and Repair of Ancient DNA." *Mutation Research* 571 (1-2): 265–76.
- Moreno-Mayar, J. Víctor, Lasse Vinner, Peter de Barros Damgaard, Constanza de la Fuente, Jeffrey Chan, Jeffrey P. Spence, Morten E. Allentoft, et al. 2018. "Early Human Dispersals within the Americas." *Science* 362 (6419). <https://doi.org/10.1126/science.aav2621>.
- Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. "MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads." *Nucleic Acids Research* 40 (20): e155.
- Narasimhan, Vagheesh M., Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, et al. 2019. "The Formation of Human Populations in South and Central Asia." *Science* 365 (6457). <https://doi.org/10.1126/science.aat7487>.
- O'Sullivan, Niall J., Matthew D. Teasdale, Valeria Mattiangeli, Frank Maixner, Ron Pinhasi, Daniel G. Bradley, and Albert Zink. 2016. "A Whole Mitochondria Analysis of the Tyrolean Iceman's Leather Provides Insights into the Animal Sources of Copper Age Clothing." *Scientific Reports* 6 (August): 31279.
- Otoni, Claudio, Linus Girdland Flink, Allowen Evin, Christina Geörg, Bea De Cupere, Wim Van Neer, László Bartosiewicz, et al. 2013. "Pig Domestication and Human-Mediated Dispersal in Western Eurasia Revealed through Ancient DNA and Geometric Morphometrics." *Molecular Biology and Evolution* 30 (4): 824–32.
- Pääbo, Svante. 1985. "Preservation of DNA in Ancient Egyptian Mummies." *Journal of Archaeological Science* 12 (6): 411–17.
- Palkopoulou, Eleftheria, Swapan Mallick, Pontus Skoglund, Jacob Enk, Nadin Rohland, Heng Li, Ayça Omrak, et al. 2015. "Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth." *Current Biology: CB* 25 (10): 1395–1400.

- Pawlik, Alfred F., and Jürgen P. Thissen. 2011. "Hafted Armatures and Multi-Component Tool Design at the Micoquian Site of Inden-Altdorf, Germany." *Journal of Archaeological Science* 38 (7): 1699–1708.
- Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. "IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth." *Bioinformatics* 28 (11): 1420–28.
- Persing, D. H., S. R. Telford 3rd, P. N. Rys, D. E. Dodge, T. J. White, S. E. Malawista, and A. Spielman. 1990. "Detection of *Borrelia burgdorferi* DNA in Museum Specimens of Ixodes Dammini Ticks." *Science* 249 (4975): 1420–23.
- Petersen, Thomas Nordahl, Oksana Lukjancenko, Martin Christen Frølund Thomsen, Maria Maddalena Sperotto, Ole Lund, Frank Møller Aarestrup, and Thomas Sicheritz-Pontén. 2017. "MGmapper: Reference Based Mapping and Taxonomy Annotation of Metagenomics Sequence Reads." *PLoS One* 12 (5): e0176469.
- Poinar, Hendrik N., Raul J. Cano, and George O. Poinar. 1993. "DNA from an Extinct Plant." *Nature* 363 (6431): 677–677.
- Poinar, Hendrik N., Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D. E. Macphee, Bernard Buigues, Alexei Tikhonov, et al. 2006. "Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA." *Science* 311 (5759): 392–94.
- Prüfer, Kay, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlević, Mateja Hajdinjak, Benjamin Vernot, et al. 2017. "A High-Coverage Neandertal Genome from Vindija Cave in Croatia." *Science* 358 (6363): 655–58.
- Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. "Shotgun Metagenomics, from Sampling to Analysis." *Nature Biotechnology* 35 (9): 833–44.
- Ramos-Madrigal, Jazmín, Anne Kathrine Wiborg Runge, Laurent Bouby, Thierry Lacombe, José Alfredo Samaniego Castruita, Anne-Françoise Adam-Blondon, Isabel Figueiral, et al. 2019. "Palaeogenomic Insights into the Origins of French Grapevine Diversity." *Nature Plants* 5 (6): 595–603.
- Ramos-Madrigal, Jazmín, Bruce D. Smith, J. Víctor Moreno-Mayar, Shyam Gopalakrishnan, Jeffrey Ross-Ibarra, M. Thomas P. Gilbert, and Nathan Wales. 2016. "Genome Sequence of a 5,310-Year-Old Maize Cob Provides Insights into the Early Stages of Maize Domestication." *Current Biology: CB* 26 (23): 3195–3201.
- Ransome, Hilda M. 2004. *The Sacred Bee in Ancient Times and Folklore*. Courier Corporation.
- Regnell, Mats, Marie-José Gaillard, Thomas Seip Bartholin, and Per Karsten. 1995. "Reconstruction of Environment and History of Plant Use during the Late Mesolithic (Ertebølle Culture) at the Inland Settlement of Bökeberg III, Southern Sweden." *Vegetation History and Archaeobotany* 4 (2): 67–91.
- Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60.
- Reid, Gordon. 1994. "The Preparation and Preservation of Collections." In *Manual of Natural History Curatorship*, 28–56.
- Roca, Alfred L., Yasuko Ishida, Nikolas Nikolaidis, Sergios-Orestis Kolokotronis, Stephen Fratpietro, Kristin Stewardson, Shannon Hensley, Michele Tisdale, Gennady Boeskorov, and Alex D. Greenwood. 2009. "Genetic Variation at Hair Length Candidate Genes in Elephants and the Extinct Woolly Mammoth." *BMC Evolutionary Biology* 9 (September): 232.
- Rogers, Rebekah L., and Montgomery Slatkin. 2017. "Excess of Genomic Defects in a Woolly Mammoth on Wrangel Island." *PLoS Genetics* 13 (3): e1006601.
- Rollo, F., A. Amici, R. Salvi, and A. Garbuglia. 1988. "Short but Faithful Pieces of Ancient DNA." *Nature* 335 (6193): 774.

- Russell, Joanne, Martin Mascher, Ian K. Dawson, Stylianos Kyriakidis, Cristiane Calixto, Fabian Freund, Micha Bayer, et al. 2016. “Exome Sequencing of Geographically Diverse Barley Landraces and Wild Relatives Gives Insights into Environmental Adaptation.” *Nature Genetics* 48 (9): 1024–30.
- Salminen, Timo. 2013. “Man, His Time, Artefacts, and Places: Collection of Articles Dedicated to Richard Indreko.” *Fornvännen* 108 (4): 290–91.
- Sánchez-Quinto, Federico, and Carles Lalueza-Fox. 2015. “Almost 20 Years of Neanderthal Palaeogenetics: Adaptation, Admixture, Diversity, Demography and Extinction.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1660): 20130374.
- Sandoval-Velasco, Marcela, Inge K. C. Lundstrøm, Nathan Wales, María C. Ávila-Arcos, Hannes Schroeder, and M. Thomas P. Gilbert. 2017. “Relative Performance of Two DNA Extraction and Library Preparation Methods on Archaeological Human Teeth Samples.” *STAR: Science & Technology of Archaeological Research* 3 (1): 80–88.
- Sankararaman, Sriram, Swapan Mallick, Nick Patterson, and David Reich. 2016. “The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans.” *Current Biology: CB* 26 (9): 1241–47.
- Sankararaman, Sriram, Nick Patterson, Heng Li, Svante Pääbo, and David Reich. 2012. “The Date of Interbreeding between Neandertals and Modern Humans.” *PLoS Genetics* 8 (10): e1002947.
- Sawyer, Susanna, Gabriel Renaud, Bence Viola, Jean-Jacques Hublin, Marie-Theres Gansauge, Michael V. Shunkov, Anatoly P. Derevianko, Kay Prüfer, Janet Kelso, and Svante Pääbo. 2015. “Nuclear and Mitochondrial DNA Sequences from Two Denisovan Individuals.” *Proceedings of the National Academy of Sciences of the United States of America* 112 (51): 15696–700.
- Schlichtherle, Helmut, and Barbara Wahlster. 1986. “Archäologie in Seen Und Mooren.” *Den Pfahlbauten Auf Der Spur. Theiss, Stuttgart*.
- Schoch, W., H. Kroll, and R. Pasternak. 1995. “Analysis of Plant Glue from the Stone and Bronze Ages.” In *Res Archaeobotanicae. Proceed 9th Symp Internat Workgroup Palaeoethnobot Kiel*, 301–8.
- Schroeder, H. E. 1969. “Formation and Inhibition of Dental Calculus.” *Journal of Periodontology* 40 (11): 643–46.
- Seutin, Gilles, Bradley N. White, and Peter T. Boag. 1991. “Preservation of Avian Blood and Tissue Samples for DNA Analyses.” *Canadian Journal of Zoology* 69 (1): 82–90.
- Skoglund, Pontus, Bernd H. Northoff, Michael V. Shunkov, Anatoli P. Derevianko, Svante Pääbo, Johannes Krause, and Mattias Jakobsson. 2014. “Separating Endogenous Ancient DNA from Modern Day Contamination in a Siberian Neandertal.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (6): 2229–34.
- Slon, Viviane, Bence Viola, Gabriel Renaud, Marie-Theres Gansauge, Stefano Benazzi, Susanna Sawyer, Jean-Jacques Hublin, et al. 2017. “A Fourth Denisovan Individual.” *Science Advances* 3 (7): e1700186.
- Stuart, Bryan L., Kerri A. Dugan, Marc W. Allard, and Maureen Kearney. 2006. “Extraction of Nuclear DNA from Bone of Skeletonized and Fluid-preserved Museum Specimens.” *Systematics and Biodiversity* 4 (2): 133–36.
- Tito, Raul Y., Dan Knights, Jessica Metcalf, Alexandra J. Obregon-Tito, Lauren Cleeland, Fares Najjar, Bruce Roe, et al. 2012. “Insights from Characterizing Extinct Human Gut Microbiomes.” *PloS One* 7 (12): e51146.
- Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. “MetaPhlan2 for Enhanced Metagenomic Taxonomic Profiling.” *Nature Methods* 12 (10): 902–3.

- Vågene, Åshild J., Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, et al. 2018. “Salmonella Enterica Genomes from Victims of a Major Sixteenth-Century Epidemic in Mexico.” *Nature Ecology & Evolution* 2 (3): 520–28.
- Wadsworth, Caroline, Noemi Procopio, Cecilia Anderung, José-Miguel Carretero, Eneko Iriarte, Cristina Valdiosera, Rengert Elburg, Kirsty Penkman, and Michael Buckley. 2017. “Comparing Ancient DNA Survival and Proteome Content in 69 Archaeological Cattle Tooth and Bone Samples from Multiple European Sites.” *Journal of Proteomics* 158 (March): 1–8.
- Warinner, Christina, Alexander Herbig, Allison Mann, James A. Fellows Yates, Clemens L. Weiß, Hernán A. Burbano, Ludovic Orlando, and Johannes Krause. 2017. “A Robust Framework for Microbial Archaeology.” *Annual Review of Genomics and Human Genetics* 18 (August): 321–56.
- Warinner, Christina, Camilla Speller, and Matthew J. Collins. 2015. “A New Era in Palaeomicrobiology: Prospects for Ancient Dental Calculus as a Long-Term Record of the Human Oral Microbiome.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1660): 20130376.
- Weyrich, Laura S., Keith Dobney, and Alan Cooper. 2015. “Ancient DNA Analysis of Dental Calculus.” *Journal of Human Evolution* 79 (February): 119–24.
- White, Lauren C., Kieren J. Mitchell, and Jeremy J. Austin. 2018. “Ancient Mitochondrial Genomes Reveal the Demographic History and Phylogeography of the Extinct, Enigmatic Thylacine (*Thylacinus Cynocephalus*).” *Journal of Biogeography* 45 (1): 1–13.
- Willerslev, Eske, Anders J. Hansen, Jonas Binladen, Tina B. Brand, M. Thomas P. Gilbert, Beth Shapiro, Michael Bunce, Carsten Wiuf, David A. Gilichinsky, and Alan Cooper. 2003. “Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments.” *Science* 300 (5620): 791–95.
- Wood, Jamie R., Andrea Crown, Theresa L. Cole, and Janet M. Wilmshurst. 2016. “Microscopic and Ancient DNA Profiling of Polynesian Dog (*kuī*) Coprolites from Northern New Zealand.” *Journal of Archaeological Science: Reports* 6 (April): 496–505.
- Yang, Dongya Y., and Kathy Watt. 2005. “Contamination Controls When Preparing Archaeological Remains for Ancient DNA Analysis.” *Journal of Archaeological Science* 32 (3): 331–36.

PhD contributions

Chapter 2: Complete genome of historic Honshū wolf reveals Pleistocene heritage

This chapter is a manuscript that has been submitted for publication. Mikkel Sinding performed the sampling and laboratory work. Shyam Gopalakrishnan performed the adapter removal and alignment of the raw data. For this project I carried out all of the remaining bioinformatic analyses, i.e. the admixture analyses, the D-statistics analyses, the phasing of the data, the haplotype-aware clustering of the phased data, and the estimation of past admixture events. I furthermore created all figures and tables. I decided the best analysis strategy with Shyam Gopalakrishnan and interpreted the results with input from Shyam Gopalakrishnan and Mikkel Sinding. I wrote the manuscript with contributions from Nathan Wales, Tom Gilbert, Mikkel Sinding, and Shyam Gopalakrishnan.

Chapter 3: Unsealing the jars - characterizing gut microbial DNA preservation in fluid-preserved museum specimens

Chapter 3 is the first draft of an ongoing project. Jessica Thomas performed the sampling and Jessica Thomas and Marcela Sandoval carried out the laboratory work. I performed all of the bioinformatic analyses, i.e. the adapter removal and quality control of the raw data, the alignment of the data to the various reference genomes, the metataxonomic assignments, the principal coordinate analysis, and the authentication of the assignments. I furthermore created all figures and tables. I decided the best analysis strategy with input from Nathan Wales. I wrote the manuscript with contributions from Nathan Wales and Tom Gilbert.

Chapter 4: A 5700 year-old human genome and oral microbiome from chewed birch pitch

Mikkel Sinding and Theis Jensen performed the sampling and laboratory work. Hannes Schroeder carried out the adapter removal of the raw data, the alignment to the human reference genome, and the F-statistics analysis. I carried out the admixture and principal component analysis. Shyam Gopalakrishnan and I carried out the phenotype prediction analysis. Anna Fotakis performed the MetaPhlan analysis. Åshild Vågane performed the MALT analysis. Mikkel Winther Pedersen performed the HOLI analysis. I aligned the assigned sequences from MALT and HOLI to the respective reference genomes, decided the best authentication strategy, and validated the assignments to microbial and eukaryotic taxa. Katrine Højholt Iversen carried out all of the virulence analyses. I created Figure 2, 3, and 4, Supplementary Figures 8-15, and Table 1 and Supplementary Table 1 with Theis Jensen. Hannes Schroeder wrote the main manuscript with input from the first authors. I authored the Supplementary Text on authenticating metagenomics assignments with input from Hannes Schroeder.

Chapter 2

Complete genome of historic Honshū wolf reveals Pleistocene heritage

Introduction to the study organism

Evolutionary history of wolves

One chapter of my dissertation is explicitly focused on the evolutionary history of wolves and dogs. The origin of dogs and their relationships with wolf populations is a longstanding question for archaeologists and evolutionary biology.

No carnivore has shaped human history as much as the grey wolf (*Canis lupus*). Feared and persecuted in most of Eurasia throughout history, its domesticated form, the dog (*Canis lupus familiaris*), has been our closest companion for thousands of years, facilitating several human activities, such as hunting, guarding, herding, and transportation (Lord, Schneider, and Coppinger 2016). But even though the grey wolf is among the most researched organisms, much of its population history is still disputed.

The earliest fossil records of *C. lupus* were found in Alaska and Siberia and date back to the early Pleistocene, suggesting that the origin of wolves lies in Beringia (Tedford, Wang, and Taylor 2009).

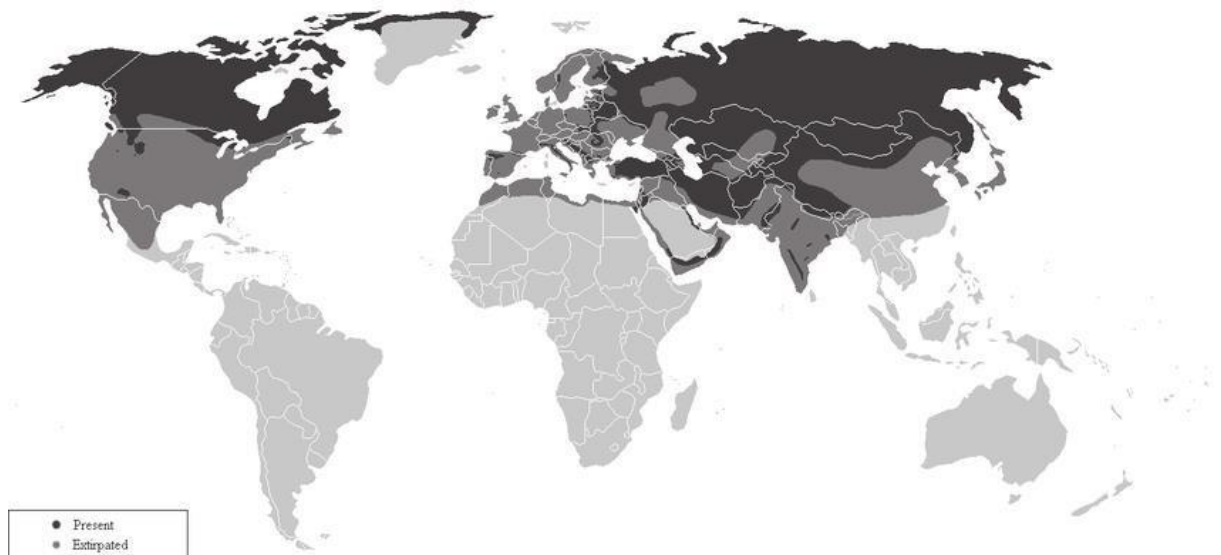


Fig. 1: The present (darkest grey) and past (lighter grey) worldwide distribution of the grey wolf (*Canis lupus*). Taken from (Jansson 2013)

Fossil records point to the presence of multiple wolf populations that inhabited the mammoth steppe in Siberia during the Late Pleistocene and Early Holocene. Isotope and morphometric analyses of skull and dentition features, suggest that unlike extant grey wolves, the Siberian Pleistocene wolf had a hypercarnivorous diet and was more specialized in hunting large prey (Baryshnikov, Mol, and Tikhonov 2009; Leonard et al. 2007). With the disappearance of the steppe habitat in the Late Pleistocene, the Siberian Pleistocene populations—along with its megafaunal prey such as the woolly mammoth—went extinct and were subsequently replaced by present-day grey wolves (Leonard et al. 2007). Nowadays, these can be found across the holarctic in vastly different ecosystems, such as the deserts of the Arabian peninsula or the Tibetan highlands (Fig. 1). However, due to conflicts with livestock owners and their reputation as man-killing beasts, grey wolves have been severely persecuted in the past two centuries and have been eradicated in large parts of their former continental ranges of Western Europe and North America, as well as islands such as those of Japan (Fritts et al. 2003). The first chapter of this thesis explores the relationship between one extinct grey wolf subspecies from Japan and ancient and modern grey wolves and dogs.

The Japanese wolf

Within the broad context of wolf evolution, I explored the history of one of the most enigmatic wolf populations, the Honshū wolf or Japanese wolf (*Canis lupus hodophilax*). This wolf population was endemic to the Japanese islands of Honshū, Kyushu, and Shikoku (Fig. 2). It was one of the smallest subspecies of grey wolf, as it stood only 56-58 cm to the withers. In fact, its morphological appearance was so strikingly different from continental grey wolves that its taxonomic status used to be controversial, and led Yoshinori Imaizuma to argue that the Japanese wolf should be considered a distinct species (Imaizuma 1970). This has since been disproven, however, based on mitogenomic studies.

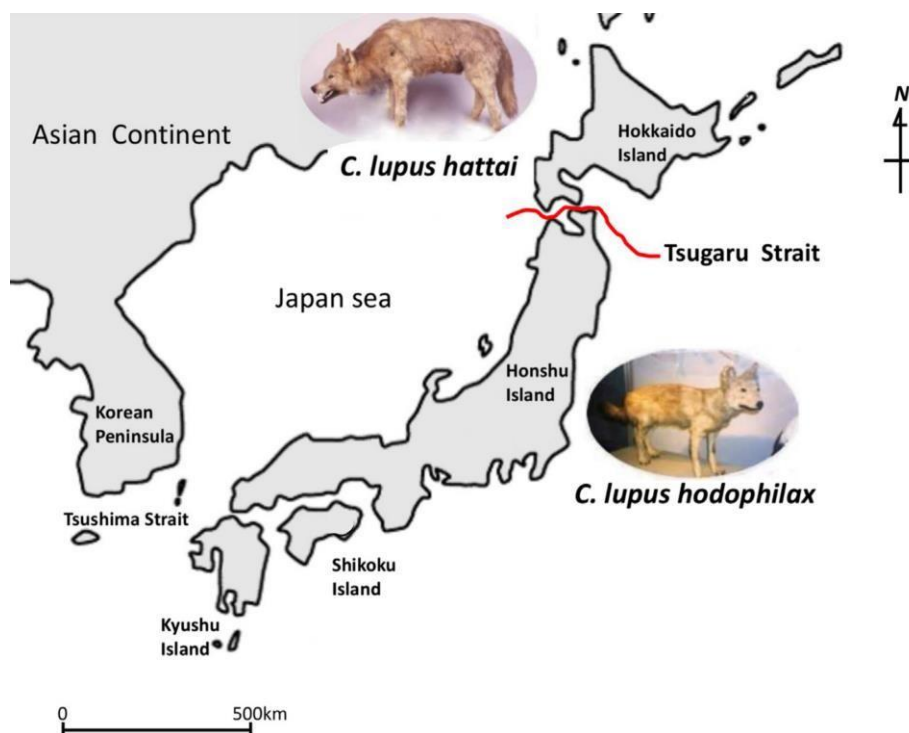


Fig. 2. The former habitat of the Japanese wolf (*C. lupus hodophilax*) extended over three of Japan's main islands: Honshū, Shikoku, and Kyushu. The Tsugaru strait separated the habitat of the Japanese wolf and the Hokkaido wolf (*C. lupus hattai*). Modified from (Matsumura, Inoshima, and Ishiguro 2014)

Unlike wolves in most of Eurasia, Honshū wolves were not perceived as a threat in medieval Japan. Instead they were appreciated for killing crop-damaging wildlife such as deer and boars, revered as deities with dedicated shrines, and the birth of pups was occasion for celebration (Fritts et al. 2003). Their reputation as benign guardians of travellers and farmers changed dramatically, however, when rabid dogs were introduced from Korea to Japan in the

late 17th century and infected the local wolf population. This led to a sharp increase in wolf attacks on livestock and humans, and in response the first bounty system was introduced in 1701. When Japan underwent fundamental societal changes during the Meiji restoration (1868-1912), the sudden rise of deforestation and urbanization added fuel to the fire, and the eradication of wolves in Japan became a national policy. The combination of firearms and baits laced with strychnine proved to be so effective that the population collapsed within one generation, and the last Honshū wolf reportedly died in 1905 in Nara Prefecture (Walker 2009).

Today, a few skins and bones are all that remain of the Honshū wolf. Previous research has revealed poor DNA preservation in several of these due to conservation practices that never accounted for future aDNA studies (Walker 2009). Nevertheless, Ishiguro and colleagues were able to extract and sequence the mitochondrial D-loop control region of seven Honshū wolf bone samples. The generated data was then compared to the control region of 78 dogs, and the authors observed that some domestic Japanese dog individuals carried the Honshū wolf haplotype, suggesting that male dogs and female wolves hybridized in the past, which is highly unusual for wolf-dog hybridizations (Ishiguro, Inoshima, and Shigehara 2009). Further studies of the Japanese wolf mitogenome by Matsamura and colleagues revealed that the Japanese wolves appear to be ancestral to extant grey wolf populations, and colonized the Japanese archipelago about 25,000–125,000 years before present (Matsamura, Inoshima, and Ishiguro 2014). A later study by Koblmüller and colleagues confirmed the ancestral phylogenetic placement, but based on a phylogenomic study with a larger wolf reference panel combined with the sea level changes in the Tsushima strait between Korea and Japan, they estimated a much later colonization event of less than 20,000 years ago (Koblmüller et al. 2016). All previous studies on the Honshu wolf are based on the mitochondrial genome, which is a single genetic marker from the maternal lineage and consequently does not allow an in-depth analysis of admixture and gene flow. We therefore generated nuclear DNA from a Honshu wolf museum specimen that was dated to the 19th century in order to gather new insights into the population history of this enigmatic subspecies in light of a dataset comprising of modern and ancient wolf and dog genomes.

References

- Baryshnikov, Gennady F., Dick Mol, and Alexei N. Tikhonov. 2009. "Finding of the Late Pleistocene Carnivores in Taimyr Peninsula (Russia, Siberia) with Paleocological Context." *Russian Journal of Theriology* 8 (2): 107–13.
- Fritts, Steven H., Robert O. Stephenson, Robert D. Hayes, and Luigi Boitani. 2003. "Wolves and Humans," USGS Northern Prairie Wildlife Research Center, .
<https://digitalcommons.unl.edu/usgsnpwrc/317/>.
- Ishiguro, N., Y. Inoshima, and N. Shigehara. 2009. "Mitochondrial DNA Analysis of the Japanese Wolf (*Canis Lupus Hodophilax* Temminck, 1839) and Comparison with Representative Wolf and Domestic Dog" *Zoological Science*. <http://www.bioone.org/doi/abs/10.2108/zsj.26.765>.
- Jansson, Eeva. 2013. "Past and Present Genetic Diversity and Structure of the Finnish Wolf Population." *Acta Universitatis Ouluensis. Series A. Scientiae Rerum Naturalium* 608.
- Koblmüller, Stephan, Carles Vilà, Belen Lorente-Galdos, Marc Dabad, Oscar Ramirez, Tomas Marques-Bonet, Robert K. Wayne, and Jennifer A. Leonard. 2016. "Whole Mitochondrial Genomes Illuminate Ancient Intercontinental Dispersals of Grey Wolves (*Canis Lupus*)." *Journal of Biogeography* 43 (9): 1728–38.
- Leonard, Jennifer A., Carles Vilà, Kena Fox-Dobbs, Paul L. Koch, Robert K. Wayne, and Blaire Van Valkenburgh. 2007. "Megafaunal Extinctions and the Disappearance of a Specialized Wolf Ecomorph." *Current Biology: CB* 17 (13): 1146–50.
- Lord, Kathryn, Richard A. Schneider, and Raymond Coppinger. 2016. "Evolution of Working Dogs." *The Domestic Dog*. <https://doi.org/10.1017/9781139161800.004>.
- Matsumura, Shuichi, Yasuo Inoshima, and Naotaka Ishiguro. 2014. "Reconstructing the Colonization History of Lost Wolf Lineages by the Analysis of the Mitochondrial Genome." *Molecular Phylogenetics and Evolution* 80 (November): 105–12.
- Tedford, Richard H., Xiaoming Wang, and Beryl E. Taylor. 2009. "Phylogenetic Systematics of the North American Fossil Caninae (Carnivora: Canidae)." *Bulletin of the American Museum of Natural History* 325 (September): 1–218.
- Walker, Brett L. 2009. *The Lost Wolves of Japan*. University of Washington Press.
- Imaizumi, Yoshinori. 1970. "ニホンオオカミの系統的地位について." *Journal of the Mammalogical Society of Japan* 5 (1): 27–32.

Complete genome of historic Honshū wolf reveals Pleistocene heritage

Authors: Jonas Niemann^{1,2}, Mikkel-Holger S Sinding^{1,3,4,5}, Shyam Gopalakrishnan¹, Nobuyuki Yamaguchi⁶, Jazmín Ramos-Madrigal¹, Nathan Wales², M Thomas P Gilbert¹

Affiliations:

¹The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark.

²BioArch, Department of Archaeology, University of York, York, UK.

³The Qimmeq Project, University of Greenland, Nuussuaq, Greenland.

⁴Greenland Institute of Natural Resources, Nuuk, Greenland.

⁵Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland.

⁶Institute of Tropical Biodiversity and Sustainable Development, University Malaysia Terengganu, 21030 Kuala Terengganu, Malaysia

Abstract

The Japanese or Honshū wolf was one the most distinct grey wolf subspecies due to its small stature and endemism to the islands of Honshū, Shikoku, and Kyūshū. Long revered as a guardian of farmers and travellers, it was persecuted from the 17th century following a rabies epidemic, which led to its extinction in the early 20th century. Its uncertain relationship with present-day and ancient wolf populations has puzzled researchers for years, although research based on mitochondrial genomes suggests a basal placement of the Honshū wolf to all extant grey wolves. To refine the evolutionary history of the species, we sequenced the nuclear genome of one Honshū wolf specimen from the 19th century to an average depth of coverage of 3.7×. We find Honshū wolves were closely related to a lineage of Siberian wolves that went extinct in the Late Pleistocene, thereby extending the survival of this ancient lineage by over 10,000 years. We also detected significant gene flow between Japanese dogs and the Honshū wolf, corroborating previous reports on Honshū wolf dog interbreeding.

Introduction

The origin of present-day dogs and Eurasian wolves is highly contentious, as genomic analyses using both modern and ancient wolf samples have yet to identify any sample that pre-dates the Holocene (ca. 10,000 year ago) that genetically resemble their modern forms. Rather, the genomic data published to date from Pleistocene Eurasian wolf subfossil records points to the presence of a genetically more basal megafaunal wolf population across the Holarctic, that ceased to exist at the end of the Pleistocene (Skoglund et al. 2015). At present, researchers have been unable to locate the ancestral homeland of modern Eurasian wolves.

Despite the long distance dispersal ability of wolves, there was a complete population replacement (genetic turnover) from Pleistocene wolves to extant wolves around the Pleistocene-Holocene transition in Siberia (Thalmann et al. 2013; Koblmüller et al. 2016; Ersmark et al. 2016; Loog et al. 2019; Pilot et al. 2019). This suggests that the Pleistocene ancestors of present-day wolves and dogs were isolated during the Last Glacial Maximum (LGM) from the megafaunal form and subsequently colonised Eurasia and North America following the extinction of the megafaunal form.

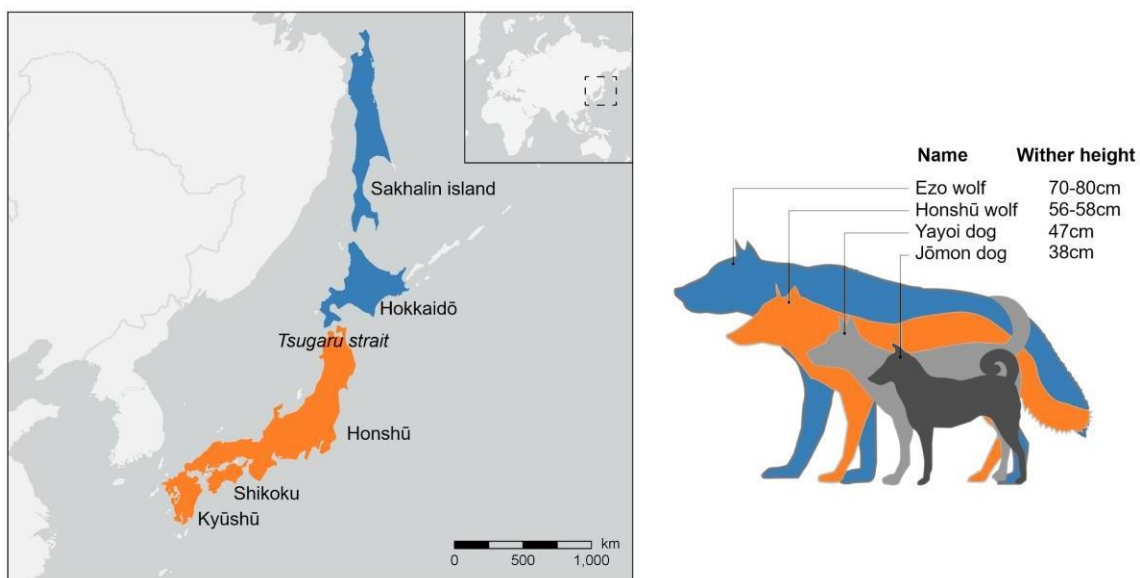


Fig 1: Geographical distribution of the extinct Honshū and Ezo wolf. The Tsugaru strait separates the former habitat of the Ezo wolf - Sakhalin and Hokkaidō - and the former habitat of the Honshū wolf - Honshū, Shikoku, and Kyūshū. Yayoi and Jōmon dogs are now extinct Japanese dog breeds that are ancestral to modern Japanese dogs (Ishiguro 2012). Map created in ArcGIS, wolf and dog outlines modified from (Ishiguro 2012)

The Japanese archipelago is one potential candidate for the LGM refugium of the ancestors of modern wolves and dogs, as land bridges between the Korean peninsula and Japan's largest island, Honshū, formed during the Pleistocene and the beginning of the Holocene (Ohshima 1990). Hokkaidō, the second largest and northernmost island of Japan, was also connected to Beringia during periods of low sea level, which occurred for instance in the Late Pleistocene (Ohshima 1990). Until their extinction at the beginning of the 20th century, Japan was inhabited by two highly phenotypically distinct endemic wolf subspecies: the Japanese or Honshū wolf (*Canis lupus hodophilax*), and the Ezo wolf (*Canis lupus hattai*). While the Honshū wolf could be found on Honshū, Kyūshū, and Shikoku, the habitat of the Ezo wolf was restricted to Hokkaidō and Sakhalin (Fig. 1) (Ishiguro, Inoshima, and Shigehara 2009). The Honshū wolf was among the smallest grey wolf subspecies in the world and appreciated in medieval Japan for killing crop-destroying wildlife (Fritts et al. 2003). A rabies epidemic in the 17th century caused an increase in wolf attacks, setting the human persecution of the Honshū wolf in motion, which culminated in their extinction by 1905 (Walker 2009).

The deep Tsugaru strait between Honshū and Hokkaidō is a major zoogeographical barrier between the two islands, also known as Blakinston's Line (Dobson 1994). As a result, the fauna on Honshū, with its snow macaques (*Macaca fuscata*) and Asian black bears (*Ursus thibetanus*), has similarities to Southeastern Asia, while the fauna on Hokkaidō, which includes the Ussuri Brown bear (*Ursus arctos lasiotus*), resembles the biological diversity in Northeastern Asia. As a consequence of this barrier, there is no evidence for an overlap between the habitats of the Japanese and the Ezo wolf that most likely colonized the Japanese archipelago from the Korean peninsula and Siberia respectively.

The exact phylogenetic placement of both subspecies is speculative, as only the mitochondrial genomes have been sequenced in previous studies (Koblmüller et al. 2016; Matsumura, Inoshima, and Ishiguro 2014). These suggest a basal phylogenetic placement of the Honshū wolf to all modern wolves, and a placement of the Ezo wolf in the North American wolf clade. The mitochondrial genome is however only one marker, and it does not allow the quantification of admixture, which is especially of interest given that both subspecies are potential candidate populations that link Pleistocene wolves and present-day Eurasian wolves.

In order to explore the evolutionary history of the enigmatic Honshū wolf population, we sequenced the nuclear genome of one of the two subspecies, the Honshū wolf (*Canis lupus hodophilax*), to reassess the relationship between Honshū wolves and other wolves and to test the hypothesis that Japan was the LGM refugium for the ancestors of present-day wolves.

Results and Discussion

We generated a 3.7× genome of a Honshū wolf sample, obtained from the Natural History Museum, London, shot in the 1800s in Chichibu District, Kotsuki, Northwest of Tokyo, Japan. First, we investigated the evolutionary relationship between the historic Honshū wolf and other wolves and dogs with a whole-genome admixture analysis using NGSAdmix (Fig. 2). For all predicted ancestry clusters, the Honshū wolf consistently has a highly similar to identical admixture profile to the Pleistocene Siberian wolves in the reference panel. In contrast to all other modern wolf populations, we find the Pleistocene wolf clade contributed substantially to the Honshū wolf genome. Irrespective of the number of ancestry clusters used in the NGSAdmix analysis, none of the present-day wolf populations show a genetic contribution to the Honshū wolf that exceeds the contribution from the Pleistocene wolves. However, there is some evidence of gene flow with dogs.

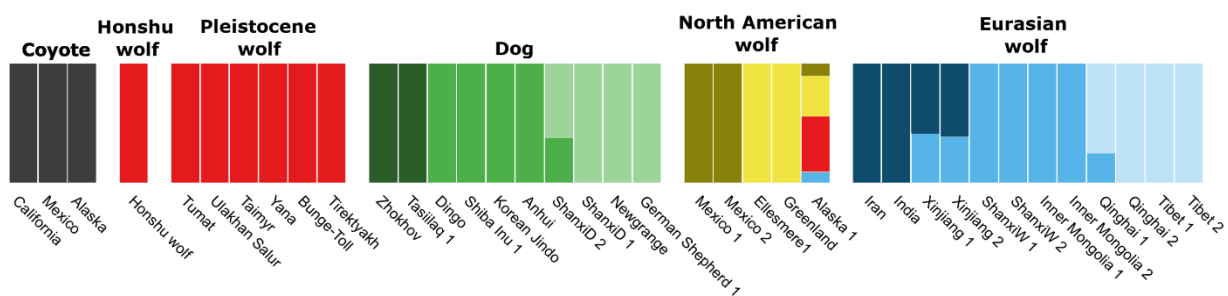


Fig. 2: Admixture plot for $K=10$. Vertical bars represent single individuals, different colours indicate estimated ancestry components. The Honshū wolf forms a cluster with all other Pleistocene wolves (see also Figure S1).

To further explore the admixture landscape between the Honshū wolf and ancient and present-day wolf and dog populations, we subsequently used D-statistics to formally test for gene flow between these groups. The D-statistic, also known as ABBA-BABA test, estimates gene flow between closely related species by comparing the number of shared derived and ancestral alleles between an ingroup (H1, H2, and H3) and the outgroup. In the D-statistic implemented in *angsd*, negative values express gene flow between H1 and H3, positive values indicate gene flow between H2 and H3, and F-values around 0 indicate that there is no excess allele sharing between H1 or H3 and H2 (Zheng and Janke 2018). The D-statistics provide support for excess allele sharing between the Honshū wolf and Greenland dogs, Asian dogs, Pleistocene wolves, and Chinese wolves. We already observed shared genetic ancestry between Pleistocene wolves and the Honshū wolf in the NGSAdmix analysis, so to further investigate wolf and dog populations that might be more genetically similar to the Honshū wolf

than other Pleistocene wolves, we created a scatter plot for the D-statistics with the Honshū wolf and Pleistocene wolf in H3. The results suggest that the Honshū wolf and Pleistocene wolves are equally distantly related to modern Eurasian and North American wolves, with the exception of some Chinese wolves that share more variant sites with the Honshū wolf than they do with any Pleistocene wolf. A potential explanation for this is the substantial admixture between East Asian wolves and dogs (Zhang et al. 2016).

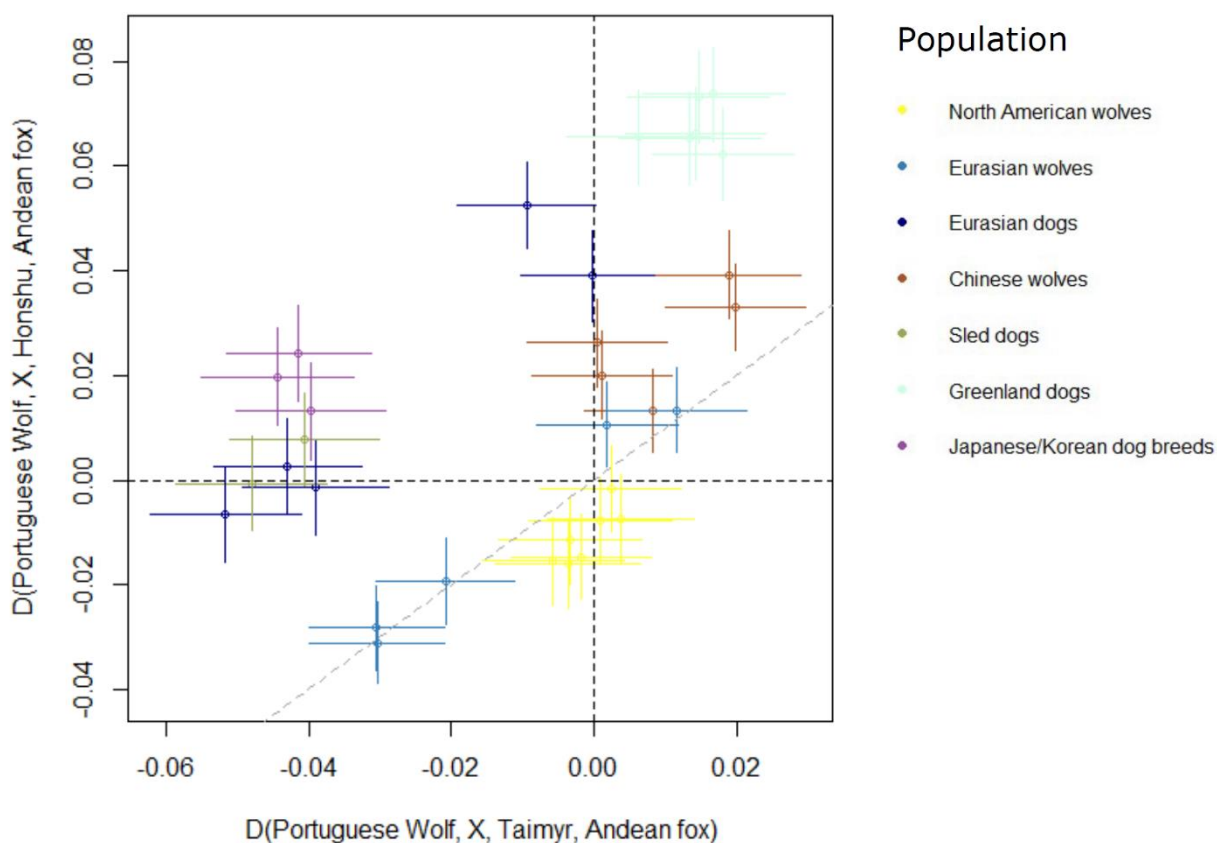


Fig. 3: D-statistics scatter plot for the Portuguese wolf in H1, samples from the reference dataset in H2 (X), and Honshū wolf (y-axis) or the Pleistocene wolf Bunge-Toll (x-axis) in H3. Vertical and horizontal error bars correspond to three standard errors for the tests in the y- and x-axis, respectively. The test involving samples with error bars that intersect the grey dotted line differ insignificantly between the Honshū wolf and Bunge-Toll in H3.

All dog individuals included in this analysis share significantly more alleles with the Honshū wolf than with the Pleistocene wolf, with Japanese dogs, Greenland dogs, and Chinese dogs having the closest genetic affinity with the Honshū wolf. We therefore hypothesize that our Honshū wolf individual was most likely admixed with Japanese dogs, as the excess of shared alleles with the Greenland dogs can be explained by the introgression

from Pleistocene wolves to Arctic dogs (Skoglund et al. 2015), and the Chinese dogs could likewise be shown to have a significant wolf contribution (Zhang et al. 2016).

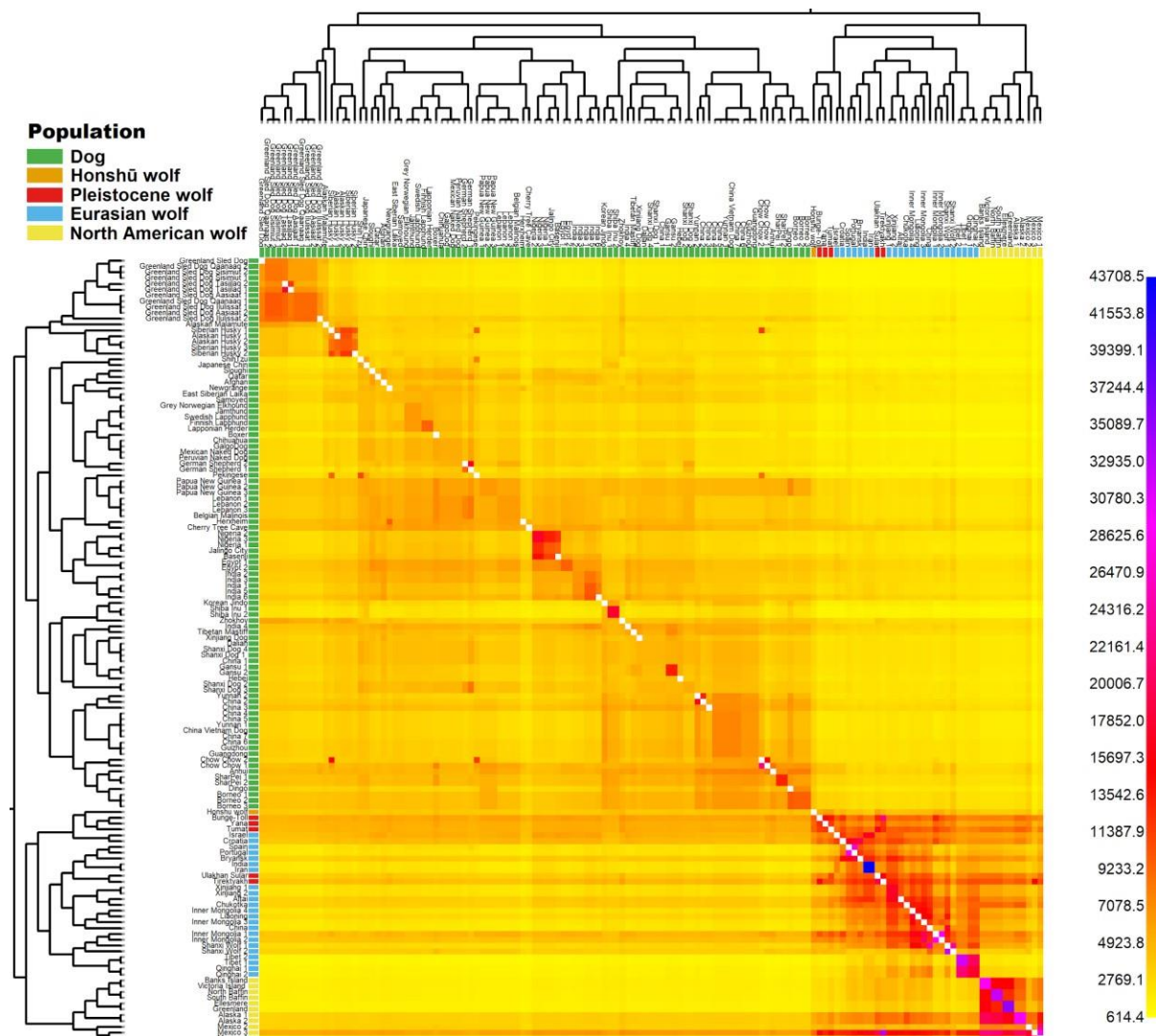


Fig. 4. Heatmap and phylogeny based on shared chromosome segments. Lower left half of the coancestry matrix is based on the unlinked model, while the upper triangle shows values for the linked model. Higher values in the scale express a higher relatedness. Coloured boxes behind sample name indicate individual's population (legend upper left).

In order to more robustly identify population structure among the wolf and dog samples, we used the haplotype-aware clustering tool fineSTRUCTURE (Lawson et al. 2012). Haplotype-aware analyses incorporate the information of which variants were inherited from a single parent and can therefore be highly informative for admixture analyses. In the phylogeny based on a similarity matrix, the Honshū wolf was positioned in the same clade as three other Pleistocene wolves - Tumat, Yana, and Bunge-Toll, further corroborating our earlier findings (Fig. 4). To further verify our findings of genetic affinity of the Honshū wolf to the Pleistocene Siberian wolves, we performed unsupervised dimension reduction on the haplotype data using principal component analysis (PCA). The Honshū wolf clustered together with all Pleistocene

wolves along the first principal component (PC1). Among all the wolves included in the analysis, it placed closest to the dog cluster in the first two principal components (Fig. 5).

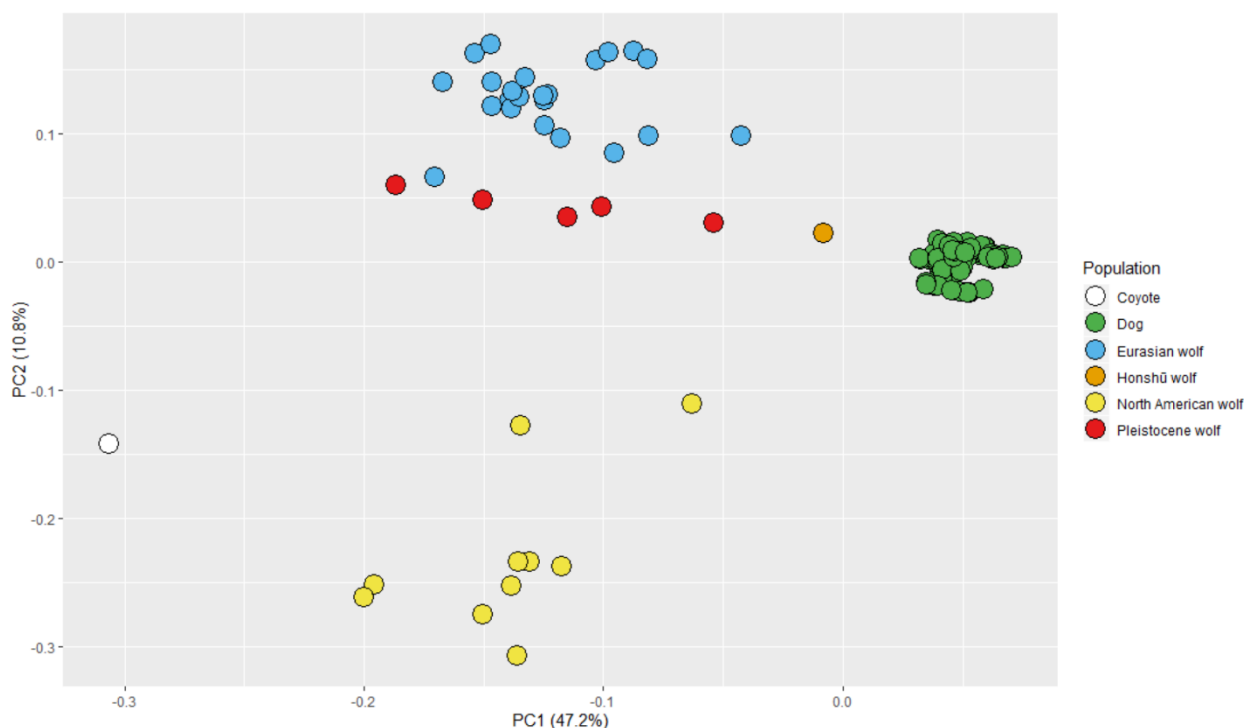


Fig. 5: Principal component analysis based on the fineSTRUCTURE coancestry matrix. The Honshū wolf forms a cluster with all other Pleistocene wolves and displays a close affinity with dogs.

To further examine the population history of the Honshū wolves we tested eight putatively related populations: Japanese dogs, Chinese dogs, Greenland dogs, sled dogs, Honshū wolves, Pleistocene wolves, Eurasian wolves, and North American wolves. The chromosomes of a subset of each of these populations were then painted with the best fitting haplotypes of all remaining individuals. The resulting chromosome paintings could then be used as input for GLOBETROTTER (Hellenthal et al. 2014), which uses the haplotype sharing information to describe and date admixture events involving pre-defined populations (surrogate populations) leading to the population of interest (target population). As GLOBETROTTER requires the data of multiple individuals in the target population to infer admixture dates, we were unable to use the Honshū wolf as a target population. Instead, we chose to run GLOBETROTTER with Japanese dogs as the target population in order to potentially detect gene flow between the Honshū wolves and local dog populations. Using the Chinese dogs, Greenland dogs, sled dogs, Honshū wolves, Pleistocene wolves, Eurasian wolves, and North American wolves as surrogate populations, we estimated that the modern Japanese dog genome can be best described as a mixture of 93% Chinese dog and 7% Honshū wolf. The most likely scenario leading to this admixed population is a single admixture

event, occurring approximately 25 generations ago, between a population that is 9% Chinese dog and 91% Honshū wolf, and a population that is 100% Chinese dog. While these preliminary results indicate that Honshū wolves significantly contributed to modern Japanese dog genomes, it is most likely that the large contribution of Japanese dogs to the Honshū wolf genome confounds the results. Further studies with larger sample sizes of Honshū wolves are therefore needed to positively determine the introgression from Honshū wolves to Japanese dog breeds.

Finally, using the Markov chain Monte Carlo algorithm implemented in SOURCEFIND, we modelled each of the eight populations used in the GLOBETROTTER analysis - Japanese dogs, Chinese dogs, Greenland dogs, sled dogs, Honshū wolves, Pleistocene wolves, Eurasian wolves, and North American wolves - as a mixture of the remaining seven populations, i.e. all the populations except the one being modelled. The chromosome painting of the population of interest was split into 100 subsections, and each subsection was assigned to the best fitting counterpart from one of the other populations.

Using this method we estimated that the Honshū wolf genome can be partitioned into a 52% contribution from Pleistocene wolves, 47% contribution from dogs, and a 1% contribution from present-day Eurasian wolves. Furthermore we detected a 15% contribution from the Honshū wolf to the Japanese dog genome, but found no evidence for haplotype sharing between the Honshū wolf and Chinese dogs. As explained above, the inference of shared ancestry in highly admixed and ill-defined populations such as wolves and dogs is computationally challenging, and the inclusion of more Honshū wolf genomes is necessary to obtain more statistically sound estimates of gene flow between dogs and the Honshū wolf. That being said, a previous mitochondrial study also documented the introgression from the Honshū wolf to some Japanese dogs (Ishiguro, Inoshima, and Shigehara 2009).

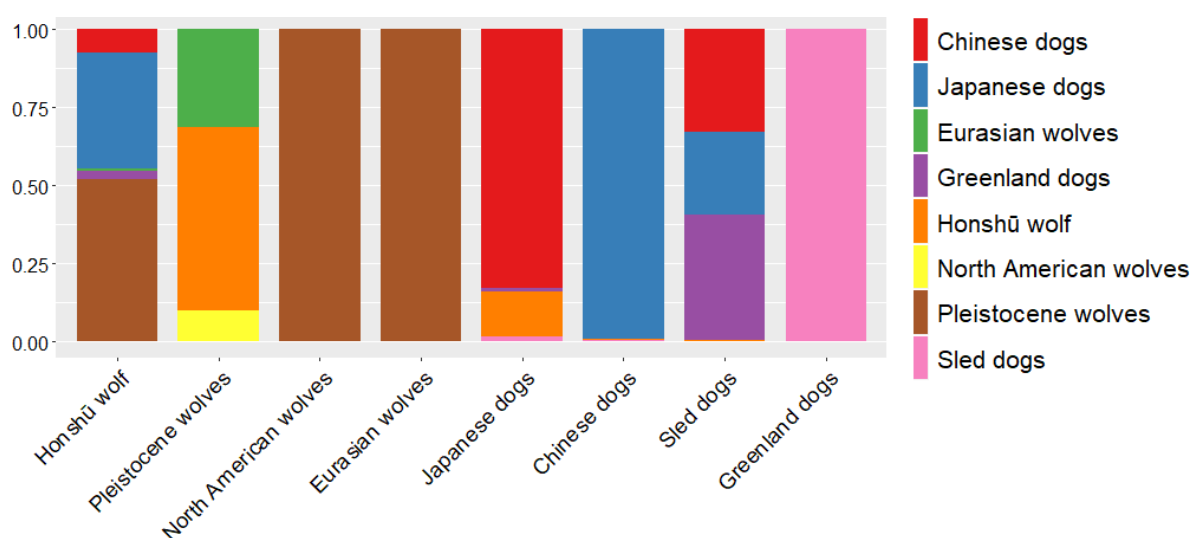


Fig. 6: Estimated proportional contribution (y-axis) from the surrogate populations (right) to the respective target population (x-axis).

Conclusion

The results of our analyses show that the recently extinct Honshū wolf is not in the same phylogenetic clade as present-day Eurasian wolves and that only insubstantial gene flow occurred between present-day wolves and the Honshū wolf. We therefore deem it unlikely that the habitat of Honshū wolves was an LGM refugium for the common ancestor of modern wolves and dogs, as the colonization of Japan by the Honshū wolf is estimated to predate the LGM.

However, we made the unexpected discovery that the Honshū wolf specimen we sampled can be best described as a hybrid between Pleistocene wolves and Japanese dogs. Until now, Pleistocene wolves were thought to have gone extinct around the beginning of the Holocene, but the strong genetic affinity between Honshū wolves (*Canis lupus hodophilax*) and Pleistocene wolves suggests rather that the Japanese archipelago had been a refugium for Pleistocene wolves for thousands of years, where their descendants only went extinct about 100 years ago.

As the Honshū wolf specimen was one of the last of its kind after centuries of human persecution, which resulted in a drastic population decline in the 19th century, it is more than likely that the extent of dog introgression we detected was significantly lower in the Honshū wolf population before they were actively hunted. It is therefore necessary to sequence and analyse the genomes of additional Honshū wolf specimens, especially those that predate the population decline, to obtain a more accurate representation of the genetic makeup of the Honshū wolf. As of now, the high proportion of dog variants in the Honshū wolf specimen hinders our ability to quantify or even reliably detect Honshū wolf introgression into Japanese dog breeds.

Finally, Hokkaidō and Sakhalin island remain potential candidates for LGM refugia, as our analyses only covered the more southern islands Honshū, Shikoku, and Kyūshū. Analysing the yet understudied Ezo wolf genome might therefore be the key to resolve the mystery of the absent ancestors of present-day dogs and wolves.

Material & Methods

Sampling collection

We sampled dry tissue from the inside of the paw of a tanned hide from a Honshū wolf, in the collections of the Natural History Museum - London, the specimen was shot in the 1800s in Chichibu District, Kotsuki, Japan and enter the museum records in 1886.

DNA extraction and Shotgun Sequence Data Generation

2 sup samples of about 500mg historical tanned museum hides were extracted and purified according to (Carøe et al. 2017), in short digested in a proteinase K containing buffer following (Gilbert et al. 2007), treated using phenol and chloroform, DNA was bound to a Minelute columns (Qiagen) using a modified binding apparatus as described in (Dabney et al. 2013) with a buffer following (Allentoft et al. 2015). The column was subsequently washed in PE buffer (Qiagen) and eluted in EB buffer (Qiagen) according to the manufacturer's guidelines. Double stranded DNA libraries were made using the "single-tube" library building protocols BEST (Carøe et al. 2017). The libraries were sequenced on Illumina HiSeq 2500 (Illumina, San Diego, CA, USA), using PE250 bp (modern DNA) and SR50 bp (historic DNA) chemistry.

Alignment

We used the PALEOMIX (v1.2.12) (Schubert et al. 2014) pipeline to process short reads obtained for all ancient and modern samples included in this study. As part of this pipeline, we trimmed the reads and removed adapters using AdapterRemoval2 (v2.2.0) (Schubert, Lindgreen, and Orlando 2016). Paired-end reads overlapping more than 10 base pairs - calculated using the sequences at the 3' end of the first read and the 5' end of the second read of the pair - were merged into a single long read (--collapse option). Adapter trimmed reads that were shorter than 25 bp were discarded. These processed reads were mapped against the wolf reference genome (Gopalakrishnan et al. 2017) and to the dog reference genome (CanFam3.1) using the alignment tool, bwa aln (v0.7.15; aln algorithm) (Li and Durbin 2009). Duplicate reads and reads that mapped to multiple locations in the reference

genome were discarded using picard tools (v1.128, <https://broadinstitute.github.io/picard>). In order to improve the local mapping of reads that span indels, we used GATK (v3.8.0) (McKenna et al. 2010) to perform an indel realignment step on the mapped reads for each of the samples, using no external indel databases. All analyses were performed using the alignments against

the wolf reference genome, unless stated otherwise. The wolf reference genome was used in order to avoid potential reference biases when comparing a mix of ancient and modern samples.

Genotype likelihoods and admixture analysis

We computed the genotype likelihoods at variant sites using ANGSD v0.929-19 (Korneliussen, Albrechtsen, and Nielsen 2014). Sites with base qualities lower than 20 and sequences with a mapping quality lower than 20 were discarded. Only biallelic transversions with data present in at least 30 out of the 37 samples were retained. All sites with minor allele frequencies below 0.01 were excluded. The final dataset consisted of 4,915,722 sites. The genotype likelihoods were then used to estimate admixture proportions between the different samples using NGSAdmix (Skotte, Korneliussen, and Albrechtsen 2013). The admixture proportions were estimated for 2 to 12 clusters. For each cluster, 100 replicates were computed and the admixture proportions of the replicate with the best likelihood were plotted using pong.

D-statistics

To further explore the gene flow between the Honshū wolf, Pleistocene wolves, and extant wolves and dogs, we computed D-statistics using ANGSD. Only biallelic transversions with a coverage higher than 3 and a base quality above 20 were considered. The Andean fox (*Lycalopex culpaeus*) was used as the outgroup for all D-statistics configurations. A weighted block jackknife procedure over 5Mb blocks was used to assess the significance of the tests. We visualized the D-statistics with the Portuguese wolf in H1, modern wolves and dogs in H2, and the Honshū wolf and Pleistocene wolf in H3 in a scatter plot.

Genotype Calling

For each sample, we used the aligned reads to generate a VCF file using GATK's HaplotypeCaller (v3.8.0, (Poplin et al. 2018)). We ran HaplotypeCaller for each sample separately using a minimum base quality score of 20 and a minimum mapping quality score of 30. Further, we ran haplotype caller with the options "--output_mode EMIT_ALL_SITES

-- ERC BP_RESOLUTION" to obtain genotype calls at all sites including sites that were not variable in the sample. Subsequently, we generated a GVCF file using the GenerateGVCFs function in GATK, while still outputting genotypes at all sites. As a final step in the variant calling, we combined the GVCFs from the different samples to get joint variant calls using the SelectVariants function, and at this step, we retained only bi-allelic SNP variants while discarding indels, multi-allelic SNPs and structural variants.

Only biallelic sites with a minimum coverage of 5, a missingness of less than 50% across all individuals, and a minimum quality score of 20 were retained for further analyses. Heterozygous sites with an allele ratio of below 0.33 and above 0.66 were excluded. The final dataset consisted of 30,466,729 sites.

In order to more sensitively detect gene flow and admixture, we simultaneously phased the filtered variant sites of the 136 individuals using Shapelt2 (Delaneau et al. 2013). The recombination maps for each chromosome of the dog genome were downloaded from https://github.com/clcampbell/dog_recombination (Campbell et al. 2016).

fineSTRUCTURE

To obtain an estimate for the global mutation and switch rate we ran ChromopainterV2 on four chromosomes of ten individuals and calculated a weighted average. These estimates were then used in a second ChromopainterV2 analysis to identify shared haplotypes among the samples, whereas each individual can be a donor and recipient of haplotypes. A Markov chain Monte Carlo (MCMC) clustering algorithm was then used with 1 million burn in iterations followed by another 1 million iterations to cluster individuals based on their haplotype sharing. Every 1000th iteration was sampled. The MCMC iteration with the highest observed posterior likelihood was used to infer a phylogeny using a hill-climbing algorithm with 10,000 iterations. A Principal Component Analysis (PCA) was then performed on the resulting linked coancestry matrix.

GLOBETROTTER

Based on the fineSTRUCTURE clustering, we chose seven populations to infer potential admixture events occurring in the ancestral history of Japanese and Korean dog breeds. The previously estimated global mutation and switch rates were used to run Chromopainter v2 with the target and surrogate populations as recipients and the remaining populations as donors.

SOURCEFIND

The Chromopainter v2 results for the previously defined eight wolf and dog populations were furthermore used to predict the admixture proportions for each population, where the remaining six populations act as potential source populations. Unlike GLOBETROTTER, SOURCEFIND (Chacón-Duque et al. 2018) uses a Markov chain Monte Carlo algorithm to estimate the most likely mixture model resulting in the target population. A total of 200,000 MCMC iterations were used, with 50,000 burn-in iterations. Every 5,000th iteration was then sampled, and the mean of each source population contribution was calculated for the resulting 30 sampled iterations.

Author contributions

M.T.P.G, M.-H.S.S. conceived the study. M.-H.S.S. did the ancient DNA lab work. J.N, S.G. performed the bioinformatic analysis. M.-H.S.S. contributed with sample collection. J.R.-M. provided computation expertise. J.N., M.-H.S.S., S.G., N.W., M.T.P.G. supervised the work. J.N, M.-H.S.S., S.G., J.R.-M., N.Y., M.T.P.G. interpreted the results. J.N, M.-H.S.S., S.G., M.T.P.G. wrote the manuscript with input from all authors. All authors read and approved the manuscript.

Acknowledgements

The authors thank the Danish National High-throughput Sequencing Centre for assistance in generating the sequencing data. We thank the London National History Museum and its curators Louise Tomsett and Richard Sabin for assisting with the sample collection. This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 676154 (ArchSci2020).

References

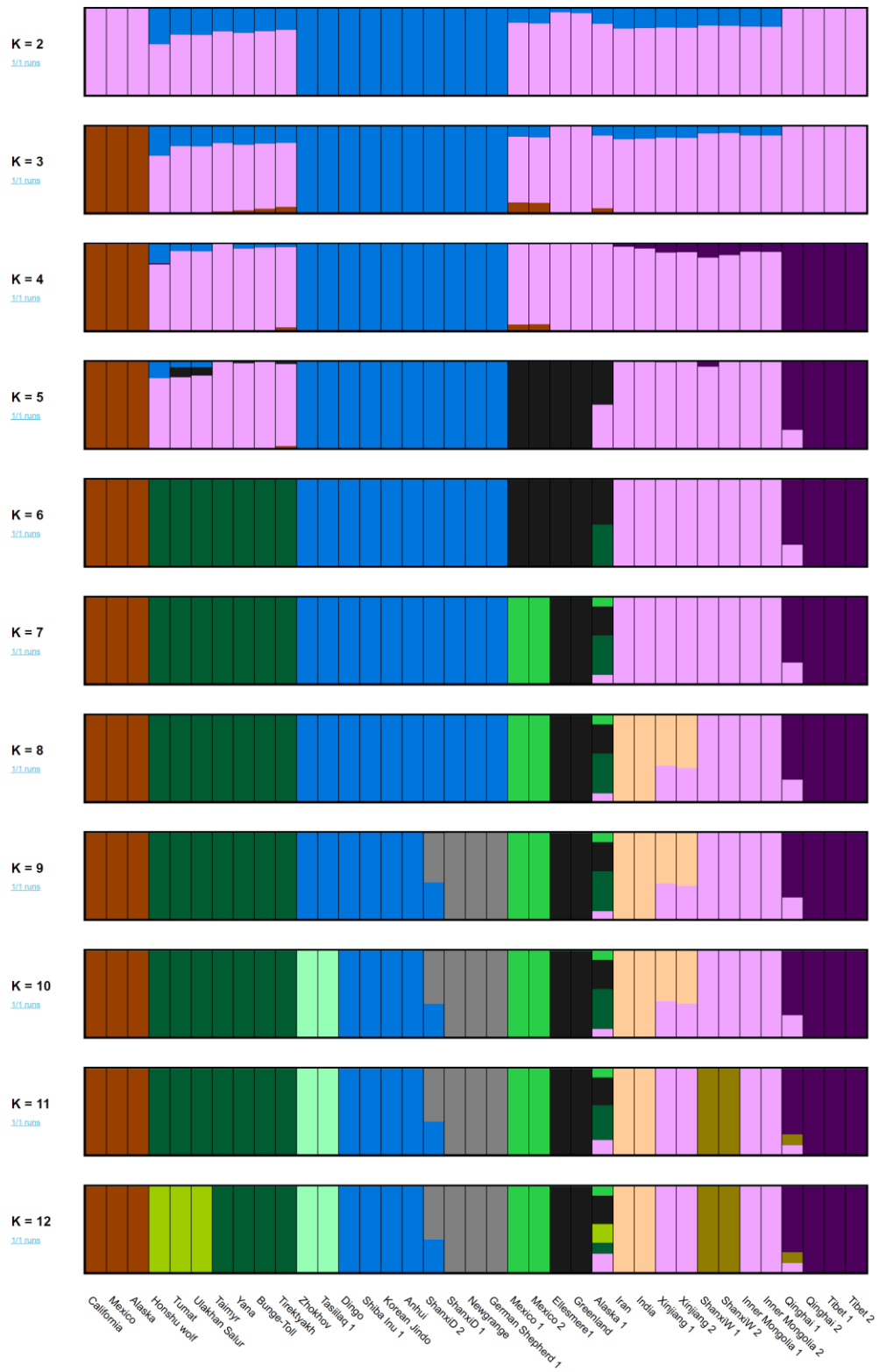
- Allentoft, M. E., Sikora, M., Sjögren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., Vinner, L. and Malaspinas, A.S (2015). Population genomics of bronze age Eurasia. *Nature*, 522(7555), 167-172.
- Campbell, Christopher L., Claude Bhérier, Bernice E. Morrow, Adam R. Boyko, and Adam Auton. 2016. "A Pedigree-Based Map of Recombination in the Domestic Dog Genome." *G3* 6 (11): 3517–24.

- Carøe, Christian, Shyam Gopalakrishnan, Lasse Vinner, Sarah S. T. Mak, Mikkel Holger S. Sinding, José A. Samaniego, Nathan Wales, Thomas Sicheritz-Pontén, and M. Thomas P. Gilbert. 2018. "Single-Tube Library Preparation for Degraded DNA." *Methods in Ecology and Evolution*. <https://doi.org/10.1111/2041-210x.12871>.
- Chacón-Duque, Juan-Camilo, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuña-Alonzo, Rodrigo Barquera, Mirsha Quinto-Sánchez, et al. 2018. "Latin Americans Show Wide-Spread Converso Ancestry and Imprint of Local Native Ancestry on Physical Appearance." *Nature Communications*. <https://doi.org/10.1038/s41467-018-07748-z>.
- Dabney, Jesse, Michael Knapp, Isabelle Glocke, Marie-Theres Gansauge, Antje Weihmann, Birgit Nickel, Cristina Valdiosera, et al. 2013. "Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave Bear Reconstructed from Ultrashort DNA Fragments." *Proceedings of the National Academy of Sciences of the United States of America* 110 (39): 15758–63.
- Delaneau, Olivier, Bryan Howie, Anthony J. Cox, Jean-François Zagury, and Jonathan Marchini. 2013. "Haplotype Estimation Using Sequencing Reads." *The American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2013.09.002>.
- Dobson, M. 1994. "Patterns of Distribution in Japanese Land Mammals." *Mammal Review*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2907.1994.tb00137.x?casa_token=yn9d40AAiREAAAAA:HZYnKhibv3DhqK4VbF5gM_J5eM1_mRhNNhtZsls9D1q1MY51aVmLKig504-ErAdG4cQzb8mNE1l-VuPJ.
- Ersmark, Erik, Cornelya F. C. Klütsch, Yvonne L. Chan, Mikkel-Holger S. Sinding, Steven R. Fain, Natalia A. Illarionova, Mattias Oskarsson, et al. 2016. "From the Past to the Present: Wolf Phylogeography and Demographic History Based on the Mitochondrial Control Region." *Frontiers in Ecology and Evolution* 4 (December): 163.
- Fritts, Steven H., Robert O. Stephenson, Robert D. Hayes, and Luigi Boitani. 2003. "Wolves and Humans," USGS Northern Prairie Wildlife Research Center, . <https://digitalcommons.unl.edu/usgsnpwrc/317/>.
- Gilbert, M. Thomas P., Lynn P. Tomsho, Snjezana Rendulic, Michael Packard, Daniela I. Drautz, Andrei Sher, Alexei Tikhonov, et al. 2007. "Whole-Genome Shotgun Sequencing of Mitochondria from Ancient Hair Shafts." *Science* 317 (5846): 1927–30.
- Gopalakrishnan, Shyam, Jose A. Samaniego Castruita, Mikkel-Holger S. Sinding, Lukas F. K. Kuderna, Jannikke Raikkonen, Bent Petersen, Thomas Sicheritz-Ponten, et al. 2017. "The Wolf Reference Genome Sequence (Canis Lupus Lupus) and Its Implications for Canis Spp. Population Genomics." *BMC Genomics* 18 (June). <https://doi.org/10.1186/s12864-017-3883-3>.
- Hellenthal, Garrett, George B. J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. 2014. "A Genetic Atlas of Human Admixture History." *Science* 343 (6172): 747–51.
- Ishiguro, Naotaka. 2012. "絶滅した日本のオオカミの遺伝的系統." *日本獣医師会雑誌* 65 (3): 225–31.
- Ishiguro, N., Y. Inoshima, and N. Shigehara. 2009. "Mitochondrial DNA Analysis of the Japanese Wolf (Canis Lupus Hodophilax Temminck, 1839) and Comparison with Representative Wolf and Domestic Dog" *Zoological Science*. <http://www.bioone.org/doi/abs/10.2108/zsj.26.765>.
- Koblmüller, Stephan, Carles Vilà, Belen Lorente-Galdos, Marc Dabad, Oscar Ramirez, Tomas Marques-Bonet, Robert K. Wayne, and Jennifer A. Leonard. 2016. "Whole Mitochondrial Genomes Illuminate Ancient Intercontinental Dispersals of Grey Wolves (Canis Lupus)." *Journal of Biogeography* 43 (9): 1728–38.
- Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15 (November): 356.
- Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. "Inference of Population Structure Using Dense Haplotype Data." *PLoS Genetics* 8 (1): e1002453.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.

- Loog, Liisa, Olaf Thalmann, Mikkel-holger S. Sinding, Verena J. Schuenemann, Angela Perri, Mietje Germonpré, Herve Bocherens, et al. 2019. "Ancient DNA Suggests Modern Wolves Trace Their Origin to a Late Pleistocene Expansion from Beringia." *Molecular Ecology*. <https://doi.org/10.1111/mec.15329>.
- Matsumura, Shuichi, Yasuo Inoshima, and Naotaka Ishiguro. 2014. "Reconstructing the Colonization History of Lost Wolf Lineages by the Analysis of the Mitochondrial Genome." *Molecular Phylogenetics and Evolution* 80 (November): 105–12.
- McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data." *Genome Research* 20 (9): 1297–1303.
- Ohshima, Kazuo. 1990. "The History of Straits around the Japanese Islands in the Late-Quaternary." *The Quaternary Research (Daiyonki-Kenkyu)* 29 (3): 193–208.
- Pilot, Małgorzata, Andre E. Moura, Innokentiy M. Okhlopkov, Nikolay V. Mamaev, Abdulaziz N. Alagaili, Osama B. Mohammed, Eduard G. Yavruyan, et al. 2019. "Global Phylogeographic and Admixture Patterns in Grey Wolves and Genetic Legacy of An Ancient Siberian Lineage." *Scientific Reports* 9 (1): 17328.
- Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." *bioRxiv*. <https://doi.org/10.1101/201178>.
- Schubert, Mikkel, Luca Ermini, Clio Der Sarkissian, Hákon Jónsson, Aurélien Ginolhac, Robert Schaefer, Michael D. Martin, et al. 2014. "Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX." *Nature Protocols* 9 (5): 1056–82.
- Schubert, Mikkel, Stinus Lindgreen, and Ludovic Orlando. 2016. "AdapterRemoval v2: Rapid Adapter Trimming, Identification, and Read Merging." *BMC Research Notes* 9 (February): 88.
- Skoglund, Pontus, Erik Ersmark, Eleftheria Palkopoulou, and Love Dalén. 2015. "Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds." *Current Biology: CB* 25 (11): 1515–19.
- Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen. 2013. "Estimating Individual Admixture Proportions from next Generation Sequencing Data." *Genetics* 195 (3): 693–702.
- Thalmann, O., B. Shapiro, P. Cui, V. J. Schuenemann, S. K. Sawyer, D. L. Greenfield, M. B. Germonpré, et al. 2013. "Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs." *Science* 342 (6160): 871–74.
- Walker, Brett L. 2009. *The Lost Wolves of Japan*. University of Washington Press.
- Zhang, Z., J. Xing, C. Vilà, and T. Marques-Bonet. 2016. "Worldwide Patterns of Genomic Variation and Admixture in Gray Wolves." *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*. <http://genome.cshlp.org/content/26/2/163.short>.
- Zheng, Yichen, and Axel Janke. "Gene flow analysis method, the D-statistic, is robust in a wide parameter space." *BMC bioinformatics* 19.1 (2018): 10.

Supplementary Materials

Figure S1: Extended admixture results



Supplementary Tables

Table S1: Whole genome data

| Name | Population | ID | Source |
|------------------|------------------|--------------------|----------------------------|
| Honshu wolf | Honshu wolf | 1886_Honshu | This study |
| Afghan | Dog | AfghanDog | (Shannon et al. 2015) |
| Alaska 1 | Wolf | Alaska1 | (Sinding et al. 2018) |
| Alaskan Husky 1 | Dog | AlaskanHusky_SY001 | (Wiedmer et al. 2016) |
| Alaskan Husky 2 | Dog | AlaskanHusky_SY018 | (Wiedmer et al. 2016) |
| Alaskan Malamute | Dog | AlaskanMalamuteDog | (Decker et al. 2015) |
| Alaska 2 | Wolf | AlaskanWolf | (Cahill et al. 2016) |
| Anhui | Dog | AnhuiDog | (Wang et al. 2016) |
| Banks Island | Wolf | BanksIsland | (Sinding et al. 2018) |
| Basenji | Dog | BasenjiDog | (Freedman et al. 2014) |
| Belgian Malinois | Dog | BM | (Wang et al. 2013) |
| Boxer | Dog | BoxerDog | (Lindblad-Toh et al. 2005) |
| Qinghai 1 | Wolf | CAN11 | (W. Zhang et al. 2014) |
| Qinghai 2 | Wolf | CAN16 | (W. Zhang et al. 2014) |
| Xinjiang 1 | Wolf | CAN24 | (W. Zhang et al. 2014) |
| Xinjiang 2 | Wolf | CAN30 | (W. Zhang et al. 2014) |
| Tibet 2 | Wolf | CAN32 | (W. Zhang et al. 2014) |
| Inner Mongolia 3 | Wolf | CAN6 | (W. Zhang et al. 2014) |
| Inner Mongolia 4 | Wolf | CAN7 | (W. Zhang et al. 2014) |
| Tibet 1 | Wolf | CAN9A | (W. Zhang et al. 2014) |
| Yana | Pleistocene wolf | CGG23 | In preparation |
| Bunge-Toll | Pleistocene wolf | CGG29 | In preparation |
| Tirektyakh | Pleistocene wolf | CGG32 | In preparation |
| Ulakhan Salur | Pleistocene wolf | CGG33 | In preparation |
| Zhokhov | Ancient dog | CGG6 | In preparation |
| Chihuahua | Dog | ChihuahuaDog | (Wang et al. 2016) |
| China Vietnam | Dog | ChinaVietnam4Dog | (Wang et al. 2016) |
| Chow Chow 1 | Dog | ChowChow01 | (Decker et al. 2015) |
| Chow Chow 2 | Dog | ChowChow02 | (Decker et al. 2015) |
| China 1 | Dog | CI1 | (Wang et al. 2013) |
| China 2 | Dog | CI2 | (Wang et al. 2013) |

| | | | |
|---------------------------------|-------------|--------------------------|------------------------------|
| China 3 | Dog | CI3 | (Wang et al. 2013) |
| Cherry Tree Cave | Ancient dog | CTC | (Botigué et al. 2017) |
| Dalian | Dog | DalianDog | (Wang et al. 2016) |
| East Siberian Laika | Dog | EastSiberianLaikaDog | (Wang et al. 2016) |
| Egypt 1 | Dog | EG44 | (Auton et al. 2013) |
| Egypt 2 | Dog | EG49 | (Auton et al. 2013) |
| Ellesmere | Wolf | Ellesmere1 | (Gopalakrishnan et al. 2018) |
| Galgo Español | Dog | GalgoDog | (Wang et al. 2016) |
| Gansu 1 | Dog | Gansu2Dog | (Wang et al. 2016) |
| Gansu 2 | Dog | Gansu3Dog | (Wang et al. 2016) |
| Greenland | Wolf | Greenland_wolf_A1 | (Gopalakrishnan et al. 2018) |
| Greenland Sled dog | Dog | GreenlandDog | (Wang et al. 2016) |
| Greenland Sled Dog Aasiaat 1 | Dog | Greenlandic_dog_Mums | In preparation |
| Greenland Sled Dog Aasiaat 2 | Dog | Greenlandic_dog_Pondus | In preparation |
| Grey Norwegian Elkhound | Dog | GreyNorwegianElkhoundDog | (Wang et al. 2016) |
| German Shepherd 1 | Dog | GS | (Wang et al. 2013) |
| German Shepherd 2 | Dog | GShepDog | (Wang et al. 2016) |
| Guangdong | Dog | GuangdongDog | (Wang et al. 2016) |
| Guizhou | Dog | GuizhouDog | (Wang et al. 2016) |
| Altai | Wolf | GW1 | (Wang et al. 2013) |
| Chukotka | Wolf | GW2 | (Wang et al. 2013) |
| Bryansk | Wolf | GW3 | (Wang et al. 2013) |
| Inner Mongolia 1 | Wolf | GW4 | (Wang et al. 2013) |
| Hebei | Dog | HebeiDog | (Wang et al. 2016) |
| Herxheim | Ancient dog | HXH | (Botigué et al. 2017) |
| Spain | Wolf | IberianWolf | (Z. Zhang et al. 2016) |
| India 1 | Dog | ID125 | (Auton et al. 2013) |
| India 2 | Dog | ID137 | (Auton et al. 2013) |
| India 3 | Dog | ID165 | (Auton et al. 2013) |
| India 4 | Dog | ID168 | (Auton et al. 2013) |
| India 5 | Dog | ID60 | (Auton et al. 2013) |
| India 6 | Dog | ID91 | (Auton et al. 2013) |
| Greenland Sled Dog Illulissat 1 | Dog | Illulissat_GS16 | In preparation |
| Greenland Sled Dog Illulissat 2 | Dog | Illulissat_GS31 | In preparation |
| Borneo 1 | Dog | IN18 | (Auton et al. 2013) |

| | | | |
|------------------------------|--------|------------------------|------------------------------|
| Borneo 2 | Dog | IN23 | (Auton et al. 2013) |
| Borneo 3 | Dog | IN29 | (Auton et al. 2013) |
| India | Wolf | IndiaWolf | (Z. Zhang et al. 2016) |
| Inner Mongolia 2 | Wolf | InnerMongoliaWolf | (Wang et al. 2016) |
| Iran | Wolf | IranWolf | (Z. Zhang et al. 2016) |
| Japanese Chin | Dog | JapaneseChin | (Marchant et al. 2017) |
| Korean Jindo | Dog | KoreanJindo | (Kim et al. 2012) |
| Lapponian Herder | Dog | LapponianHerderDog | (Wang et al. 2016) |
| Lebanon 1 | Dog | LB74 | (Auton et al. 2013) |
| Lebanon 2 | Dog | LB79 | (Auton et al. 2013) |
| Lebanon 3 | Dog | LB85 | (Auton et al. 2013) |
| Mexico 1 | Wolf | Mexican_wolf | (Z. Zhang et al. 2016) |
| Mexican Naked Dog | Dog | MexicanNakedDog | (Wang et al. 2016) |
| Mexico 2 | Wolf | MexicanWolf | (Gopalakrishnan et al. 2018) |
| Mexico | Coyote | Mexico | (Gopalakrishnan et al. 2018) |
| Newgrange | Wolf | Newgrange | (Frantz et al. 2016) |
| North Baffin | Wolf | NorthBaffin | (Sinding et al. 2018) |
| China | Wolf | Novembre_Chinese_Wolf | (Freedman et al. 2014) |
| Croatia | Wolf | Novembre_Croatian_Wolf | (Freedman et al. 2014) |
| Dingo | Dog | Novembre_Dingo | (Freedman et al. 2014) |
| Israel | Wolf | Novembre_Israeli_Wolf | (Freedman et al. 2014) |
| Pekingese | Dog | Pekingese | (Decker et al. 2015) |
| Peruvian Naked Dog | Dog | PeruvianNakedDog | (Wang et al. 2016) |
| Papua New Guinea 1 | Dog | PG115 | (Auton et al. 2013) |
| Papua New Guinea 2 | Dog | PG122 | (Auton et al. 2013) |
| Papua New Guinea 3 | Dog | PG84 | (Auton et al. 2013) |
| Portugal | Wolf | PortugueseWolf | (Z. Zhang et al. 2016) |
| Qatar | Dog | QA27 | (Auton et al. 2013) |
| Greenland Sled Dog Qaanaaq 1 | Dog | Qaanaaq_Q11 | In preparation |
| Greenland Sled Dog Qaanaaq 2 | Dog | Qaanaaq_QSON | In preparation |
| China 4 | Dog | SAMN03168368 | (Wang et al. 2016) |
| China 5 | Dog | SAMN03168369 | (Wang et al. 2016) |
| China 6 | Dog | SAMN03168370 | (Wang et al. 2016) |
| China 7 | Dog | SAMN03168372 | (Wang et al. 2016) |
| Nigeria 1 | Dog | SAMN03168373 | (Wang et al. 2016) |

| | | | |
|-------------------------------|------------------|--------------------|--------------------------|
| Nigeria 2 | Dog | SAMN03168374 | (Wang et al. 2016) |
| Nigeria 3 | Dog | SAMN03168375 | (Wang et al. 2016) |
| Jämthund | Dog | SAMN03168383 | (Wang et al. 2016) |
| Finnish Lapphund | Dog | SAMN03168391 | (Wang et al. 2016) |
| Liaoning | Dog | SAMN03168394 | (Wang et al. 2016) |
| Samoyed | Dog | SamoyedDog | (Wang et al. 2016) |
| Shanxi 1 | Dog | Shanxi1Dog | (Wang et al. 2016) |
| Shanxi 1 | Wolf | Shanxi1Wolf | (Wang et al. 2016) |
| Shanxi 2 | Dog | Shanxi2Dog | (Wang et al. 2016) |
| Shanxi 2 | Wolf | Shanxi2Wolf | (Wang et al. 2016) |
| Shanxi 3 | Dog | Shanxi3Dog | (Wang et al. 2016) |
| Shanxi 4 | Dog | Shanxi4Dog | (Wang et al. 2016) |
| SharPei 1 | Dog | SharPei01 | (Metzger et al. 2017) |
| SharPei 2 | Dog | SharPei02 | (Metzger et al. 2017) |
| Shiba Inu 1 | Dog | ShibaInuFemale | (Kolicheski et al. 2017) |
| Shiba Inu 2 | Dog | ShibaInuMale | (Kolicheski et al. 2017) |
| ShihTzu | Dog | ShihTzu | (Marchant et al. 2017) |
| Siberian Husky 3 | Dog | SiberianHusky_SYXX | (Wiedmer et al. 2016) |
| Siberian Husky 2 | Dog | SiberianHusky01 | (Decker et al. 2015) |
| Siberian Husky 1 | Dog | SiberianHuskyDog | (Wang et al. 2016) |
| Greenland Sled Dog Sisimut 1 | Dog | Sisimiut_02_06_16 | In preparation |
| Greenland Sled Dog Sisimut 2 | Dog | Sisimiut_19_05_16 | In preparation |
| Sloughi | Dog | SloughiDog | (Wang et al. 2016) |
| South Baffin | Wolf | SouthBaffin | (Sinding et al. 2018) |
| Swedish Lapphund | Dog | SwedishLapphundDog | (Wang et al. 2016) |
| Taimyr | Pleistocene Wolf | Taimyr | (Skoglund et al. 2015) |
| Jalingo City | Dog | TarabaDog | (Wang et al. 2016) |
| Greenland Sled Dog Tasiilaq 1 | Dog | Tasiilaq_51602 | In preparation |
| Greenland Sled Dog Tasiilaq 2 | Dog | Tasiilaq_51603 | In preparation |
| Tumat | Wolf | Tumat | In preparation |
| Victoria Island | Wolf | VictoriaIsland | (Sinding et al. 2018) |
| Tibetan Mastiff | Dog | Wang_TM | (Wang et al. 2013) |
| Xinjiang | Dog | XinjiangDog | (Wang et al. 2016) |
| Yunnan 1 | Dog | Yunnan1Dog | (Wang et al. 2016) |
| Yunnan 2 | Dog | Yunnan2Dog | (Wang et al. 2016) |

Table S2: D-statistics showing close affinity between Honshu wolf and Pleistocene wolves

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|---------------|-----------------------------|-------------|------------|------------|--------------|---------------|----------------|------------|
| Alaska 1 | Yana | Honshu wolf | 224522 | 193254 | 0.07484394 | 0.07484394 | 0.003164044 | 23.65452 |
| Alaska 1 | Bunge-Toll | Honshu wolf | 235439 | 204470 | 0.07039865 | 0.07039865 | 0.003551522 | 19.82211 |
| Yana | Bunge-Toll | Honshu wolf | 205660 | 207346 | -0.004082265 | -0.004082265 | 0.002728602 | -1.496101 |
| Alaska 1 | Tirektyakh | Honshu wolf | 229693 | 218776 | 0.02434282 | 0.02434282 | 0.003199007 | 7.609493 |
| Yana | Tirektyakh | Honshu wolf | 202726 | 225328 | -0.05280175 | -0.05280175 | 0.002556573 | -20.65333 |
| Bunge-Toll | Tirektyakh | Honshu wolf | 210213 | 232128 | -0.04954323 | -0.04954323 | 0.002733916 | -18.12171 |
| Alaska 1 | Ulakhan Salur | Honshu wolf | 235662 | 197849 | 0.08722501 | 0.08722501 | 0.003257177 | 26.77932 |
| Yana | Ulakhan Salur | Honshu wolf | 214296 | 210974 | 0.007811508 | 0.007811508 | 0.002762085 | 2.82812 |
| Bunge-Toll | Ulakhan Salur | Honshu wolf | 227844 | 222412 | 0.01206425 | 0.01206425 | 0.002883582 | 4.183771 |
| Tirektyakh | Ulakhan Salur | Honshu wolf | 244677 | 217319 | 0.05921696 | 0.05921696 | 0.00276085 | 21.44882 |
| Alaska 1 | Greenland Sled Dog Aasiat 2 | Honshu wolf | 230968 | 193309 | 0.08876041 | 0.08876041 | 0.003932082 | 22.57339 |
| Yana | Greenland Sled Dog Aasiat 2 | Honshu wolf | 214672 | 210537 | 0.00972463 | 0.00972463 | 0.003844818 | 2.529282 |
| Bunge-Toll | Greenland Sled Dog Aasiat 2 | Honshu wolf | 225693 | 220504 | 0.01162939 | 0.01162939 | 0.003971125 | 2.928488 |
| Tirektyakh | Greenland Sled Dog Aasiat 2 | Honshu wolf | 244145 | 217204 | 0.05839614 | 0.05839614 | 0.003783755 | 15.43338 |
| Ulakhan Salur | Greenland Sled Dog Aasiat 2 | Honshu wolf | 222712 | 223489 | -0.001741368 | -0.001741368 | 0.003917419 | -0.4445192 |
| Alaska 1 | India | Honshu wolf | 210199 | 209507 | 0.001648773 | 0.001648773 | 0.003562212 | 0.4628509 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|--------------------|--------------|-------------|------------|------------|--------------|---------------|----------------|-----------|
| Yana | India | Honshu wolf | 198280 | 229670 | -0.073334969 | -0.073334969 | 0.002952662 | -24.84188 |
| Bunge-Toll | India | Honshu wolf | 209614 | 240840 | -0.06932117 | -0.06932117 | 0.003249302 | -21.33418 |
| Tirektyakh | India | Honshu wolf | 224618 | 234520 | -0.0215665 | -0.0215665 | 0.002846478 | -7.576556 |
| Ulakhan Salur | India | Honshu wolf | 205114 | 242759 | -0.08405285 | -0.08405285 | 0.003218989 | -26.11157 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | India | Honshu wolf | 197930 | 235730 | -0.08716506 | -0.08716506 | 0.004275788 | -20.38573 |
| Alaska 1 | Korean Jindo | Honshu wolf | 230590 | 203684 | 0.06195628 | 0.06195628 | 0.004161139 | 14.88926 |
| Yana | Korean Jindo | Honshu wolf | 215033 | 224565 | -0.02168345 | -0.02168345 | 0.004303214 | -5.038895 |
| Bunge-Toll | Korean Jindo | Honshu wolf | 227287 | 236610 | -0.02009713 | -0.02009713 | 0.004555753 | -4.430826 |
| Tirektyakh | Korean Jindo | Honshu wolf | 245073 | 232296 | 0.02676546 | 0.02676546 | 0.004198333 | 6.375258 |
| Ulakhan Salur | Korean Jindo | Honshu wolf | 223541 | 238344 | -0.0320491 | -0.0320491 | 0.004384888 | -7.30899 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Korean Jindo | Honshu wolf | 168560 | 181027 | -0.03566208 | -0.03566208 | 0.003681872 | -9.685856 |
| India | Korean Jindo | Honshu wolf | 233864 | 206389 | 0.0624073 | 0.0624073 | 0.004697041 | 13.28651 |
| Alaska 1 | Portugal | Honshu wolf | 206452 | 212454 | -0.0143278 | -0.0143278 | 0.003378856 | -4.240428 |
| Yana | Portugal | Honshu wolf | 196824 | 236848 | -0.09229095 | -0.09229095 | 0.003031182 | -30.44718 |
| Bunge-Toll | Portugal | Honshu wolf | 208378 | 250119 | -0.09103876 | -0.09103876 | 0.003460589 | -26.3073 |
| Tirektyakh | Portugal | Honshu wolf | 223207 | 243561 | -0.04360625 | -0.04360625 | 0.003204878 | -13.60621 |
| Ulakhan Salur | Portugal | Honshu wolf | 202457 | 250181 | -0.1054352 | -0.1054352 | 0.003314623 | -31.80912 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Portugal | Honshu wolf | 195505 | 241325 | -0.1048921 | -0.1048921 | 0.00407938 | -25.71275 |
| India | Portugal | Honshu wolf | 204677 | 210566 | -0.01418206 | -0.01418206 | 0.003429236 | -4.135632 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|--------------------|---------------|-------------|------------|------------|--------------|---------------|----------------|-----------|
| Korean Jindo | Portugal | Honshu wolf | 202657 | 237098 | -0.07831861 | -0.07831861 | 0.00441276 | -17.74821 |
| Alaska 1 | Shanxi wolf 1 | Honshu wolf | 214267 | 211428 | 0.006669094 | 0.006669094 | 0.003354533 | 1.988084 |
| Yana | Shanxi wolf 1 | Honshu wolf | 205329 | 236525 | -0.07060251 | -0.07060251 | 0.003301337 | -21.38603 |
| Bunge-Toll | Shanxi wolf 1 | Honshu wolf | 216561 | 248456 | -0.06858889 | -0.06858889 | 0.003573685 | -19.19277 |
| Tirektyakh | Shanxi wolf 1 | Honshu wolf | 232119 | 242585 | -0.02204742 | -0.02204742 | 0.003220921 | -6.845068 |
| Ulakhan Salur | Shanxi wolf 1 | Honshu wolf | 212104 | 250270 | -0.08254357 | -0.08254357 | 0.003375643 | -24.4527 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Shanxi wolf 1 | Honshu wolf | 201591 | 238462 | -0.08378763 | -0.08378763 | 0.003761901 | -22.27269 |
| India | Shanxi wolf 1 | Honshu wolf | 218313 | 215363 | 0.006802313 | 0.006802313 | 0.003585368 | 1.897243 |
| Korean Jindo | Shanxi wolf 1 | Honshu wolf | 209216 | 234682 | -0.05736904 | -0.05736904 | 0.003878406 | -14.79191 |
| Portugal | Shanxi wolf 1 | Honshu wolf | 220381 | 211291 | 0.02105765 | 0.02105765 | 0.003161539 | 6.660571 |
| Alaska 1 | Shiba Inu 1 | Honshu wolf | 234813 | 203153 | 0.07222872 | 0.07222872 | 0.004082016 | 17.70907 |
| Yana | Shiba Inu 1 | Honshu wolf | 220531 | 225429 | -0.01098305 | -0.01098305 | 0.004222731 | -2.600935 |
| Bunge-Toll | Shiba Inu 1 | Honshu wolf | 232606 | 237097 | -0.009561361 | -0.009561361 | 0.004342929 | -2.201593 |
| Tirektyakh | Shiba Inu 1 | Honshu wolf | 250646 | 233111 | 0.03624754 | 0.03624754 | 0.004148715 | 8.737052 |
| Ulakhan Salur | Shiba Inu 1 | Honshu wolf | 228692 | 238606 | -0.02121558 | -0.02121558 | 0.004378628 | -4.845259 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Shiba Inu 1 | Honshu wolf | 173682 | 181839 | -0.02294379 | -0.02294379 | 0.003691399 | -6.215473 |
| India | Shiba Inu 1 | Honshu wolf | 239003 | 206565 | 0.07280146 | 0.07280146 | 0.004519237 | 16.10924 |
| Korean Jindo | Shiba Inu 1 | Honshu wolf | 176149 | 171720 | 0.0127318 | 0.0127318 | 0.003698132 | 3.442766 |
| Portugal | Shiba Inu 1 | Honshu wolf | 241575 | 202025 | 0.0891569 | 0.0891569 | 0.004322157 | 20.62787 |
| Shanxi wolf 1 | Shiba Inu 1 | Honshu wolf | 240011 | 209683 | 0.06744142 | 0.06744142 | 0.003943726 | 17.10094 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|---------------------------------|----------------------------------|-------------|------------|------------|--------------|---------------|----------------|-----------|
| Alaska 1 | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 231462 | 194984 | 0.08553955 | 0.08553955 | 0.003925899 | 21.78853 |
| Yana | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 215206 | 212642 | 0.005992782 | 0.005992782 | 0.003969816 | 1.509587 |
| Bunge-Toll | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 226752 | 223282 | 0.007710529 | 0.007710529 | 0.004044754 | 1.906304 |
| Tirektyakh | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 244422 | 219113 | 0.05459998 | 0.05459998 | 0.00388137 | 14.06719 |
| Ulakhan Salur | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 222954 | 225339 | -0.005320181 | -0.005320181 | 0.003881045 | -1.370811 |
| Greenland Sled Dog Aasiaat 2 | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 121249 | 122966 | -0.00703069 | -0.00703069 | 0.003631046 | -1.936271 |
| India | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 236610 | 200207 | 0.08333696 | 0.08333696 | 0.004337413 | 19.21352 |
| Korean Jindo | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 180742 | 169971 | 0.03071172 | 0.03071172 | 0.00363998 | 8.437333 |
| Portugal | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 241954 | 197613 | 0.1008743 | 0.1008743 | 0.004178953 | 24.13865 |
| Shanxi wolf 1 | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 238781 | 203352 | 0.080132 | 0.080132 | 0.003794208 | 21.11956 |
| Shiba Inu 1 | Greenland Sled Dog Tasiliag 1 | Honshu wolf | 180665 | 174736 | 0.01668256 | 0.01668256 | 0.003846212 | 4.337401 |
| Alaska 1 | Tumat | Honshu wolf | 218700 | 192933 | 0.06259702 | 0.06259702 | 0.003226135 | 19.4031 |
| Yana | Tumat | Honshu wolf | 201110 | 206888 | -0.01416183 | -0.01416183 | 0.002818714 | -5.024218 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|--------------------|-----------------|-------------|------------|------------|--------------|---------------|----------------|-----------|
| Bunge-Toll | Tumat | Honshu wolf | 212330 | 217539 | -0.01211765 | -0.01211765 | 0.00288452 | -4.200923 |
| Tirektyakh | Tumat | Honshu wolf | 229654 | 214012 | 0.03525625 | 0.03525625 | 0.002926787 | 12.04606 |
| Ulakhan Salur | Tumat | Honshu wolf | 205727 | 217407 | -0.02760355 | -0.02760355 | 0.002743453 | -10.06161 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Tumat | Honshu wolf | 205457 | 216978 | -0.02727283 | -0.02727283 | 0.003896104 | -7.000028 |
| India | Tumat | Honshu wolf | 225797 | 201438 | 0.05701546 | 0.05701546 | 0.003342252 | 17.059 |
| Korean Jindo | Tumat | Honshu wolf | 217007 | 215591 | 0.003273247 | 0.003273247 | 0.004385567 | 0.746368 |
| Portugal | Tumat | Honshu wolf | 229589 | 196056 | 0.07878161 | 0.07878161 | 0.003289485 | 23.94953 |
| Shanxi wolf 1 | Tumat | Honshu wolf | 229535 | 205022 | 0.05640917 | 0.05640917 | 0.003393063 | 16.62485 |
| Shiba Inu 1 | Tumat | Honshu wolf | 218300 | 221764 | -0.007871582 | -0.007871582 | 0.004223933 | -1.863567 |
| Greenland Sled Dog | | | | | | | | |
| Tasitlag 1 | Tumat | Honshu wolf | 206922 | 216730 | -0.02315108 | -0.02315108 | 0.003905781 | -5.927387 |
| Alaska 1 | Victoria Island | Honshu wolf | 187082 | 193276 | -0.01628466 | -0.01628466 | 0.003518714 | -4.628014 |
| Yana | Victoria Island | Honshu wolf | 198132 | 238229 | -0.09188951 | -0.09188951 | 0.003577421 | -25.68596 |
| Bunge-Toll | Victoria Island | Honshu wolf | 209721 | 250705 | -0.08901322 | -0.08901322 | 0.004004319 | -22.2293 |
| Tirektyakh | Victoria Island | Honshu wolf | 224633 | 244473 | -0.04229321 | -0.04229321 | 0.003637759 | -11.62617 |
| Ulakhan Salur | Victoria Island | Honshu wolf | 203544 | 250928 | -0.1042616 | -0.1042616 | 0.003556046 | -29.31954 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Victoria Island | Honshu wolf | 198093 | 244289 | -0.1044256 | -0.1044256 | 0.004353495 | -23.98661 |
| India | Victoria Island | Honshu wolf | 215687 | 223129 | -0.01695927 | -0.01695927 | 0.004029263 | -4.209026 |
| Korean Jindo | Victoria Island | Honshu wolf | 208025 | 242566 | -0.0766571 | -0.0766571 | 0.00457967 | -16.73856 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|--------------------|-----------------|-------------|------------|------------|-------------|---------------|----------------|----------------|
| Portugal | Victoria Island | Honshu wolf | 218065 | 218046 | 4.36E+01 | 4.36E+01 | 0.003869642 | 0.0112586 4 |
| Shanxi wolf 1 | Victoria Island | Honshu wolf | 217339 | 226579 | -0.02081465 | -0.02081465 | 0.003637518 | -5.722213 |
| Shiba Inu 1 | Victoria Island | Honshu wolf | 208545 | 247474 | -0.08536706 | -0.08536706 | 0.004469144 | -19.10144 |
| Greenland Sled Dog | | | | | | | | |
| Tasiliq 1 | Victoria Island | Honshu wolf | 199605 | 244275 | -0.1006353 | -0.1006353 | 0.004382886 | -22.96097 |
| Tumai | Victoria Island | Honshu wolf | 197460 | 231397 | -0.0791336 | -0.0791336 | 0.003812922 | -20.75406 |
| Alaska 1 | Altai | Honshu wolf | 206188 | 191701 | 0.03640965 | 0.03640965 | 0.003467085 | 10.50152 |
| Yana | Altai | Honshu wolf | 196058 | 209052 | -0.03207524 | -0.03207524 | 0.002989549 | -10.72912 |
| Bunge-Toll | Altai | Honshu wolf | 208152 | 219702 | -0.02699519 | -0.02699519 | 0.002952138 | -9.144284 |
| Tirektyakh | Altai | Honshu wolf | 221443 | 213056 | 0.01930269 | 0.01930269 | 0.002803819 | 6.884429 |
| Ulakhan Salur | Altai | Honshu wolf | 202228 | 221171 | -0.0447403 | -0.0447403 | 0.003181714 | -14.0617 |
| Greenland Sled Dog | | | | | | | | |
| Aasiaat 2 | Altai | Honshu wolf | 193690 | 215701 | -0.05376523 | -0.05376523 | 0.004018484 | -13.37948 |
| India | Altai | Honshu wolf | 204684 | 190194 | 0.03669488 | 0.03669488 | 0.003223666 | 11.38297 |
| Korean Jindo | Altai | Honshu wolf | 202075 | 215290 | -0.03166293 | -0.03166293 | 0.004539295 | -6.975297 |
| Portugal | Altai | Honshu wolf | 209523 | 191156 | 0.04583969 | 0.04583969 | 0.003310978 | 13.84476 |
| Shanxi wolf 1 | Altai | Honshu wolf | 208300 | 198450 | 0.02421635 | 0.02421635 | 0.003285305 | 7.371112 |
| Shiba Inu 1 | Altai | Honshu wolf | 202210 | 220405 | -0.04305337 | -0.04305337 | 0.0043613 | -9.871682 |
| Greenland Sled Dog | | | | | | | | |
| Tasiliq 1 | Altai | Honshu wolf | 195244 | 216311 | -0.05118878 | -0.05118878 | 0.004068149 | -12.58282 |
| Tumai | Altai | Honshu wolf | 198328 | 206563 | -0.02033881 | -0.02033881 | 0.00319397 | -6.367877 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-----------------|--------------------|-------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Victoria Island | Altai | Honshu wolf | 217638 | 196761 | 0.05037898 | 0.05037898 | 0.004004775 | 12.57973 |
| Honshu wolf | Yana | Alaska 1 | 212628 | 193254 | 0.04773308 | 0.04773308 | 0.00335295 | 14.23615 |
| Honshu wolf | Bunge-Toll | Alaska 1 | 217697 | 204470 | 0.0313312 | 0.0313312 | 0.003198166 | 9.796615 |
| Honshu wolf | Tirektyakh | Alaska 1 | 222630 | 218776 | 0.008731191 | 0.008731191 | 0.003006605 | 2.904003 |
| Honshu wolf | Ulakhan Salur | Alaska 1 | 233408 | 197849 | 0.08245431 | 0.08245431 | 0.003362883 | 24.51894 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiat 2 | Alaska 1 | 223733 | 193309 | 0.07295188 | 0.07295188 | 0.003435899 | 21.23226 |
| Honshu wolf | India | Alaska 1 | 233299 | 209507 | 0.05373008 | 0.05373008 | 0.003506273 | 15.32398 |
| Honshu wolf | Korean Jindo | Alaska 1 | 225313 | 203684 | 0.0504176 | 0.0504176 | 0.003401395 | 14.82263 |
| Honshu wolf | Portugal | Alaska 1 | 239718 | 212454 | 0.06029564 | 0.06029564 | 0.003365967 | 17.91332 |
| Honshu wolf | Shanxi wolf 1 | Alaska 1 | 238138 | 211428 | 0.05941286 | 0.05941286 | 0.003297608 | 18.01695 |
| Honshu wolf | Shiba Inu 1 | Alaska 1 | 224838 | 203153 | 0.05066695 | 0.05066695 | 0.003309813 | 15.3081 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliq 1 | Alaska 1 | 225301 | 194984 | 0.07213439 | 0.07213439 | 0.003625615 | 19.89577 |
| Honshu wolf | Tumat | Alaska 1 | 228438 | 192933 | 0.08426066 | 0.08426066 | 0.003356221 | 25.10581 |
| Honshu wolf | Victoria Island | Alaska 1 | 286277 | 193276 | 0.1939327 | 0.1939327 | 0.00454457 | 42.67349 |
| Honshu wolf | Altai | Alaska 1 | 230227 | 191701 | 0.09130942 | 0.09130942 | 0.003506833 | 26.03757 |
| Honshu wolf | Alaska 1 | Yana | 212628 | 224522 | -0.02720805 | -0.02720805 | 0.003221439 | -8.445932 |
| Honshu wolf | Bunge-Toll | Yana | 232212 | 207346 | 0.05657046 | 0.05657046 | 0.003178726 | 17.79658 |
| Honshu wolf | Tirektyakh | Yana | 233840 | 225328 | 0.01853788 | 0.01853788 | 0.002993998 | 6.191717 |
| Honshu wolf | Ulakhan Salur | Yana | 231806 | 210974 | 0.0470482 | 0.0470482 | 0.003091459 | 15.21877 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| | Aasiat 2 | Yana | 213162 | 210537 | 0.006195436 | 0.006195436 | 0.003235939 | 1.914571 |
| Honshu wolf | India | Yana | 218971 | 229670 | -0.02384758 | -0.02384758 | 0.002935898 | -8.122752 |
| Honshu wolf | Korean Jindo | Yana | 212889 | 224565 | -0.02669081 | -0.02669081 | 0.003290369 | -8.111798 |
| Honshu wolf | Portugal | Yana | 220066 | 236848 | -0.03672901 | -0.03672901 | 0.003145587 | -11.67636 |
| Honshu wolf | Shanxi wolf 1 | Yana | 218221 | 236525 | -0.04025104 | -0.04025104 | 0.003147245 | -12.78929 |
| Honshu wolf | Shiba Inu 1 | Yana | 212382 | 225429 | -0.02980053 | -0.02980053 | 0.003241565 | -9.193253 |
| | Greenland Sled Dog | | | | | | | |
| | Tasiliq 1 | Yana | 213984 | 212642 | 0.003145612 | 0.003145612 | 0.003404731 | 0.9238945 |
| Honshu wolf | Tunat | Yana | 221462 | 206888 | 0.03402358 | 0.03402358 | 0.003115045 | 10.92234 |
| Honshu wolf | Victoria Island | Yana | 219607 | 238229 | -0.04067395 | -0.04067395 | 0.003503021 | -11.61111 |
| Honshu wolf | Altai | Yana | 214513 | 209052 | 0.01289294 | 0.01289294 | 0.002993469 | 4.307025 |
| Honshu wolf | Alaska 1 | Bunge-Toll | 217697 | 235439 | -0.03915381 | -0.03915381 | 0.002997578 | -13.06182 |
| Honshu wolf | Yana | Bunge-Toll | 232212 | 205660 | 0.06063873 | 0.06063873 | 0.003057412 | 19.83335 |
| Honshu wolf | Tirektyakh | Bunge-Toll | 249122 | 232128 | 0.03531221 | 0.03531221 | 0.002874941 | 12.28276 |
| Honshu wolf | Ulakhan Salur | Bunge-Toll | 236877 | 222412 | 0.03149433 | 0.03149433 | 0.002824186 | 11.15165 |
| | Greenland Sled Dog | | | | | | | |
| | Aasiat 2 | Bunge-Toll | 219902 | 220504 | -0.001366921 | -0.001366921 | 0.003134666 | -0.4360658 |
| Honshu wolf | India | Bunge-Toll | 225651 | 240840 | -0.03256011 | -0.03256011 | 0.002897204 | -11.23846 |
| Honshu wolf | Korean Jindo | Bunge-Toll | 219542 | 236610 | -0.03741735 | -0.03741735 | 0.003398379 | -11.01035 |
| Honshu wolf | Portugal | Bunge-Toll | 225334 | 250119 | -0.05212923 | -0.05212923 | 0.003136674 | -16.61927 |
| Honshu wolf | Shanxi wolf 1 | Bunge-Toll | 223804 | 248456 | -0.05220006 | -0.05220006 | 0.003071758 | -16.99355 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Shiba Inu 1 | Bunge-Toll | 218470 | 237097 | -0.04088751 | -0.04088751 | 0.003405439 | -12.00653 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Bunge-Toll | 220659 | 223282 | -0.005908443 | -0.005908443 | 0.003205376 | -1.843292 |
| Honshu wolf | Tumat | Bunge-Toll | 226912 | 217539 | 0.02108894 | 0.02108894 | 0.002844923 | 7.412833 |
| Honshu wolf | Victoria Island | Bunge-Toll | 225620 | 250705 | -0.05266362 | -0.05266362 | 0.003424884 | -15.37676 |
| Honshu wolf | Altai | Bunge-Toll | 219502 | 219702 | 0.0004553693 | -0.0004553693 | 0.002860453 | -0.1591948 |
| Honshu wolf | Alaska 1 | Tirektyakh | 222630 | 229693 | -0.01561495 | -0.01561495 | 0.002916736 | -5.353569 |
| Honshu wolf | Yana | Tirektyakh | 233840 | 202726 | 0.07126987 | 0.07126987 | 0.00298035 | 23.91326 |
| Honshu wolf | Bunge-Toll | Tirektyakh | 249122 | 210213 | 0.08470724 | 0.08470724 | 0.003076695 | 27.5319 |
| Honshu wolf | Ulakhan Salur | Tirektyakh | 244001 | 217319 | 0.05783838 | 0.05783838 | 0.002704219 | 21.3882 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiat 2 | Tirektyakh | 222114 | 217204 | 0.01117641 | 0.01117641 | 0.003014321 | 3.707771 |
| Honshu wolf | India | Tirektyakh | 233533 | 234520 | -0.002108736 | -0.002108736 | 0.002855537 | -0.7384726 |
| Honshu wolf | Korean Jindo | Tirektyakh | 221855 | 232296 | -0.02299015 | -0.02299015 | 0.003168312 | -7.256276 |
| Honshu wolf | Portugal | Tirektyakh | 232008 | 243561 | -0.024293 | -0.024293 | 0.002826444 | -8.5949 |
| Honshu wolf | Shanxi wolf 1 | Tirektyakh | 229106 | 242585 | -0.02857591 | -0.02857591 | 0.00291826 | -9.792107 |
| Honshu wolf | Shiba Inu 1 | Tirektyakh | 221942 | 233111 | -0.02454439 | -0.02454439 | 0.003057549 | -8.027473 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Tirektyakh | 223226 | 219113 | 0.009298298 | 0.009298298 | 0.003065855 | 3.032856 |
| Honshu wolf | Tumat | Tirektyakh | 229310 | 214012 | 0.03450765 | 0.03450765 | 0.003013878 | 11.44958 |
| Honshu wolf | Victoria Island | Tirektyakh | 231471 | 244473 | -0.02731834 | -0.02731834 | 0.003147697 | -8.678835 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|---------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Altai | Tirekiyakh | 225106 | 213056 | 0.02750124 | 0.02750124 | 0.002959363 | 9.29296 |
| Honshu wolf | Alaska 1 | Ulakhan Salur | 233408 | 235662 | -0.004805253 | -0.004805253 | 0.003395548 | -1.415163 |
| Honshu wolf | Yana | Ulakhan Salur | 231806 | 214296 | 0.03925111 | 0.03925111 | 0.003305898 | 11.87306 |
| Honshu wolf | Bunge-Toll | Ulakhan Salur | 236877 | 227844 | 0.01943747 | 0.01943747 | 0.003306863 | 5.877918 |
| Honshu wolf | Tirekiyakh | Ulakhan Salur | 244001 | 244677 | -0.001383324 | -0.001383324 | 0.002982159 | -0.4638666 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiaat 2 | Ulakhan Salur | 231506 | 223489 | 0.01761997 | 0.01761997 | 0.00327319 | 5.38312 |
| Honshu wolf | India | Ulakhan Salur | 239620 | 242759 | -0.006507331 | -0.006507331 | 0.003266024 | -1.992432 |
| Honshu wolf | Korean Jindo | Ulakhan Salur | 230940 | 238344 | -0.01577723 | -0.01577723 | 0.003394323 | -4.648122 |
| Honshu wolf | Portugal | Ulakhan Salur | 241019 | 250181 | -0.01865228 | -0.01865228 | 0.003141164 | -5.938016 |
| Honshu wolf | Shanxi wolf 1 | Ulakhan Salur | 237005 | 250270 | -0.02722282 | -0.02722282 | 0.003192598 | -8.526855 |
| Honshu wolf | Shiba Inu 1 | Ulakhan Salur | 231360 | 238606 | -0.01541814 | -0.01541814 | 0.003199242 | -4.819309 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Ulakhan Salur | 232218 | 225339 | 0.01503419 | 0.01503419 | 0.003331896 | 4.512204 |
| Honshu wolf | Tumat | Ulakhan Salur | 244968 | 217407 | 0.05960746 | 0.05960746 | 0.003359873 | 17.74099 |
| Honshu wolf | Victoria Island | Ulakhan Salur | 241098 | 250928 | -0.01997862 | -0.01997862 | 0.003572975 | -5.591593 |
| Honshu wolf | Altai | Ulakhan Salur | 233993 | 221171 | 0.02817007 | 0.02817007 | 0.003254382 | 8.656041 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Alaska 1 | Aasiaat 2 | 223733 | 230968 | -0.01591156 | -0.01591156 | 0.004055863 | -3.9231 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Yana | Aasiaat 2 | 213162 | 214672 | -0.003529406 | -0.003529406 | 0.004324994 | -0.8160488 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|----------------------------------|---------------------------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Bunge-Toll | Greenland Sled Dog Aasiaat 2 | 219902 | 225693 | -0.01299611 | -0.01299611 | 0.004202213 | -3.092682 |
| Honshu wolf | Tirektyakh | Greenland Sled Dog Aasiaat 2 | 222114 | 244145 | -0.04725056 | -0.04725056 | 0.003905629 | -12.09807 |
| Honshu wolf | Ulakhan Salur | Greenland Sled Dog Aasiaat 2 | 231506 | 222712 | 0.01936075 | 0.01936075 | 0.004097295 | 4.725251 |
| Honshu wolf | India | Greenland Sled Dog Aasiaat 2 | 230644 | 235730 | -0.01090541 | -0.01090541 | 0.004244882 | -2.569073 |
| Honshu wolf | Korean Jindo | Greenland Sled Dog Aasiaat 2 | 314536 | 181027 | 0.2694087 | 0.2694087 | 0.005019386 | 53.67364 |
| Honshu wolf | Portugal | Greenland Sled Dog Aasiaat 2 | 233827 | 241325 | -0.01578021 | -0.01578021 | 0.004115229 | -3.834589 |
| Honshu wolf | Shanxi wolf 1 | Greenland Sled Dog Aasiaat 2 | 235718 | 238462 | -0.005786832 | -0.005786832 | 0.004175212 | -1.385997 |
| Honshu wolf | Shiba Inu 1 | Greenland Sled Dog Aasiaat 2 | 312843 | 181839 | 0.2648247 | 0.2648247 | 0.00496049 | 53.38679 |
| Honshu wolf | Greenland Sled Dog Tasiliag 1 | Greenland Sled Dog Aasiaat 2 | 414009 | 122966 | 0.5420047 | 0.5420047 | 0.006051975 | 89.55833 |
| Honshu wolf | Tumat | Greenland Sled Dog Aasiaat 2 | 227082 | 216978 | 0.02275368 | 0.02275368 | 0.00419834 | 5.419685 |
| Honshu wolf | Victoria Island | Greenland Sled Dog Aasiaat 2 | 233179 | 244289 | -0.02326858 | -0.02326858 | 0.004387112 | -5.303848 |
| Honshu wolf | Altai | Greenland Sled Dog Aasiaat 2 | 226523 | 215701 | 0.02447176 | 0.02447176 | 0.004351177 | 5.624171 |
| Honshu wolf | Alaska 1 | India | 233299 | 210199 | 0.05208592 | 0.05208592 | 0.003507296 | 14.85073 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|---------------------------------|--------------|------------|------------|--------------|---------------|----------------|-----------|
| Honshu wolf | Yana | India | 218971 | 198280 | 0.04958886 | 0.04958886 | 0.003186521 | 15.56207 |
| Honshu wolf | Bunge-Toll | India | 225651 | 209614 | 0.03684422 | 0.03684422 | 0.003191791 | 11.54343 |
| Honshu wolf | Tirektyakh | India | 233533 | 224618 | 0.01945865 | 0.01945865 | 0.003066253 | 6.346068 |
| Honshu wolf | Ulakhan Salur | India | 239620 | 205114 | 0.07758795 | 0.07758795 | 0.003254165 | 23.84266 |
| Honshu wolf | Greenland Sled Dog Aasiat 2 | India | 230644 | 197930 | 0.07633221 | 0.07633221 | 0.003415274 | 22.35024 |
| Honshu wolf | Korean Jindo | India | 237722 | 206389 | 0.07055218 | 0.07055218 | 0.003344597 | 21.09438 |
| Honshu wolf | Portugal | India | 261948 | 210566 | 0.1087418 | 0.1087418 | 0.003946003 | 27.55745 |
| Honshu wolf | Shanxi wolf 1 | India | 247471 | 215363 | 0.0693726 | 0.0693726 | 0.003398419 | 20.4132 |
| Honshu wolf | Shiba Inu 1 | India | 235789 | 206565 | 0.06606474 | 0.06606474 | 0.003456569 | 19.11281 |
| Honshu wolf | Greenland Sled Dog Tasiliq 1 | India | 231712 | 200207 | 0.07294192 | 0.07294192 | 0.003653725 | 19.96371 |
| Honshu wolf | Tumat | India | 229978 | 201438 | 0.06615425 | 0.06615425 | 0.003334713 | 19.83806 |
| Honshu wolf | Victoria Island | India | 242171 | 223129 | 0.04092413 | 0.04092413 | 0.00358452 | 11.41691 |
| Honshu wolf | Altai | India | 254923 | 190194 | 0.1454202 | 0.1454202 | 0.003823221 | 38.03604 |
| Honshu wolf | Alaska 1 | Korean Jindo | 225313 | 230590 | -0.01157483 | -0.01157483 | 0.004212746 | -2.747574 |
| Honshu wolf | Yana | Korean Jindo | 212889 | 215033 | -0.005010259 | -0.005010259 | 0.004146756 | -1.208236 |
| Honshu wolf | Bunge-Toll | Korean Jindo | 219542 | 227287 | -0.01733325 | -0.01733325 | 0.004227764 | -4.099863 |
| Honshu wolf | Tirektyakh | Korean Jindo | 221855 | 245073 | -0.04972501 | -0.04972501 | 0.004054002 | -12.26566 |
| Honshu wolf | Ulakhan Salur | Korean Jindo | 230940 | 223541 | 0.01628011 | 0.01628011 | 0.004175709 | 3.898765 |
| Honshu wolf | Greenland Sled Dog Aasiat 2 | Korean Jindo | 314536 | 168560 | 0.3021677 | 0.3021677 | 0.004606469 | 65.59637 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|---------------------------------|--------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | India | Korean Jindo | 237722 | 233864 | 0.008180904 | 0.008180904 | 0.004439013 | 1.842956 |
| Honshu wolf | Portugal | Korean Jindo | 244423 | 237098 | 0.01521221 | 0.01521221 | 0.004342126 | 3.503402 |
| Honshu wolf | Shanxi wolf 1 | Korean Jindo | 244910 | 234682 | 0.02132646 | 0.02132646 | 0.004238707 | 5.03136 |
| Honshu wolf | Shiba Inu 1 | Korean Jindo | 335381 | 171720 | 0.3227385 | 0.3227385 | 0.005133442 | 62.8698 |
| Honshu wolf | Greenland Sled Dog Tasiliq 1 | Korean Jindo | 315474 | 169971 | 0.2997312 | 0.2997312 | 0.004722928 | 63.46299 |
| Honshu wolf | Tunat | Korean Jindo | 230230 | 21591 | 0.03283605 | 0.03283605 | 0.004359522 | 7.53203 |
| Honshu wolf | Victoria Island | Korean Jindo | 236379 | 242566 | -0.01291798 | -0.01291798 | 0.00452913 | -2.852198 |
| Honshu wolf | Altai | Korean Jindo | 229892 | 215290 | 0.03280007 | 0.03280007 | 0.00450177 | 7.286038 |
| Honshu wolf | Alaska 1 | Portugal | 239718 | 206452 | 0.07455902 | 0.07455902 | 0.003949202 | 18.87952 |
| Honshu wolf | Yana | Portugal | 220066 | 196824 | 0.05575092 | 0.05575092 | 0.0035199 | 15.83878 |
| Honshu wolf | Bunge-Toll | Portugal | 225334 | 208378 | 0.03909507 | 0.03909507 | 0.003267377 | 11.96528 |
| Honshu wolf | Tirektyakh | Portugal | 232008 | 223207 | 0.01933372 | 0.01933372 | 0.003271286 | 5.910129 |
| Honshu wolf | Ulakhan Salur | Portugal | 241019 | 202457 | 0.08695397 | 0.08695397 | 0.003362927 | 25.85664 |
| Honshu wolf | Greenland Sled Dog Aasiaat 2 | Portugal | 233827 | 195505 | 0.08925959 | 0.08925959 | 0.003519826 | 25.35909 |
| Honshu wolf | India | Portugal | 261948 | 204677 | 0.1227345 | 0.1227345 | 0.00402615 | 30.48434 |
| Honshu wolf | Korean Jindo | Portugal | 244423 | 202657 | 0.09341952 | 0.09341952 | 0.003296848 | 28.33601 |
| Honshu wolf | Shanxi wolf 1 | Portugal | 259118 | 211291 | 0.1016711 | 0.1016711 | 0.003543451 | 28.69268 |
| Honshu wolf | Shiba Inu 1 | Portugal | 244985 | 202025 | 0.09610523 | 0.09610523 | 0.003402535 | 28.24519 |
| Honshu wolf | Greenland Sled Dog Tasiliq 1 | Portugal | 235798 | 197613 | 0.08810344 | 0.08810344 | 0.003935951 | 22.38428 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|---------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Tumat | Portugal | 237950 | 196056 | 0.09652862 | 0.09652862 | 0.003497719 | 27.59759 |
| Honshu wolf | Victoria Island | Portugal | 250915 | 218046 | 0.07008898 | 0.07008898 | 0.004258735 | 16.4577 |
| Honshu wolf | Altai | Portugal | 248156 | 191156 | 0.1297483 | 0.1297483 | 0.003945672 | 32.88371 |
| Honshu wolf | Alaska 1 | Shanxi wolf 1 | 238138 | 214267 | 0.05276467 | 0.05276467 | 0.003482344 | 15.15206 |
| Honshu wolf | Yana | Shanxi wolf 1 | 218221 | 205329 | 0.03043796 | 0.03043796 | 0.00319097 | 9.53878 |
| Honshu wolf | Bunge-Toll | Shanxi wolf 1 | 223804 | 216561 | 0.01644772 | 0.01644772 | 0.003142452 | 5.23404 |
| Honshu wolf | Tirektyakh | Shanxi wolf 1 | 229106 | 232119 | -0.006532603 | -0.006532603 | 0.002913231 | -2.242391 |
| Honshu wolf | Ulakhan Salur | Shanxi wolf 1 | 237005 | 212104 | 0.05544534 | 0.05544534 | 0.002972966 | 18.64984 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiat 2 | Shanxi wolf 1 | 235718 | 201591 | 0.07803864 | 0.07803864 | 0.003404639 | 22.92127 |
| Honshu wolf | India | Shanxi wolf 1 | 247471 | 218313 | 0.06259983 | 0.06259983 | 0.003781222 | 16.55545 |
| Honshu wolf | Korean Jindo | Shanxi wolf 1 | 244910 | 209216 | 0.07859933 | 0.07859933 | 0.003384275 | 23.22487 |
| Honshu wolf | Portugal | Shanxi wolf 1 | 259118 | 220381 | 0.0807864 | 0.0807864 | 0.003951643 | 20.44375 |
| Honshu wolf | Shiba Inu 1 | Shanxi wolf 1 | 243258 | 209683 | 0.07412665 | 0.07412665 | 0.003375225 | 21.96199 |
| | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasilag 1 | Shanxi wolf 1 | 236663 | 203352 | 0.07570424 | 0.07570424 | 0.003471407 | 21.80794 |
| Honshu wolf | Tumat | Shanxi wolf 1 | 235925 | 205022 | 0.07008325 | 0.07008325 | 0.003141673 | 22.30762 |
| Honshu wolf | Victoria Island | Shanxi wolf 1 | 248280 | 226579 | 0.04569988 | 0.04569988 | 0.003494839 | 13.07639 |
| Honshu wolf | Altai | Shanxi wolf 1 | 248269 | 198450 | 0.111522 | 0.111522 | 0.003505764 | 31.81104 |
| Honshu wolf | Alaska 1 | Shiba Inu 1 | 224838 | 234813 | -0.02170125 | -0.02170125 | 0.004405689 | -4.925733 |
| Honshu wolf | Yana | Shiba Inu 1 | 212382 | 220531 | -0.01882364 | -0.01882364 | 0.004125495 | -4.56276 |
| Honshu wolf | Bunge-Toll | Shiba Inu 1 | 218470 | 232606 | -0.0313384 | -0.0313384 | 0.004208627 | -7.446229 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|--------------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Tirektyakh | Shiba Inu 1 | 221942 | 250646 | -0.06073789 | -0.06073789 | 0.003989951 | -15.22272 |
| Honshu wolf | Ulakhan Salur | Shiba Inu 1 | 231360 | 228692 | 0.005799344 | 0.005799344 | 0.004190363 | 1.383972 |
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiaat 2 | Shiba Inu 1 | 312843 | 173682 | 0.2860305 | 0.2860305 | 0.005062208 | 56.50312 |
| Honshu wolf | India | Shiba Inu 1 | 235789 | 239003 | -0.00676928 | -0.00676928 | 0.004818382 | -1.404886 |
| Honshu wolf | Korean Jindo | Shiba Inu 1 | 335381 | 176149 | 0.3112858 | 0.3112858 | 0.005577721 | 55.80877 |
| Honshu wolf | Portugal | Shiba Inu 1 | 244985 | 241575 | 0.007008385 | 0.007008385 | 0.004324409 | 1.620657 |
| Honshu wolf | Shanxi wolf 1 | Shiba Inu 1 | 243258 | 240011 | 0.006718825 | 0.006718825 | 0.004099859 | 1.638794 |
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Shiba Inu 1 | 315440 | 174736 | 0.2870479 | 0.2870479 | 0.005241871 | 54.76059 |
| Honshu wolf | Tumat | Shiba Inu 1 | 227754 | 221764 | 0.01332538 | 0.01332538 | 0.00432404 | 3.081698 |
| Honshu wolf | Victoria Island | Shiba Inu 1 | 236060 | 247474 | -0.02360537 | -0.02360537 | 0.004643352 | -5.083693 |
| Honshu wolf | Altai | Shiba Inu 1 | 228763 | 220405 | 0.01860774 | 0.01860774 | 0.004586078 | 4.05744 |
| Honshu wolf | Alaska 1 | Greenland Sled Dog | | | | | | |
| Honshu wolf | Alaska 1 | Tasiliag 1 | 225301 | 231462 | -0.0134884 | -0.0134884 | 0.004329667 | -3.115342 |
| Honshu wolf | Yana | Greenland Sled Dog | | | | | | |
| Honshu wolf | Yana | Tasiliag 1 | 213984 | 215206 | -0.002847224 | -0.002847224 | 0.004211796 | -0.6760118 |
| Honshu wolf | Bunge-Toll | Greenland Sled Dog | | | | | | |
| Honshu wolf | Bunge-Toll | Tasiliag 1 | 220659 | 226752 | -0.01361835 | -0.01361835 | 0.004213919 | -3.231754 |
| Honshu wolf | Tirektyakh | Greenland Sled Dog | | | | | | |
| Honshu wolf | Tirektyakh | Tasiliag 1 | 223226 | 244422 | -0.04532469 | -0.04532469 | 0.004017792 | -11.281 |
| Honshu wolf | Ulakhan Salur | Greenland Sled Dog | | | | | | |
| Honshu wolf | Ulakhan Salur | Tasiliag 1 | 232218 | 222954 | 0.02035275 | 0.02035275 | 0.004294007 | 4.739803 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------------------|---------------------------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Greenland Sled Dog Aasiat 2 | Greenland Sled Dog Tasilaq 1 | 414009 | 121249 | 0.5469512 | 0.5469512 | 0.005882145 | 92.98499 |
| Honshu wolf | India | Greenland Sled Dog Tasilaq 1 | 231712 | 236610 | -0.01045862 | -0.01045862 | 0.004196704 | -2.492102 |
| Honshu wolf | Korean Jindo | Greenland Sled Dog Tasilaq 1 | 315474 | 180742 | 0.2715189 | 0.2715189 | 0.004907606 | 55.32613 |
| Honshu wolf | Portugal | Greenland Sled Dog Tasilaq 1 | 235798 | 241954 | -0.01288535 | -0.01288535 | 0.004085147 | -3.154194 |
| Honshu wolf | Shanxi wolf 1 | Greenland Sled Dog Tasilaq 1 | 236663 | 238781 | -0.004454783 | -0.004454783 | 0.00401964 | -1.108254 |
| Honshu wolf | Shiba Inu 1 | Greenland Sled Dog Tasilaq 1 | 315440 | 180665 | 0.2716663 | 0.2716663 | 0.004915783 | 55.26409 |
| Honshu wolf | Tumat | Greenland Sled Dog Tasilaq 1 | 228370 | 216730 | 0.02615143 | 0.02615143 | 0.004272279 | 6.121188 |
| Honshu wolf | Victoria Island | Greenland Sled Dog Tasilaq 1 | 234079 | 244275 | -0.02131476 | -0.02131476 | 0.004498463 | -4.738231 |
| Honshu wolf | Altai | Greenland Sled Dog Tasilaq 1 | 227209 | 216311 | 0.02457161 | 0.02457161 | 0.004499704 | 5.460717 |
| Honshu wolf | Alaska 1 | Tumat | 228438 | 218700 | 0.02177851 | 0.02177851 | 0.003311536 | 6.576559 |
| Honshu wolf | Yana | Tumat | 221462 | 201110 | 0.04816221 | 0.04816221 | 0.003379523 | 14.25118 |
| Honshu wolf | Bunge-Toll | Tumat | 226912 | 212330 | 0.0331981 | 0.0331981 | 0.003310176 | 10.0291 |
| Honshu wolf | Tirektyakh | Tumat | 229310 | 229654 | 0.0007495141 | -0.0007495141 | 0.003141551 | -0.2385809 |
| Honshu wolf | Ulakhan Salur | Tumat | 244968 | 205727 | 0.08706775 | 0.08706775 | 0.00321811 | 27.05555 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|---------------------------------|-----------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Greenland Sled Dog Aasiat 2 | Tumat | 227082 | 205457 | 0.04999549 | 0.04999549 | 0.00329465 | 15.17475 |
| Honshu wolf | India | Tumat | 229978 | 225797 | 0.009173386 | 0.009173386 | 0.00328752 | 2.790366 |
| Honshu wolf | Korean Jindo | Tumat | 230230 | 217007 | 0.02956598 | 0.02956598 | 0.003418636 | 8.648473 |
| Honshu wolf | Portugal | Tumat | 237950 | 229589 | 0.017883 | 0.017883 | 0.003361974 | 5.319196 |
| Honshu wolf | Shanxi wolf 1 | Tumat | 235925 | 229535 | 0.01372835 | 0.01372835 | 0.003166486 | 4.335518 |
| Honshu wolf | Shiba Inu 1 | Tumat | 227754 | 218300 | 0.02119474 | 0.02119474 | 0.003349138 | 6.328418 |
| Honshu wolf | Greenland Sled Dog Tasiliq 1 | Tumat | 228370 | 206922 | 0.04927267 | 0.04927267 | 0.003333087 | 14.7829 |
| Honshu wolf | Victoria Island | Tumat | 237522 | 231397 | 0.01306196 | 0.01306196 | 0.00367697 | 3.55237 |
| Honshu wolf | Altai | Tumat | 226084 | 206563 | 0.04511992 | 0.04511992 | 0.003330454 | 13.54768 |
| Honshu wolf | Alaska 1 | Victoria Island | 286277 | 187082 | 0.2095555 | 0.2095555 | 0.004538035 | 46.1776 |
| Honshu wolf | Yana | Victoria Island | 219607 | 198132 | 0.0514077 | 0.0514077 | 0.003209822 | 16.01575 |
| Honshu wolf | Bunge-Toll | Victoria Island | 225620 | 209721 | 0.0365208 | 0.0365208 | 0.003210568 | 11.37518 |
| Honshu wolf | Tirektyakh | Victoria Island | 231471 | 224633 | 0.01499219 | 0.01499219 | 0.003128621 | 4.791951 |
| Honshu wolf | Ulakhan Salur | Victoria Island | 241098 | 203544 | 0.08445896 | 0.08445896 | 0.003209921 | 26.31185 |
| Honshu wolf | Greenland Sled Dog Aasiat 2 | Victoria Island | 233179 | 198093 | 0.08135469 | 0.08135469 | 0.003564449 | 22.82392 |
| Honshu wolf | India | Victoria Island | 242171 | 215687 | 0.05784326 | 0.05784326 | 0.003542353 | 16.32905 |
| Honshu wolf | Korean Jindo | Victoria Island | 236379 | 208025 | 0.06380231 | 0.06380231 | 0.003462021 | 18.42921 |
| Honshu wolf | Portugal | Victoria Island | 250915 | 218065 | 0.07004563 | 0.07004563 | 0.003321656 | 21.08756 |
| Honshu wolf | Shanxi wolf 1 | Victoria Island | 248280 | 217339 | 0.06645133 | 0.06645133 | 0.003228239 | 20.58439 |

| H1 | H2 | H3 | ABBA sites | BABA sites | Dstat score | JackEst score | Standard Error | Z-score |
|-------------|--------------------|-----------------|-------------------|-------------------|--------------------|----------------------|-----------------------|----------------|
| Honshu wolf | Shiba Inu 1 | Victoria Island | 236060 | 208545 | 0.06188639 | 0.06188639 | 0.003488905 | 17.73806 |
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Victoria Island | 234079 | 199605 | 0.07949106 | 0.07949106 | 0.003586822 | 22.16197 |
| Honshu wolf | Tumat | Victoria Island | 237522 | 197460 | 0.09210036 | 0.09210036 | 0.003395799 | 27.12185 |
| Honshu wolf | Altai | Victoria Island | 237158 | 196761 | 0.09309802 | 0.09309802 | 0.003732639 | 24.94161 |
| Honshu wolf | Alaska 1 | Altai | 230227 | 206188 | 0.05508289 | 0.05508289 | 0.003751196 | 14.68409 |
| Honshu wolf | Yana | Altai | 214513 | 196058 | 0.04494959 | 0.04494959 | 0.003302945 | 13.60894 |
| Honshu wolf | Bunge-Toll | Altai | 219502 | 208152 | 0.02654015 | 0.02654015 | 0.003222797 | 8.235129 |
| Honshu wolf | Tirektyakh | Altai | 225106 | 221443 | 0.008202907 | 0.008202907 | 0.003140926 | 2.61162 |
| Honshu wolf | Ulakhan Salur | Altai | 233993 | 202228 | 0.07281859 | 0.07281859 | 0.003251958 | 22.39223 |
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Aasiat 2 | Altai | 226523 | 193690 | 0.07813418 | 0.07813418 | 0.003406297 | 22.93816 |
| Honshu wolf | India | Altai | 254923 | 204684 | 0.1093086 | 0.1093086 | 0.003848968 | 28.39946 |
| Honshu wolf | Korean Jindo | Altai | 229892 | 202075 | 0.06439612 | 0.06439612 | 0.003521256 | 18.28783 |
| Honshu wolf | Portugal | Altai | 248156 | 209523 | 0.08441069 | 0.08441069 | 0.003724216 | 22.66536 |
| Honshu wolf | Shanxi wolf 1 | Altai | 248269 | 208300 | 0.08754208 | 0.08754208 | 0.003320194 | 26.36656 |
| Honshu wolf | Shiba Inu 1 | Altai | 228763 | 202210 | 0.06161175 | 0.06161175 | 0.003500553 | 17.60057 |
| Honshu wolf | Greenland Sled Dog | | | | | | | |
| Honshu wolf | Tasiliag 1 | Altai | 227209 | 195244 | 0.07566522 | 0.07566522 | 0.003507498 | 21.57242 |
| Honshu wolf | Tumat | Altai | 226084 | 198328 | 0.06539872 | 0.06539872 | 0.003476491 | 18.8117 |
| Honshu wolf | Victoria Island | Altai | 237158 | 217638 | 0.04292034 | 0.04292034 | 0.003828465 | 11.21085 |

Table S3. Populations used in GLOBETROTTER and SOURCEFIND analyses

| ID | Assigned population | Donor/Surrogate |
|----------------------|----------------------------|------------------------|
| AfghanDog | Eurasian dog | Donor |
| Alaska1 | North American wolf | Donor |
| AlaskanHusky_SY001 | Sled dog | Surrogate |
| AlaskanHusky_SY018 | Sled dog | Donor |
| AlaskanMalamuteDog | Sled dog | Donor |
| AlaskanWolf | North American wolf | Surrogate |
| AnhuiDog | Chinese village dog | Donor |
| BanksIsland | North American wolf | Surrogate |
| BasenjiDog | African dog | Donor |
| BM | Eurasian dog | Donor |
| BoxerDog | European dog | Donor |
| CGG23 | Pleistocene wolf | Donor |
| CGG29 | Pleistocene wolf | Surrogate |
| CGG32 | Pleistocene wolf | Donor |
| CGG33 | Pleistocene wolf | Surrogate |
| CGG6 | Ancient sled dog | Donor |
| ChihuahuaDog | Latin American dog | Donor |
| ChinaVietnam4Dog | Chinese village dog | Donor |
| ChowChow01 | Chinese dog | Donor |
| ChowChow02 | Chinese dog | Donor |
| CI1 | Chinese village dog | Donor |
| CI2 | Chinese village dog | Donor |
| CI3 | Chinese village dog | Donor |
| CTC | Ancient European dog | Donor |
| DalianDog | Chinese village dog | Donor |
| EastSiberianLaikaDog | Siberian dog | Donor |
| EG44 | Egyptian village dog | Donor |
| EG49 | Egyptian village dog | Donor |
| Ellesmere1 | North American wolf | Donor |
| GalgoDog | Latin American dog | Donor |
| Gansu2Dog | Chinese village dog | Donor |

| | | |
|--------------------------|------------------------|-----------|
| Gansu3Dog | Chinese village dog | Donor |
| Greenland_wolf_A1 | North American wolf | Donor |
| GreenlandDog | Greenland dog | Surrogate |
| Greenlandic_dog_Mums | Greenland dog | Donor |
| Greenlandic_dog_Pondus | Greenland dog | Surrogate |
| GreyNorwegianElkhoundDog | Scandinavian dog | Donor |
| GS | European dog | Donor |
| GShepDog | European dog | Donor |
| GuangdongDog | Chinese village dog | Donor |
| GuizhouDog | Chinese village dog | Donor |
| GW1 | Eurasian wolf | Surrogate |
| GW2 | Eurasian wolf | Donor |
| GW3 | Eurasian wolf | Donor |
| GW4 | Eurasian wolf | Surrogate |
| HebeiDog | Chinese village dog | Donor |
| Honshu | Honshu wolf | Surrogate |
| HXH | Ancient European dog | Donor |
| IberianWolf | Eurasian wolf | Donor |
| ID125 | Indian village dog | Donor |
| ID137 | Indian village dog | Donor |
| ID165 | Indian village dog | Donor |
| ID168 | Indian village dog | Donor |
| ID60 | Indian village dog | Donor |
| ID91 | Indian village dog | Donor |
| Ilulissat_GS16 | Greenland dog | Surrogate |
| Ilulissat_GS31 | Greenland dog | Donor |
| IN18 | Indonesian village dog | Donor |
| IN23 | Indonesian village dog | Donor |
| IN29 | Indonesian village dog | Donor |
| IndiaWolf | Eurasian wolf | Surrogate |
| InnerMongoliaWolf | Eurasian wolf | Donor |
| IranWolf | Eurasian wolf | Donor |
| JapaneseChin | Chinese dog | Surrogate |
| KoreanJindo | Japanese dog | Surrogate |

| | | |
|------------------------|----------------------------|-----------|
| LapponianHerderDog | Scandinavian dog | Donor |
| LB74 | Middle-Eastern village dog | Donor |
| LB79 | Middle-Eastern village dog | Donor |
| LB85 | Middle-Eastern village dog | Donor |
| Mexican wolf | Mexican wolf | Donor |
| Mexican_wolf | Mexican wolf | Donor |
| MexicanNakedDog | Latin American dog | Donor |
| Newgrange | Ancient European dog | Donor |
| NorthBaffin | North American wolf | Donor |
| Novembre_Chinese_Wolf | Eurasian wolf | Donor |
| Novembre_Croatian_Wolf | Eurasian wolf | Donor |
| Novembre_Israeli_Wolf | Eurasian wolf | Donor |
| Pekingese | Chinese dog | Surrogate |
| PeruvianNakedDog | Latin American dog | Donor |
| PG115 | Papuan village dog | Donor |
| PG122 | Papuan village dog | Donor |
| PG84 | Papuan village dog | Donor |
| PortugueseWolf | Eurasian wolf | Surrogate |
| QA27 | Middle-Eastern village dog | Donor |
| Qaanaaq_Q11 | Greenland dog | Donor |
| Qaanaaq_QSON | Greenland dog | Surrogate |
| SAMN03168368 | Chinese village dog | Donor |
| SAMN03168369 | Chinese village dog | Donor |
| SAMN03168370 | Chinese village dog | Donor |
| SAMN03168372 | Chinese village dog | Donor |
| SAMN03168373 | Nigerian dog | Donor |
| SAMN03168374 | Nigerian dog | Donor |
| SAMN03168375 | Nigerian dog | Donor |
| SAMN03168383 | Scandinavian dog | Donor |
| SAMN03168391 | Scandinavian dog | Donor |
| SAMN03168394 | Eurasian wolf | Donor |
| SamoyedDog | Siberian dog | Donor |
| Shanxi1Dog | Chinese village dog | Donor |
| Shanxi1Wolf | Eurasian wolf | Donor |

| | | |
|--------------------|---------------------|-----------|
| Shanxi2Dog | Chinese village dog | Donor |
| Shanxi2Wolf | Eurasian wolf | Donor |
| Shanxi3Dog | Chinese village dog | Donor |
| Shanxi4Dog | Chinese village dog | Donor |
| SharPei01 | Chinese dog | Donor |
| SharPei02 | Chinese dog | Donor |
| ShibaInuFemale | Japanese dog | Surrogate |
| ShibaInuMale | Japanese dog | Surrogate |
| ShihTzu | Chinese dog | Surrogate |
| SiberianHusky_SYXX | Sled dog | Donor |
| SiberianHusky01 | Sled dog | Surrogate |
| SiberianHuskyDog | Sled dog | Donor |
| Sisimiut_02_06_16 | Greenland dog | Donor |
| Sisimiut_19_05_16 | Greenland dog | Surrogate |
| SloughiDog | African dog | Donor |
| SouthBaffin | North American wolf | Donor |
| SwedishLapphundDog | Scandinavian dog | Donor |
| TarabaDog | Nigerian dog | Donor |
| Tasiilaq_51602 | Greenland dog | Donor |
| Tasiilaq_51603 | Greenland dog | Surrogate |
| Tumat | Pleistocene wolf | Donor |
| VictorialIsland | North American wolf | Donor |
| Wang_TM | Chinese dog | Donor |
| XinjiangDog | Chinese dog | Surrogate |
| Yunnan1Dog | Chinese village dog | Donor |
| Yunnan2Dog | Chinese village dog | Donor |

Table S4. GLOBETROTTER results for the best fitting admixture event (1-date admixture)

| Generations | Source 1 proportion | R2 Fit | Best Matching Source 1 | Best Matching Source 2 |
|-------------|---------------------|--------------|------------------------|------------------------|
| 24.95247758 | 0.07 | 0.6976536549 | Honshu wolf | Chinese dogs |

| | | |
|------------|----------------|--------------|
| proportion | Chinese dogs | Honshu wolf |
| 0.07 | 0.08957398834 | 0.9104260117 |
| proportion | Honshu wolf | Chinese dogs |
| 0.93 | 0.008228450642 | 0.9917715494 |

References

- Auton, Adam, Ying Rui Li, Jeffrey Kidd, Kyle Oliveira, Julie Nadel, J. Kim Holloway, Jessica J. Hayward, et al. 2013. "Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs." *PLoS Genetics* 9 (12): e1003984.
- Botigué, Laura R., Shiya Song, Amelie Scheu, Shyamalika Gopalan, Amanda L. Pendleton, Matthew Oetjens, Angela M. Taravella, et al. 2017. "Ancient European Dog Genomes Reveal Continuity since the Early Neolithic." *Nature Communications* 8 (July): 16082.
- Cahill, James A., Zhenxin Fan, Ilan Gronau, Jacqueline Robinson, John P. Pollinger, Beth Shapiro, Jeff Wall, Robert K. Wayne, and Others. 2016. "Whole-Genome Sequence Analysis Shows That Two Endemic Species of North American Wolf Are Admixtures of the Coyote and Gray Wolf." *Science Advances* 2 (7): e1501714.
- Decker, Brennan, Brian W. Davis, Maud Rimbault, Adrienne H. Long, Eric Karlins, Vidhya Jagannathan, Rebecca Reiman, et al. 2015. "Comparison against 186 Canid Whole-Genome Sequences Reveals Survival Strategies of an Ancient Clonally Transmissible Canine Tumor." *Genome Research* 25 (11): 1646–55.
- Frantz, Laurent A. F., Victoria E. Mullin, Maud Pionnier-Capitan, Ophélie Lebrasseur, Morgane Ollivier, Angela Perri, Anna Linderholm, et al. 2016. "Genomic and Archaeological Evidence Suggest a Dual Origin of Domestic Dogs." *Science* 352 (6290): 1228–31.
- Freedman, Adam H., Ilan Gronau, Rena M. Schweizer, Diego Ortega-Del Vecchyo, Eunjung Han, Pedro M. Silva, Marco Galaverni, et al. 2014. "Genome Sequencing Highlights the Dynamic Early History of Dogs." *PLoS Genetics* 10 (1): e1004016.
- Gopalakrishnan, Shyam, Mikkel-Holger S. Sinding, Jazmín Ramos-Madrugal, Jonas Niemann, Jose A. Samaniego Castruita, Filipe G. Vieira, Christian Carøe, et al. 2018. "Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*." *Current Biology: CB*, October. <https://doi.org/10.1016/j.cub.2018.08.041>.
- Kim, Ryong Nam, Dae-Soo Kim, Sang-Haeng Choi, Byoung-Ha Yoon, Aram Kang, Seong-Hyeuk Nam, Dong-Wook Kim, et al. 2012. "Genome Analysis of the Domestic Dog (Korean Jindo) by Massively Parallel Sequencing." *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes* 19 (3): 275–87.
- Kolichski, A., G. S. Johnson, N. A. Villani, D. P. O'Brien, T. Mhlanga-Mutangadura, D. A. Wenger, K. Mikoloski, et al. 2017. "GM 2 Gangliosidosis in Shiba Inu Dogs with an In-Frame Deletion in *HEXB*." *Journal of Veterinary Internal Medicine / American College of Veterinary Internal Medicine* 31 (5): 1520–26.
- Lindblad-Toh, Kerstin, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, et al. 2005. "Genome Sequence, Comparative Analysis and Haplotype Structure of the Domestic Dog." *Nature* 438 (7069): 803–19.

- Marchant, Thomas W., Edward J. Johnson, Lynn McTeir, Craig I. Johnson, Adam Gow, Tiziana Liuti, Dana Kuehn, et al. 2017. "Canine Brachycephaly Is Associated with a Retrotransposon-Mediated Missplicing of SMOC2." *Current Biology: CB* 27 (11): 1573–84.e6.
- Metzger, Julia, Anna Nolte, Ann-Kathrin Uhde, Marion Hewicker-Trautwein, and Ottmar Distl. 2017. "Whole Genome Sequencing Identifies Missense Mutation in MTBP in Shar-Pei Affected with Autoinflammatory Disease (SPAID)." *BMC Genomics* 18 (1): 348.
- Shannon, Laura M., Ryan H. Boyko, Marta Castelhano, Elizabeth Corey, Jessica J. Hayward, Corin McLean, Michelle E. White, et al. 2015. "Genetic Structure in Village Dogs Reveals a Central Asian Domestication Origin." *Proceedings of the National Academy of Sciences of the United States of America* 112 (44): 13639–44.
- Sinding, Mikkel-Holger S., Shyam Gopalakrishnan, Filipe G. Vieira, Jose A. Samaniego Castruita, Katrine Raundrup, Mads Peter Heide Jørgensen, Morten Meldgaard, et al. 2018. "Population Genomics of Grey Wolves and Wolf-like Canids in North America." *PLoS Genetics* 14 (11): e1007745.
- Skoglund, Pontus, Erik Ersmark, Eleftheria Palkopoulou, and Love Dalén. 2015. "Ancient Wolf Genome Reveals an Early Divergence of Domestic Dog Ancestors and Admixture into High-Latitude Breeds." *Current Biology: CB* 25 (11): 1515–19.
- Wang, Guo-Dong, Weiwei Zhai, He-Chuan Yang, Ruo-Xi Fan, Xue Cao, Li Zhong, Lu Wang, et al. 2013. "The Genomics of Selection in Dogs and the Parallel Evolution between Dogs and Humans." *Nature Communications* 4: 1860.
- Wang, Guo-Dong, Weiwei Zhai, He-Chuan Yang, Lu Wang, Li Zhong, Yan-Hu Liu, Ruo-Xi Fan, et al. 2016. "Out of Southern East Asia: The Natural History of Domestic Dogs across the World." *Cell Research* 26 (1): 21–33.
- Wiedmer, Michaela, Anna Oevermann, Stephanie E. Borer-Germann, Daniela Gorgas, G. Diane Shelton, Michaela Drögemüller, Vidhya Jagannathan, Diana Henke, and Tosso Leeb. 2016. "A RAB3GAP1 SINE Insertion in Alaskan Huskies with Polyneuropathy, Ocular Abnormalities, and Neuronal Vacuolation (POANV) Resembling Human Warburg Micro Syndrome 1 (WARBM1)." *G3: Genes|Genomes|Genetics*. <https://doi.org/10.1534/g3.115.022707>.
- Zhang, Wenping, Zhenxin Fan, Eunjung Han, Rong Hou, Liang Zhang, Marco Galaverni, Jie Huang, et al. 2014. "Hypoxia Adaptations in the Grey Wolf (*Canis Lupus Chanco*) from Qinghai-Tibet Plateau." *PLoS Genetics* 10 (7): e1004466.
- Zhang, Z., J. Xing, C. Vilà, and T. Marques-Bonet. 2016. "Worldwide Patterns of Genomic Variation and Admixture in Gray Wolves." *Genome / National Research Council Canada = Genome / Conseil National de Recherches Canada*. <http://genome.cshlp.org/content/26/2/163.short>.

Chapter 3

Unsealing the jars - characterizing gut microbial DNA preservation in fluid-preserved museum specimens

Unsealing the jars - characterizing gut microbial DNA preservation in fluid-preserved museum specimens

Jonas Niemann^{1,2}, Jessica E Thomas³, Marcela Sandoval-Velasco¹, Jan Bolding Kristensen⁴, Sarah Mak¹, Michael Knapp⁵, Nathan Wales², M Thomas P Gilbert^{1,6}

¹The GLOBE Institute, University of Copenhagen, Copenhagen, Denmark.

²BioArch, Department of Archaeology, University of York, York, UK.

³Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, United Kingdom;

⁴Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark;

⁵Department of Anatomy, University of Otago, Dunedin, New Zealand

⁶University Museum, Norwegian University of Science and Technology, Trondheim, Norway.

Abstract

Fluid-preserved collections are a vital part of natural history museums. Aside from their valuable contribution to the conservation of anatomical features of a wide range of organisms, previous biomolecular studies furthermore demonstrated that it is possible to extract host DNA from fluid-preserved specimens. In this pilot study we expand on these findings and investigate the DNA preservation of fluid-preserved gut microbiomes, i.e. DNA originating from the sampled organism itself and not from its microbiome. To do this, we sequenced stomach and intestine tissues of six historical seabird specimens dating from 1873 to 1944. We succeeded in recovering the gut microbiome of a razorbill dating back to 1916, highlighting the potential for future metagenomic studies on fluid-preserved samples.

Introduction

Natural history collections are not only valuable for documenting the phenotypic diversity and anatomy of species, but since the emergence of high-throughput sequencing and recent advances in the analysis of heavily degraded DNA, they have increasingly been used to obtain and analyse the genomes of ancient and historical samples (Green and Speller 2017), in a sense “unlocking the vault” (Bi et al. 2013). While most studies focus on the endogenous DNA of the specimen (Dabney et al. 2013; Noonan et al. 2006), there have also been efforts over the last few decades to examine single pathogens (Bos et al. 2011) as well as entire microbial communities (microbiomes) from archaeological remains (Warinner et al. 2014). For

example, analyses of ancient DNA (aDNA) extracted from dental calculus (Warinner, Speller, and Collins 2015; Preus et al. 2011), coprolites (Poinar et al. 2001; Iñiguez et al. 2006), and mummies (Lugli et al. 2017) has shed light into the past diet and health of archaic and prehistoric humans. In these materials, it is hypothesised that mineralisation, and/or rapid tissue desiccation, preserves both the microbial and host DNA, ultimately enabling the long term preservation of the dietary or microbial taxa present in the sample (Weyrich et al. 2017).

In addition to skeletal remains and dried/tanned materials such as skins, museum collections often also contain extensive ‘wet’ collections of formalin-, ethanol- or other spirit-preserved soft tissues. In particular, zoologists have preserved animals in alcohol since at least the 17th century (Down 1989), and today most natural history museums house large collections of fluid-preserved organisms. For example, the Zoological Museum of Copenhagen alone holds more than 10,000 fluid-preserved bird specimens. Researchers have already succeeded in recovering the genomes of fluid-preserved specimens (Shokralla, Singer, and Hajibabaei 2010; Miller et al. 2013) and even host-associated pathogens (Hühns et al. 2017; Devault et al. 2014). It is important to recognize that all specimens in natural history collections once served as host to a diverse microbial communities. At present it is unknown whether such microbiomes are preserved alongside the tissues of the fluid-preserved organisms; however, if microbial tissues or biomolecules persist in the collection jars, researchers would have a valuable new option to explore past biological diversity. For example, given that fluid- preserved samples date back centuries, the recovery of fluid-preserved microbiomes could not only be used to catalogue the diversity of microbes found associated with host organisms, but also be used to identify shifts in any species’ microbial profile over time. Also, as fluid preservation generally retains the external and internal structure of the specimen, there is the potential to selectively sample organ-specific microbiomes of historical specimens, and identify their diet by sampling the

content of the digestive system. Moreover, as many specimens were collected before the Industrial Revolution, the discovery of antibiotics, and recent global warming, fluid-preserved specimens may represent a potential treasure trove with regards to information on the impact of humans on the environment.

So far no metagenomic studies have characterised historic-era fluid-preserved microbiomes, and indeed it is currently unclear whether fluid-preserved, historic soft tissues even retain microbiome information. This could potentially be due simply to microbial DNA not being preserved in such materials. While good preservation of host DNA has been observed in samples that were stored and fixed in 95-100% ethyl alcohol for 2-5 years (Mandrioli, Borsatti, and Mola 2006; Chakraborty, Sakai, and Iwatsuki 2006), certainly there are some tissue preservation practices that are known to induce DNA damage, such as the widespread use of formaldehyde during fixation (Down 1989; Srinivasan, Sedmak, and Jewell 2002) or exposure to ultraviolet light. An even greater challenge is perhaps that a range of preservatives and additives (e.g. seawater, brandy, vinegar, mercury; see Simmons, J. (2014) *Fluid Preservation: a comprehensive review*: pp 199-279) have been used in the past, many of which were unrecorded, and thus have unknown ramifications for the DNA preservation.

To explore the potential of fluid-preserved specimens for microbiome research, we performed a targeted investigation of soft tissues of six historical seabird specimens, dating from 1873 to 1944. We assessed whether it is possible to recover the gut microbiome profile from these samples, using shotgun sequencing of liquid-preserved stomach and intestine tissue, and subsequent analysis with metagenomic pipelines to distinguish authentically historic microbiomes.

Material and methods

Sample Collection

Samples were obtained from museum specimens stored in the Zoological Collections at the Natural History Museum of Denmark, Copenhagen (Table 1). Specimens had been collected between 1873 and 1944. For tissue sampling, specimens were removed from their storage jars, placed in a sterile tray and dissected using a sterile, disposable scalpel. A new scalpel was used for each specimen. Once the gut cavity was open, samples were collected from the intestines or stomach. Liquid contents were pipetted into labelled tubes, and solid samples were placed in labelled tubes using a scalpel.

Table 1. List of specimens with sample type and collection year

| Sample ID | Species name | Common name | Tissue type | Sample type | Collection year |
|-----------|----------------------------|-----------------|-------------------|---------------|-----------------|
| 1-I | <i>Phalacrocorax carbo</i> | Great cormorant | <i>Intestines</i> | <i>liquid</i> | 1873 |
| 1-S | <i>Phalacrocorax carbo</i> | Great cormorant | <i>Stomach</i> | <i>liquid</i> | 1873 |
| 2-I | <i>Alca torda</i> | Razorbill | <i>Intestines</i> | <i>liquid</i> | 1908 |
| 2-S | <i>Alca torda</i> | Razorbill | <i>Stomach</i> | <i>solid</i> | 1908 |
| 3-I | <i>Cephus grylle</i> | Black guillemot | <i>Intestines</i> | <i>liquid</i> | 1908 |
| 3-S | <i>Cephus grylle</i> | Black guillemot | <i>Stomach</i> | <i>solid</i> | 1908 |
| 4-I | <i>Uria aalge</i> | Common murre | <i>Intestines</i> | <i>liquid</i> | 1944 |
| 4-S | <i>Uria aalge</i> | Common murre | <i>Stomach</i> | <i>solid</i> | 1944 |
| 5-I | <i>Alca torda</i> | Razorbill | <i>Intestines</i> | <i>liquid</i> | 1884 |
| 5-S | <i>Alca torda</i> | Razorbill | <i>Stomach</i> | <i>liquid</i> | 1884 |
| 6-I | <i>Alca torda</i> | Razorbill | <i>Intestines</i> | <i>liquid</i> | 1916 |
| 6-S | <i>Alca torda</i> | Razorbill | <i>Stomach</i> | <i>solid</i> | 1916 |

DNA Extraction and Shotgun Sequence Data Generation

All laboratory work prior to polymerase chain reaction (PCR) amplification was carried out in the designated ancient DNA (aDNA) laboratories of the Natural History Museum of Denmark. Strict aDNA protocols were followed to avoid contamination. For each DNA extraction and library build, no-template controls were used to test for contamination by exogenous DNA. All post-PCR work was carried out in separate laboratory facilities (Knapp, Clarke, Horsburgh, & Matisoo-Smith, 2012).

Genomic DNA extraction method was dependent on sample type. For liquid samples initial steps to adjust the salt concentration were performed by adding 10% sample volume of 3M sodium acetate, followed by 0.7 volumes of room temperature isopropanol and vortexed to mix well. The samples were then centrifuged for 30 seconds at $12,500 \times g$. The liquid was transferred into a new tube and 1ml 70% ethanol was added. Samples were centrifuged again at $12,500 \times g$ for 15 minutes. The ethanol was then discarded and samples left to air dry to remove residual ethanol. For those which were solid, excess ethanol was removed and the first few steps of the extraction method were skipped.

Due to the nature of the samples collected, a modified version of Dabney et al.'s (2013) extraction protocol was used in which the initial digestion was carried out following the protocol by Gilbert et al. (2007). This digestion buffer is better suited to extraction from these tissues types than the Dabney et al. (2013) digestion buffer, which was optimised for DNA extraction from bone. Digestion of the samples was performed overnight at 56°C with rotation, using 1ml of the digestion buffer from Gilbert et al. (2007). Subsequent DNA purification and

elution was conducted following the approach described by Dabney et al. (2013).

After extraction, 20 μ L of DNA extract were built into Illumina libraries using a single-stranded library preparation protocol that has been specifically designed for the sequencing of ancient or damaged DNA (Gansauge and Meyer 2013) Libraries were prepared as originally described in (Gansauge and Meyer 2013) but without first removing deoxyuracils.

Libraries were indexed and amplified in 100 μ l PCR reactions, containing 15 μ l of aDNA library template, 10 μ l 10X PCR buffer, 10 μ l MgCl₂ (25 mM), 0.8 μ l BSA (20 mg/ml), 0.8 μ l dNTPs (25 mM), 2 μ l of each primer (10 μ M, inPE forward primer and indexed reverse primer), and 0.8 μ l AmpliTaq Gold DNA Polymerase (Applied Biosystems, Foster City, CA). Thermocycling conditions were 5 min at 95°C, followed by 16-20 cycles of 30s at 95°C, 30s at 60°C and 40s at 72°C, and a final 7 min elongation step at 72°C. The number of cycles was estimated for each sample using qPCR. Following amplification, samples were purified using Qiaquick columns (Qiagen) according to the manufacturers instructions and quantified using a 2200 TapeStation Instrument. Samples were pooled in equimolar amounts and sequenced on one lane of Illumina HiSeq2500 platform in 80bp single read chemistry mode.

Data processing

The kmer-based trimming tool BBDUK of the BBTools (Bushnell 2014) package with the kmer length set to 10 was used for low-quality and adapter trimming. PRINSEQ-lite (Schmieder and Edwards 2011) was then used to discard low complexity sequences using the DUST method with a sequence complexity score higher than 6, whereas a score of 7 and above indicating low complexity, and sequences shorter than 25bp.

To identify human DNA that might have contaminated the sample during preparation or DNA extraction, `bwa aln` (Li and Durbin 2009) with the seed function disabled (`-l 1024`) was used to align the samples to the human reference genome hg38. Non-human reads were extracted using `bedtools` (Quinlan and Hall 2010), and `bwa aln` was used with the settings as described above to align the sequences against the suitable published reference: sample 1 was aligned to the *Phalacrocorax carbo* (great cormorant) (Zhang et al. 2014), samples 2, 5, and 6 were aligned to the *Alca torda* (razorbill) (Feng et al., in review), sample 3 was aligned to the *Cephus grylle* (black guillemot) (Feng et al., in review), and sample 4 was aligned to the *Uria lomvia* (thick-billed murre) (Feng et al., in review), the closest relative to *Uria aalge* (common murre). The alignments were then authenticated as described below.

Metataxonomic assignment

MALT v0.4.1 (MEGAN Alignment Tool) (Vågene et al. 2018) was used to characterize the microbial profile of the samples. All archaeal, viral, and bacterial reference sequences that are either complete or representative were downloaded from NCBI on 06.05.2019 and indexed using `malt-build` to build a custom database. `Malt-run` was then used with minimum percent identity (`--minPercentIdentity`) set to 95, the minimum support (`--minSupport`) parameter set to 10, and the top percent value (`--topPercent`) set as 1, other parameters were set to default. The resulting `rma6` files were visualized with MEGAN6 (Huson et al. 2007), the assigned reads of the five most abundant taxa of each sample were extracted and then aligned to its respective reference genome. MALT was also used to screen for non-host metazoan DNA in the sample by using a custom database with all metazoan mitochondrial sequences obtained from NCBI on 14.05.2020.

Authentication of taxonomic assignments

A mapping quality score filter of 30 was applied to the alignment and the MarkDuplicates function of Picard-tools v2.20.2 (Broad Institute n.d.) was used to remove duplicate reads. To authenticate the alignments, mapDamage 2.0.4 (Jónsson et al. 2013) was used to estimate the post-mortem deamination rates and bedtools was then used to calculate the breadth and average depth of coverage. To assess the evenness of coverage, the average depth of coverage of the microbial alignments was calculated in 1000 bp windows and then visualized using Circos v0.69-6 (Krzywinski et al. 2009). The edit distance distribution was obtained using samtools, visualized with R, and the distributions were then used to calculate the negative difference proportion (Huebler et al. 2019). Only taxa with a negative difference proportion larger than 0.9, with cytosine deamination, and an even coverage were regarded as being present in the samples.

Results

Sequencing output

Table 2. List of generated sequences per sample after quality filtering

| Sample ID | Raw reads | Filtered reads | Fragment length | hg38 reads | Bird reads |
|-----------|------------|----------------|-----------------|------------|------------|
| 1-I | 33,170,984 | 16,869,767 | 44 | 456,714 | 42,281 |
| 1-S | 45,972,838 | 28,615,754 | 52 | 992,508 | 23,293 |
| 2-I | 27,385,338 | 7,170,718 | 39 | 220,922 | 228,691 |
| 2-S | 26,785,906 | 4,667,828 | 36 | 177,437 | 29,310 |
| 3-I | 30,755,048 | 15,703,757 | 43 | 684,585 | 29,880 |
| 3-S | 27,628,076 | 7,996,017 | 35 | 1,354,307 | 70,644 |
| 4-I | 41,269,956 | 18,660,685 | 41 | 649,716 | 48,826 |
| 4-S | 43,241,500 | 6,262,800 | 38 | 621,968 | 34,043 |
| 5-I | 33,399,220 | 17,877,997 | 47 | 3,479,775 | 65,311 |
| 5-S | 41,694,030 | 2,507,557 | 44 | 207,214 | 34,785 |
| 6-I | 29,306,390 | 19,806,778 | 66 | 112,725 | 16,545,114 |
| 6-S | 35,131,018 | 24,252,592 | 62 | 140,489 | 17,581,152 |

We generated a total of 415,740,304 paired-end reads, with an average of 34.6 million reads per sample (Table 2). However, for most samples a large proportion of the reads (between 30-95%, library dependent) had to be discarded during quality control, as they were either adapter-dimers or of very short read length (<25bp). This high level likely indicates that the DNA in the samples had undergone considerable degradation in the time since, a feature which is also suggested by the deamination profiles, which are unusually high for historic samples (Fig 1b).

Endogenous host DNA

Between 0.1% and 83.5% of the sequences of each sample could be aligned to the reference genome of the corresponding seabird species (Figure 1a). Out of the six samples, the razorbill that was collected 1916 (Sample 6-I and 6-S) had the largest proportion of endogenous DNA. After removing duplicates and applying a mapping quality filter of 30, 29.4 million sequences (45.6% of the total number of reads) aligned to the razorbill genome, covering 56% of the genome with an mean depth of coverage of 1.6X. Its edit distance distribution is strictly declining, which is expected when the sequences are aligned to the correct reference genome. The only other two samples with a strictly declining edit distance distribution are sample 2-I and 5-S (both razorbills from 1884 and 1908, Figure 1b). We also observed significant cytosine deamination at the first read positions for sample 2-I, 5-S, 6-I, and 6-S (Figure 1c). The three other individuals did not exhibit a strictly declining edit distance in either of their two samples, which indicates we did not recover significant amounts of endogenous DNA from the great cormorant, black guillemot, or common murre samples.

Human DNA

The proportion of sequences aligning to the human reference genome ranged between 0.2 and 14.8% per sample. The cytosine deamination pattern for each of the samples was insignificant, indicating that the contamination did not occur during the collection of the specimens (Supplementary Table S2). The alignments of sample 2 and sample 5 did not exhibit a decreasing edit distance distribution and had a short mean read length and are therefore likely false positive assignments. The mean depth of coverage was too low to determine whether the human DNA of the remaining samples originates from the same individual.

Non-host metazoan DNA

Aside from sample 3-I, none of the samples had a significant number of sequences that were assigned to non-avian mitochondrial reference genomes. 538 sequences from sample 2-I were assigned to the mitochondrial genome of the *Gasterosteus aculeatus* (three-spined stickleback). All non-human and non-avian sequences of sample 2-I were subsequently aligned to the reference genome of the three-spined stickleback, resulting in 259,729 sequences with a mapping quality above 30 aligning to the genome. The alignment to the nuclear reference genome the three-spined stickleback genome showed significant cytosine deamination patterns (Supplementary Fig. S3) and a strictly decreasing edit distance distribution (Supplementary Fig. S4).

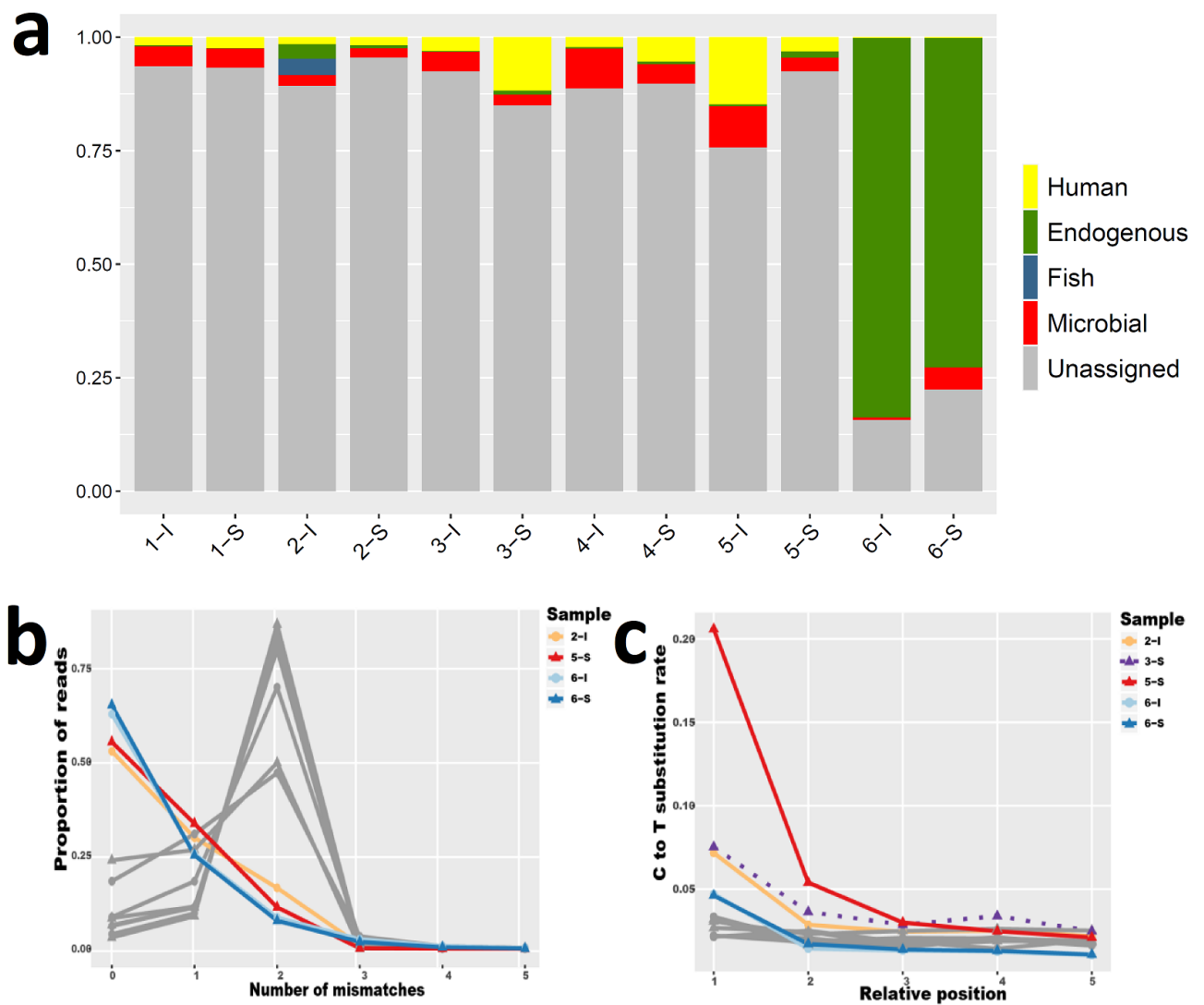


Fig. 1. Endogenous host DNA. **a)** Proportion of reads assigned to human, host, and microbial reference genomes per sample. **b)** Edit distance distribution of the reads aligning the host reference genome. **c)** Cytosine deamination observed at the first five positions of the reads aligning to the host reference genome.

Microbial DNA

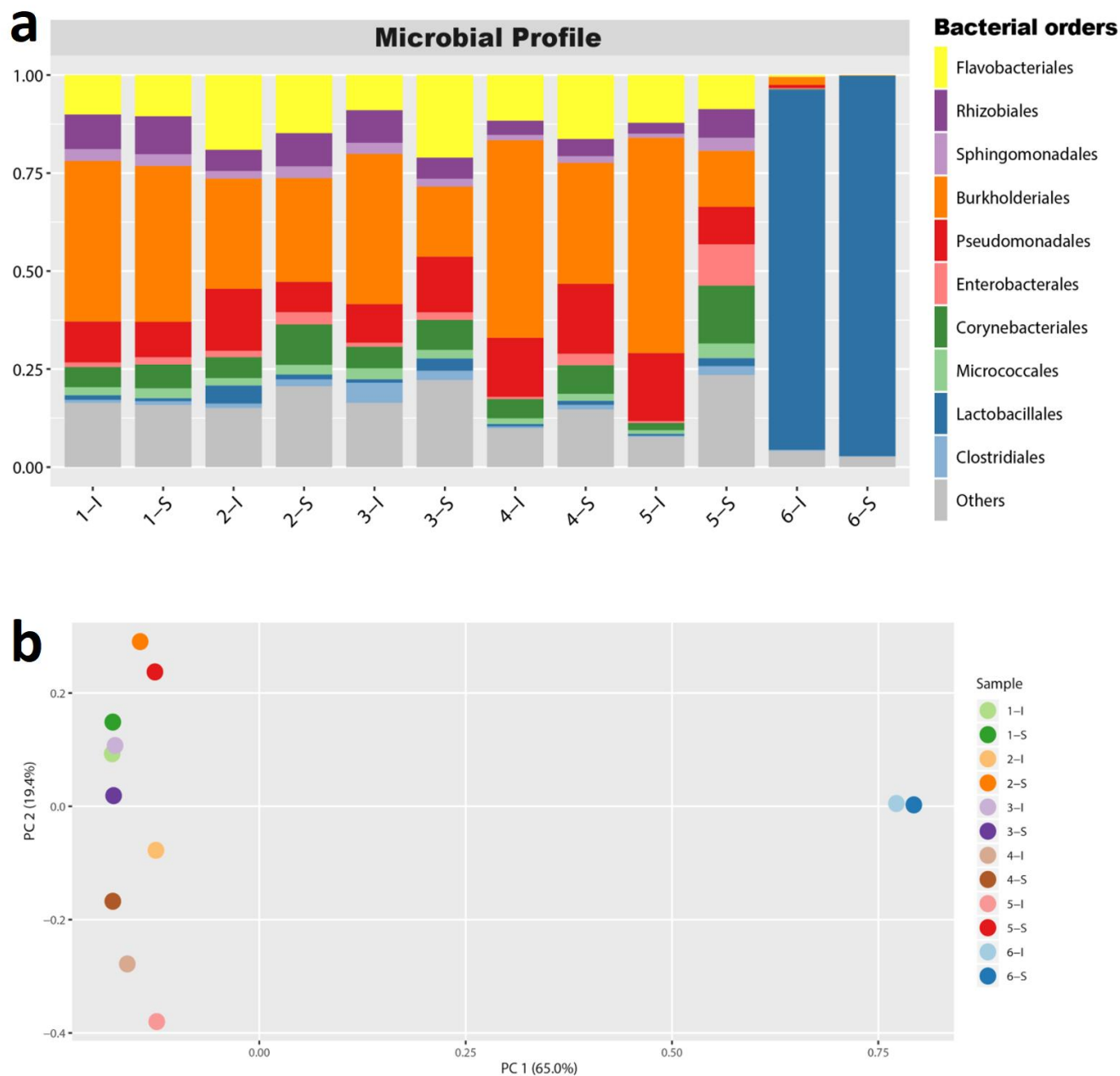


Fig. 2. Microbial profile. a) Microbial composition on order level for the 12 samples. b) PCoA using Bray-Curtis on species level for the 12 samples.

A total of ~8 million out of the 170 million non-human and non-endogenous sequences of all samples (between 0.6-9.2% per sample) could be assigned to microbial, archaeal, and viral taxa (Figure 2a). A PCoA on species level revealed two major clusters of samples separating sample 6-I and 6-S from the remaining samples (Fig. 2b).

This is reflected in a large overlap of the most abundant taxa assigned to sample 1 to 5, with the wastewater bacterium *Cloacibacterium normanense* (Allen et al. 2006) one of the three most abundant taxa in nine samples (part of the order Flavobacteriales in Figure 2a). The majority of taxa in samples 1 to 5 are not associated with gut microbiomes, suggesting that they are not derived from the microbiome of the samples, but are either contaminants or spurious assignments.

In contrast to samples 1 to 5, the microbial profiles of sample 6-I and 6-S are heavily dominated by *Catelicoccus marimammalium*, with 85% and 95% of the assigned microbial sequences, respectively. After removing PCR duplicates and applying a mapping quality filter of 30, we observed an average depth of coverage of 5.2X and 35.5X and a breadth of coverage of 78.4% and 89.9%, respectively (Fig. 3). *C. marimammalium* was also detected in high abundance in sample 2-I where it made up 3.3% of the microbial sequences. *C. marimammalium* is an avian gut bacterium that has been previously proposed as a suitable biomolecular indicator to monitor gull fecal contamination of recreational waters due to its prevalence and abundance in seabird and waterfowl feces.

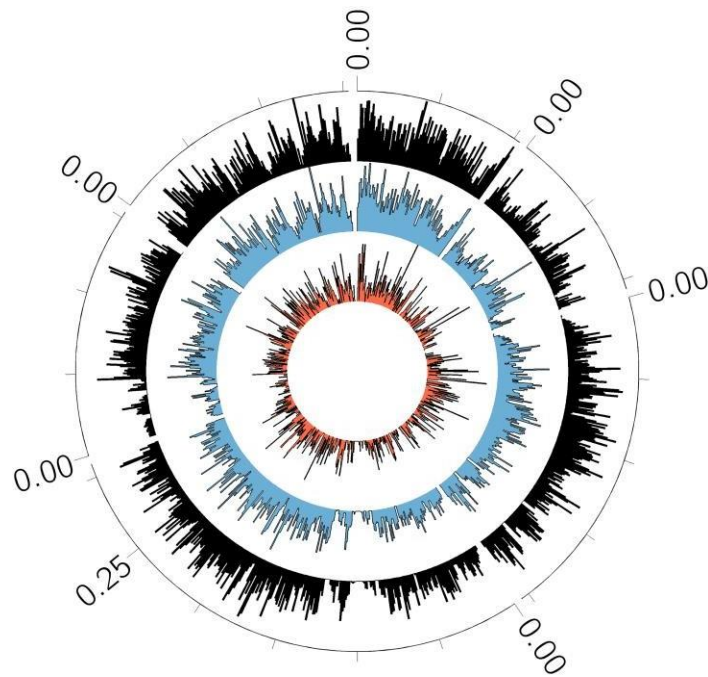


Fig. 3. Coverage plot for *Catellicoccus marimammalium* detected in sample 2-I (red), 6-I (blue), and 6-S (black).

The remaining most abundant taxa that we detected in sample 6-I and 6-S are the piscine pathogens *Aeromonas salmonicida*, *Aliivibrio salmonicida*, *Carnobacterium maltaromaticum*, and *Lactobacillus fuchuensis*, as well as *Clostridium frigidicarnis* that has previously been recovered from spoiled meat (Broda et al. 1999). Among the 11 most abundant taxa identified in sample 6-S are also three *Enterococcus* species - *E. faecium*, *E. faecalis*, and *E. columbae* - that are all associated with the gastrointestinal tract, with *E. columbae* being part of the intestinal flora of pigeons (Devriese et al. 1990). No archaeal taxa were identified in high abundance in any of the samples, and none of the viral taxa passed the evenness of coverage criterion.

Discussion

Our results show that it is, in principal, possible to recover genetic signatures of gut microbiomes from seabird specimens which had been collected and preserved in fluid up to 100 years ago. However, just as importantly, our data shows clearly that half of the specimens tested did not yield unambiguous traces of endogenous or gut metagenomic DNA. In fact, only one of the six specimens yielded a large proportion of both host and gut microbiome DNA, while another yielded host and dietary DNA.

Several features of the shotgun sequencing data highlight the challenging nature of recovering and analysing DNA from fluid-preserved specimens. The DNA fragments sequenced were very short (between 25-58bp on average per sample after removing adapter sequences), and frequently contained high proportions of adapter dimers (up to 56%) (Table 2). Highly fragmented DNA is often observed by researchers studying historic and ancient samples, and it is clear that the fluid-preservation methods applied to the samples studied here must have either caused DNA fragmentation through one of many possible pathways (Lindahl 1993) or even cross-linking of longer DNA molecules, thereby making them inaccessible to extraction methods. The high levels of adapter-dimers seen is also often seen in other aDNA datasets (e.g. Carøe, Gopalakrishnan, and Vinner 2018), reflecting the dual challenges of working with small amounts of short molecules, that firstly may require many cycles of library amplification until they reach a concentration that can be sequenced (exasperating the level of dimers) and secondly the challenge of purifying away such dimers without losing the insert-containing libraries. In many cases, this problem can be ameliorated through careful titration of adapter molarity in library preparation, but when very few template DNA molecules are isolated from a specimen, adapter-dimers are nearly inevitable.

To improve the feasibility of fluid-preserved microbiome analyses, further studies are necessary to determine what effect different preservatives have on the preservation of microbial DNA in order to a-priori select specimens with a higher likelihood of yielding a sufficient amount of well-preserved DNA. This sample set so far has failed to provide clear evidence for visual markers that correspond with microbial DNA preservation in fluid. While the colouration of eye lenses has been suggested as an indicator of the fixation in formaldehyde (Simmons 2014), our inspection of the eyes of the specimens yielded only inconclusive results (Supplementary Figure 2). Further metagenomic screening of fluid-preserved museum specimens are necessary to determine if collection date, the appearance of the specimen, or other factors provide hints on DNA preservation.

Our data also show high levels of contamination by non-endogenous microbes in most of the samples. As we observed strictly declining edit distance distributions, evenness of coverage, and post-mortem deamination patterns for the aforementioned microbes, it is unlikely that they were misassigned. Understanding the origins of these microbial contaminants could have ramifications for future metagenomic studies of museum collections and potentially curation practices. As discussed below, we considered several hypotheses for how microbial contamination could have occurred, including whether the specimens began rotting prior to immersion in the preservatives, if curators inadvertently transferred microorganisms during the preparation of the samples, if unsterilized collection jars might harbor these communities, or if fluid preservatives may be contaminated, particularly through the re-use of alcohol from other fluid-preserved specimens (Hawks 2003).

For these samples, rotting of the specimens appears to be an insignificant factor, as none of the most abundant contaminant taxa are associated with the decomposition of carcasses. Furthermore, the microbial profiles across samples 1 to 5 are strikingly similar, indicating that

the contamination originates from the same source, such as reagents that were used to produce the preservation fluid, with which the jars were regularly refilled.

One of the most telling features of the contaminant DNA is that it bears the same hallmarks of degraded DNA, with cytosine deamination and high levels of fragmentation (Pääbo 1989). The similar degradation patterns of the contaminant microbial DNA and endogenous DNA suggests the contaminants are not due to sampling the specimens for this project or from unavoidable contaminants in laboratory reagents (e.g. Salter 2014). Thus, it appears the contaminant DNA was present in the collection jars along with the historic specimens for an indeterminate amount of time. It is important to note that indistinguishable degradation signals complicates the recognition of authentic endogenous microbial taxa, and future studies should take appropriate precautions such as sampling and sequencing the preservation liquid in order to differentiate between contaminants and microbial taxa of interest.

The contamination could have occurred during the preparation of the samples, due to rotting of the material prior to immersion in the preservatives, the use of contaminated batches of fluid preservatives, or even the re-use of alcohol from other fluid-preserved specimens (Hawks 2003). One of the most critical aspects of this contaminant DNA is that it bears the same hallmarks of degraded DNA, with cytosine deamination and high levels of fragmentation (Pääbo 1989). This degradation signal complicates the recognition of authentic endogenous microbial taxa, and future studies should take appropriate precautions such as sampling and sequencing the preservation liquid in order to differentiate between contaminants and microbial taxa of interest.

One curious observation in the microbial dataset is that the most abundant taxa are gram-negative bacteria. Presently it is unclear if the high proportion of these species is an accurate depiction of the microorganisms present in the specimens, or if it instead is a bias

caused by the tendency of gram-negative cell walls to lyse in ethanol (Jones 1989). Given that even low concentrations of ethanol inhibits microbial growth (e.g. Fletcher 1983), it is unlikely that the metagenomic signals originate from taxa that were biologically active in the preservation fluid.

In the metagenomic analysis we observed a large proportion of unassigned reads, ranging from 15-93% per sample. Aside from the fact that very short fragments can often not be confidently assigned to the correct source reference genome, the number of unassigned reads can also be attributed to the fact that microbiomes of non-model organisms are often complex but not well described. This underrepresentation in public reference genome databases can lead to spurious taxonomic assignments and a large proportion of unassigned sequences. Due to the growing interest in metagenomes in recent years and thus an increasing rate of microbial genome sequencing, this issue should become less problematic in coming years. In future projects, researchers may consider exploring assembly-based approaches to reduce the proportion of unassigned reads. We were unable to use this approach in this study due to the very short fragment lengths and low depth of coverage; however, with sufficient amounts of sequencing data, it may be possible to characterize a more complete microbial profile in historic specimens.

While this project has focused on understanding the preservation of historic microbiomes in fluid-preserved collection jars, it is important to highlight that such metagenomic experiments often produce novel data that is useful for other researchers. For example, in the process of generating sequencing data to characterize the gut microbiome of an Atlantic razorbill collected in 1916, we simultaneously recovered a substantial amount of DNA from the host, amounting to an average depth of coverage of 1.6X on the nuclear genome. To our knowledge there are no whole genome sequencing projects of the razorbill aside from

the generation of the reference genome, thus our data from an historical specimen could provide valuable information to future research on this understudied seabird. The three-spined stickleback DNA identified in sample 2-I furthermore indicates that it is possible to retrieve dietary information from fluid-preserved specimens. Razorbills and other members from the Alcidae family have previously been observed to feed on three-spined sticklebacks that can be found in coastal regions as well as fresh water (Huettmann et al. 2005; Lance and Thompson 2005; Olson et al. 1979).

In summary, this project demonstrates that fluid-preserved specimens have the potential to represent a novel substrate with which to study historical microbial communities, although they must be approached with measured optimism. We were for example able to recover the genomes of multiple gut microbes of a fluid-preserved razorbill that was collected in 1916, as well as the nuclear genome of the host bird to average depth of coverage of 1.6X. While we observed that the DNA recovered from most of the specimens we analyzed was heavily degraded, future studies should be able to improve the viability of this material, facilitating the analysis of so far overlooked microbial communities.

Acknowledgements

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 676154 (ArchSci2020).

References

- Allen, Toby D., Paul A. Lawson, Matthew D. Collins, Enevold Falsen, and Ralph S. Tanner. 2006. "Cloacibacterium Normanense Gen. Nov., Sp. Nov., a Novel Bacterium in the Family Flavobacteriaceae Isolated from Municipal Wastewater." *International Journal of Systematic and Evolutionary Microbiology* 56 (Pt 6): 1311–16.
- Bi, Ke, Tyler Linderoth, Dan Vanderpool, Jeffrey M. Good, Rasmus Nielsen, and Craig Moritz. 2013. "Unlocking the Vault: Next-Generation Museum Population Genomics." *Molecular Ecology* 22 (24): 6018–32.
- Bos, Kirsten I., Verena J. Schuenemann, G. Brian Golding, Hernán A. Burbano, Nicholas Waglechner, Brian K. Coombes, Joseph B. McPhee, et al. 2011. "A Draft Genome of *Yersinia Pestis* from Victims of the Black Death." *Nature* 478 (7370): 506–10.
- Broad Institute. n.d. "Picard Tools." Picard Tools. Accessed November 14, 2019. <http://broadinstitute.github.io/picard/>.
- Broda, D. M., P. A. Lawson, R. G. Bell, and D. R. Musgrave. 1999. "Clostridium Frigidicarnis Sp. Nov., a Psychrotolerant Bacterium Associated with 'blown Pack' spoilage of Vacuum-Packed Meats." *International Journal of Systematic and Evolutionary Microbiology* 49 (4): 1539–50.
- Bushnell, B. 2014. "BBTools Software Package." URL <Http://sourceforge.Net/projects/bbmap>.
- Carøe, C., S. Gopalakrishnan, and L. Vinner. 2018. "Single-tube Library Preparation for Degraded DNA." *Methods in Ecology and Evolution / British Ecological Society*. <https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12871>.
- Dabney, Jesse, Michael Knapp, Isabelle Glocke, Marie-Theres Gansauge, Antje Weihmann, Birgit Nickel, Cristina Valdiosera, et al. 2013. "Complete Mitochondrial Genome Sequence of a Middle Pleistocene Cave Bear Reconstructed from Ultrashort DNA Fragments." *Proceedings of the National Academy of Sciences of the United States of America* 110 (39): 15758–63.
- Devault, Alison M., G. Brian Golding, Nicholas Waglechner, Jacob M. Enk, Melanie Kuch, Joseph H. Tien, Mang Shi, et al. 2014. "Second-Pandemic Strain of *Vibrio Cholerae* from the Philadelphia Cholera Outbreak of 1849." *The New England Journal of Medicine* 370 (4): 334–40.
- Devriese, L. A., K. Ceysens, U. M. Rodrigues, and M. D. Collins. 1990. "Enterococcus Columbae, a Species from Pigeon Intestines." *FEMS Microbiology Letters* 59 (3): 247–51.
- Down, Rosina. 1989. "'Old' Preservative Methods." In *Conservation of Natural History Specimens. Spirit Collections*, 33–38. bcin.ca.
- Fletcher, Madilyn. 1983. "The Effects of Methanol, Ethanol, Propanol and Butanol on Bacterial Attachment to Surfaces." *Microbiology* 129 (3): 633–41.
- Gansauge, Marie-Theres, and Matthias Meyer. 2013. "Single-Stranded DNA Library Preparation for the Sequencing of Ancient or Damaged DNA." *Nature Protocols* 8 (4): 737–48.
- Green, Eleanor Joan, and Camilla F. Speller. 2017. "Novel Substrates as Sources of Ancient DNA: Prospects and Hurdles." *Genes* 8 (7). <https://doi.org/10.3390/genes8070180>.
- Hawks, Catharine. 2003. "Re-Use Of Ethanol In Processing Biological Specimens." *Conserve O Gram*, no. 11/5 (June): 1–4.
- Huebler, Ron, Felix M. M. Key, Christina Warinner, Kirsten I. Bos, Johannes Krause, and Alexander Herbig. 2019. "HOPS: Automated Detection and Authentication of Pathogen DNA in Archaeological Remains." *bioRxiv*. <https://doi.org/10.1101/534198>.
- Huettmann, Falk, Antony W. Diamond, Brian Dalzell, and Ken Macintosh. 2005. "Winter Distribution, Ecology and Movements of Razorbills *Alca Torda* and Other Auks in the Outer Bay of Fundy, Eastern Canada." *Marine Ornithology* 33: 161–71.

- Hühns, Maja, Andreas Erbersdobler, Annette Obliers, and Paula Röpenack. 2017. "Identification of HPV Types and Mycobacterium Tuberculosis Complex in Historical Long-Term Preserved Formalin Fixed Tissues in Different Human Organs." *PloS One* 12 (1): e0170353.
- Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. "MEGAN Analysis of Metagenomic Data." *Genome Research* 17 (3): 377–86.
- Iñiguez, Alena Mayo, Karl Reinhard, Marcelo Luiz Carvalho Gonçalves, Luiz Fernando Ferreira, Adauto Araújo, and Ana Carolina Paulo Vicente. 2006. "SL1 RNA Gene Recovery from *Enterobius Vermicularis* Ancient DNA in Pre-Columbian Human Coprolites." *International Journal for Parasitology* 36 (13): 1419–25.
- Jones, Rodney P. 1989. "Biological Principles for the Effects of Ethanol." *Enzyme and Microbial Technology* 11 (3): 130–53.
- Jónsson, Hákon, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, and Ludovic Orlando. 2013. "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters." *Bioinformatics* 29 (13): 1682–84.
- Krzywinski, Martin, Jacqueline Schein, Inanç Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. 2009. "Circos: An Information Aesthetic for Comparative Genomics." *Genome Research* 19 (9): 1639–45.
- Lance, Monique M., and Christopher W. Thompson. 2005. "Overlap in Diets and Foraging of Common Murres (*Uria Aalge*) and Rhinoceros Auklets (*Cerorhinca Monocerata*) After the Breeding Season." *The Auk* 122 (3): 887–901.
- Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
- Lindahl, T. 1993. "Instability and Decay of the Primary Structure of DNA." *Nature* 362 (6422): 709–15.
- Lugli, Gabriele Andrea, Christian Milani, Leonardo Mancabelli, Francesca Turrone, Chiara Ferrario, Sabrina Duranti, Douwe van Sinderen, and Marco Ventura. 2017. "Erratum to: Ancient Bacteria of the Ötzi's Microbiome: A Genomic Tale from the Copper Age." *Microbiome* 5 (1): 23.
- Miller, Jeremy A., Kevin K. Beentjes, Peter van Helsdingen, and Steven Ijland. 2013. "Which Specimens from a Museum Collection Will Yield DNA Barcodes? A Time Series Study of Spiders in Alcohol." *ZooKeys*, no. 365 (December): 245–61.
- Noonan, James P., Graham Coop, Sridhar Kudaravalli, Doug Smith, Johannes Krause, Joe Alessi, Feng Chen, et al. 2006. "Sequencing and Analysis of Neanderthal Genomic DNA." *Science* 314 (5802): 1113–18.
- Olson, S. L., C. C. Swift, and C. Mokhiber. 1979. "An Attempt to Determine the Prey of the Great Auk (*Pinguinus Impennis*)." *The Auk*.
https://www.jstor.org/stable/4085666?casa_token=_XsiWt3SfSQAAAAA:Oozw2zpDF-DiqgBxlhEj1EqiA_x0NoolCt8qJsCSrCW65GhAytYKi80vy_NdH3psim9R82PuH0TpwUbJrF7nhLGMINAM3T3rRX9kEupNBLFWnfHZFxc.
- Poinar, H. N., M. Kuch, K. D. Sobolik, I. Barnes, A. B. Stankiewicz, T. Kuder, W. G. Spaulding, V. M. Bryant, A. Cooper, and S. Pääbo. 2001. "A Molecular Analysis of Dietary Diversity for Three Archaic Native Americans." *Proceedings of the National Academy of Sciences of the United States of America* 98 (8): 4317–22.
- Preus, Hans R., Ole J. Marvik, Knut A. Selvig, and Pia Bennike. 2011. "Ancient Bacterial DNA (aDNA) in Dental Calculus from Archaeological Human Remains." *Journal of Archaeological Science* 38 (8): 1827–31.
- Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
- Salter, Susannah J., Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. 2014. "Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses." *BMC Biology* 12 (November): 87.

- Schmieder, Robert, and Robert Edwards. 2011. "Quality Control and Preprocessing of Metagenomic Datasets." *Bioinformatics* 27 (6): 863–64.
- Shokralla, Shadi, Gregory A. C. Singer, and Mehrdad Hajibabaei. 2010. "Direct PCR Amplification and Sequencing of Specimens' DNA from Preservative Ethanol." *BioTechniques* 48 (3): 233–34.
- Simmons, John E. 2014. *Fluid Preservation: A Comprehensive Reference*. Rowman & Littlefield.
- Srinivasan, Mythily, Daniel Sedmak, and Scott Jewell. 2002. "Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids." *The American Journal of Pathology* 161 (6): 1961–71.
- Vågene, Åshild J., Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, et al. 2018. "Salmonella Enterica Genomes from Victims of a Major Sixteenth-Century Epidemic in Mexico." *Nature Ecology & Evolution* 2 (3): 520–28.
- Warinner, Christina, João F. Matias Rodrigues, Rounak Vyas, Christian Trachsel, Natallia Shved, Jonas Grossmann, Anita Radini, et al. 2014. "Pathogens and Host Immunity in the Ancient Human Oral Cavity." *Nature Genetics* 46 (4): 336–44.
- Warinner, Christina, Camilla Speller, and Matthew J. Collins. 2015. "A New Era in Palaeomicrobiology: Prospects for Ancient Dental Calculus as a Long-Term Record of the Human Oral Microbiome." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 370 (1660): 20130376.
- Weyrich, Laura S., Sebastian Duchene, Julien Soubrier, Luis Arriola, Bastien Llamas, James Breen, Alan G. Morris, et al. 2017. "Neanderthal Behaviour, Diet, and Disease Inferred from Ancient DNA in Dental Calculus." *Nature* 544 (7650): 357–61.
- Zhang, Guojie, Bo Li, Cai Li, M. Thomas P. Gilbert, Erich D. Jarvis, Jun Wang, and Avian Genome Consortium. 2014. "Comparative Genomic Data of the Avian Phylogenomics Project." *GigaScience* 3 (1): 26..

Supplementary materials



Fig S1. Picture of the five jars containing the six specimens.

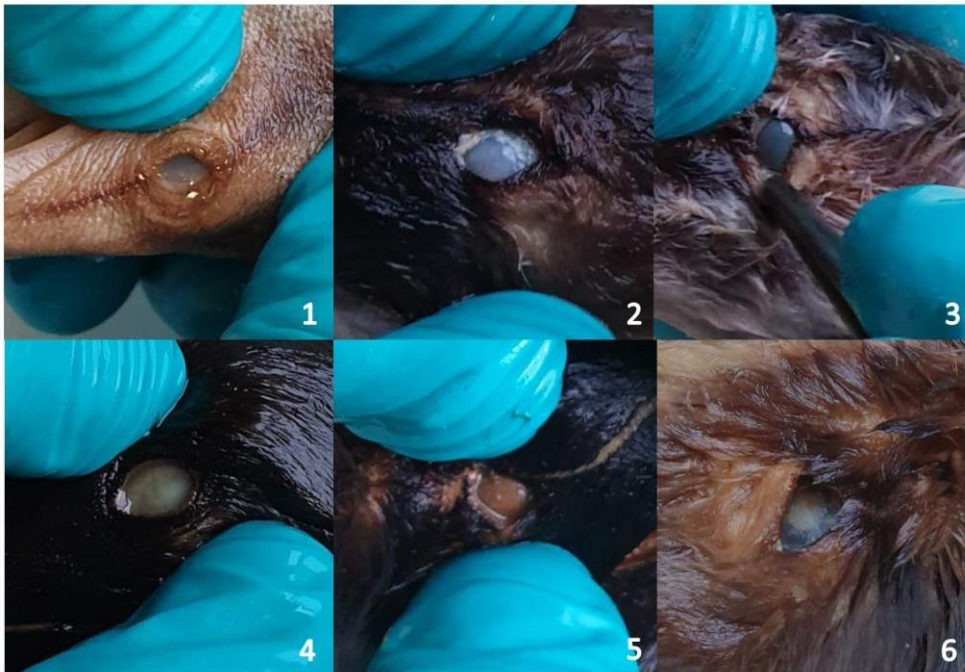


Fig S2. Pictures of the six specimens' eyes. White colouration of the eyes is associated with the fixation in a fluid that is not formaldehyde.

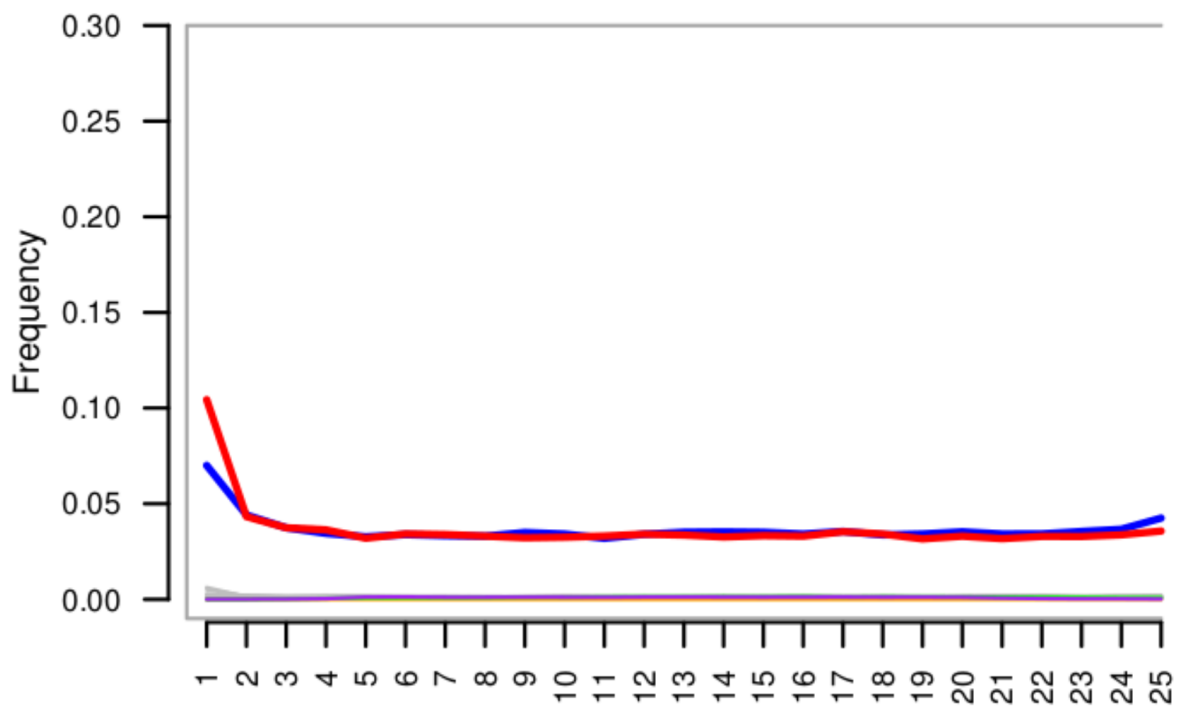


Fig S3. Cytosine deamination pattern of the *Gasterosteus aculeatus* DNA recovered from sample 2-l.

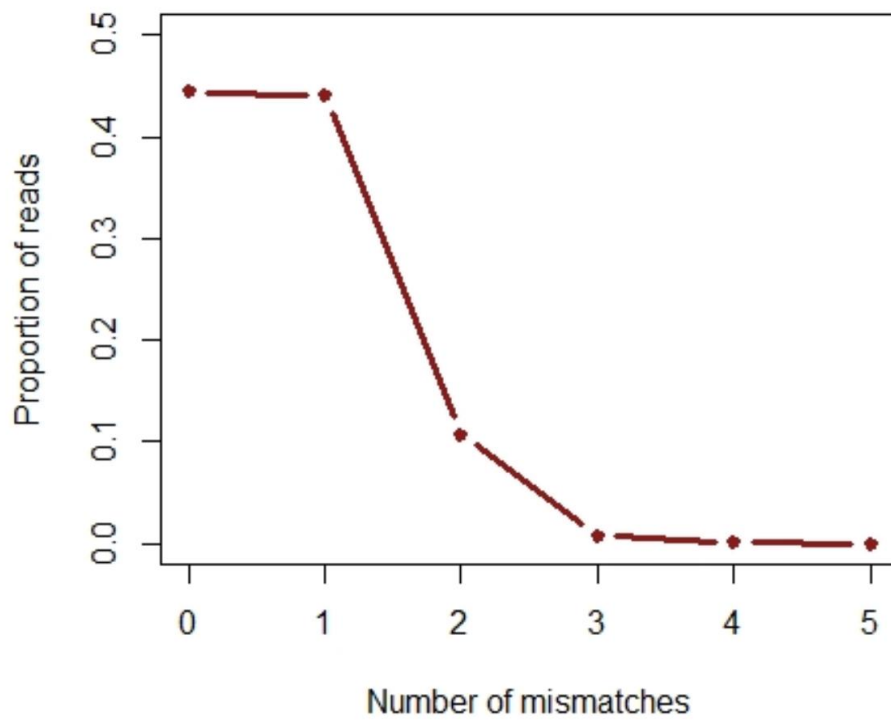


Fig. S4. Edit distance distribution of the *Gasterosteus aculeatus* DNA recovered from sample 2-l.

Supplementary tables

Table S1. Species alignment

| Sample | Species name | Common name | #Sequences | Avg. read length | Breadth of coverage | Avg. depth of coverage | δs | - $\Delta\%$ |
|--------|----------------------------|------------------------|------------|------------------|---------------------|------------------------|------------|--------------|
| 1-I | <i>Phalacrocorax carbo</i> | Great cormorant | 23,095 | 32 | 0.00 | 0.00 | 0.10 | 0.53 |
| 1-S | <i>Phalacrocorax carbo</i> | Great cormorant | 12,242 | 28 | 0.00 | 0.00 | 0.47 | 0.51 |
| 2-I | <i>Alca torda</i> | Razorbill | 172,657 | 43 | 0.00 | 0.01 | 0.38 | 1.00 |
| 2-S | <i>Alca torda</i> | Razorbill | 16,257 | 26 | 0.00 | 0.00 | 0.12 | 0.52 |
| 3-I | <i>Cephus grylle</i> | Black guillemot | 16,815 | 27 | 0.00 | 0.00 | 0.13 | 0.51 |
| 3-S | <i>Cephus grylle</i> | Black guillemot | 42,279 | 27 | 0.00 | 0.00 | 0.56 | 0.66 |
| 4-I | <i>Uria aalge</i> | Common murre/guillemot | 27,759 | 28 | 0.00 | 0.00 | 0.06 | 0.52 |
| 4-S | <i>Uria aalge</i> | Common murre/guillemot | 17,612 | 27 | 0.00 | 0.00 | 0.26 | 0.53 |
| 5-I | <i>Alca torda</i> | Razorbill | 36,220 | 36 | 0.00 | 0.00 | 0.62 | 0.62 |
| 5-S | <i>Alca torda</i> | Razorbill | 21,961 | 32 | 0.00 | 0.00 | 1.00 | 1.00 |
| 6-I | <i>Alca torda</i> | Razorbill | 14,487,714 | 66 | 0.38 | 0.82 | 0.14 | 1.00 |
| 6-S | <i>Alca torda</i> | Razorbill | 14,923,770 | 60 | 0.35 | 0.77 | 0.17 | 1.00 |

Table S3. Alignments to the human reference genome

| | Reads | Fragment length (bp) | C-T 5'/(%) | - $\Delta\%$ |
|-----|-----------|----------------------|------------|--------------|
| 1-I | 313,047 | 43 | 0.01 | 1.00 |
| 1-S | 701,149 | 60 | 0.01 | 1.00 |
| 2-I | 113,477 | 36 | 0.02 | 0.71 |
| 2-S | 84,551 | 34 | 0.02 | 0.71 |
| 3-I | 488,568 | 46 | 0.01 | 1.00 |
| 3-S | 947,089 | 40 | 0.01 | 1.00 |
| 4-I | 420,518 | 42 | 0.01 | 1.00 |
| 4-S | 342,173 | 48 | 0.01 | 1.00 |
| 5-I | 2,652,760 | 52 | 0.01 | 1.00 |
| 5-S | 80,186 | 42 | 0.01 | 1.00 |
| 6-I | 48,135 | 40 | 0.04 | 0.55 |
| 6-S | 60,503 | 38 | 0.04 | 0.55 |

Table S4. Five most abundant taxa for each sample

| Sample | Taxon | RefSeq assembly accession | #Sequences | Avg. read length | Breadth of coverage | Avg. depth of coverage | δs | $-\Delta\%$ |
|--------|---|---------------------------|------------|------------------|---------------------|------------------------|------------|-------------|
| 1-1 | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 22,627 | 60 | 24.31% | 0.50 | 0.21 | 1.00 |
| 1-1 | <i>Acinetobacter junii</i> | GCF_001941805.1 | 21,548 | 58 | 19.64% | 0.38 | 0.10 | 1.00 |
| 1-1 | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 11,265 | 48 | 7.57% | 0.17 | 0.46 | 0.72 |
| 1-1 | <i>Acidovorax</i> sp. KKS102 | GCF_000302535.1 | 9,481 | 50 | 4.92% | 0.09 | 0.19 | 0.78 |
| 1-1 | <i>Cutibacterium acnes</i> | GCF_000008345.1 | 7,782 | 48 | 8.47% | 0.15 | 0.48 | 1.00 |
| 1-S | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 25,221 | 55 | 13.47% | 0.44 | 0.44 | 0.70 |
| 1-S | <i>Acinetobacter junii</i> | GCF_001941805.1 | 22,763 | 65 | 22.62% | 0.44 | 0.20 | 1.00 |
| 1-S | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 21,111 | 64 | 24.50% | 0.50 | 0.19 | 0.94 |
| 1-S | <i>Acidovorax</i> sp. KKS102 | GCF_000302535.1 | 16,223 | 57 | 8.81% | 0.18 | 0.17 | 0.74 |
| 1-S | <i>Acinetobacter</i> sp. WCHA45 | GCF_002165255.2 | 14,817 | 60 | 14.74% | 0.29 | 0.35 | 1.00 |
| 2-1 | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 14,014 | 50 | 13.43% | 0.26 | 0.22 | 1.00 |
| 2-1 | <i>Acinetobacter junii</i> | GCF_001941805.1 | 10,376 | 51 | 8.71% | 0.16 | 0.37 | 1.00 |
| 2-1 | <i>Catellibacoccus marimamailum</i> | GCF_000313915.1 | 5,413 | 66 | 14.68% | 0.28 | 0.24 | 1.00 |
| 2-1 | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 3,777 | 44 | 2.83% | 0.05 | 0.11 | 0.75 |
| 2-1 | <i>Acidovorax</i> sp. JS42 | GCF_000015545.1 | 1,743 | 53 | 1.17% | 0.02 | 0.51 | 1.00 |
| 2-S | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 3,403 | 41 | 2.39% | 0.04 | 0.20 | 0.83 |
| 2-S | <i>Mycobacterium peregrinum</i> | GCF_001403655.1 | 1,923 | 39 | 0.62% | 0.01 | 0.32 | 1.00 |
| 2-S | <i>Escherichia coli</i> | GCF_000008865.2 | 505 | 46 | 0.24% | 0.00 | 0.24 | 1.00 |
| 2-S | <i>Acinetobacter</i> sp. WCHA45 | GCF_002165255.2 | 1,456 | 44 | 1.17% | 0.02 | 0.23 | 1.00 |
| 2-S | <i>Paraburkholderia aromaticivorans</i> | GCF_002278075.1 | 821 | 42 | 0.19% | 0.00 | 0.24 | 0.86 |

| | | | | | | | | |
|-----|---|-----------------|---------|----|--------|------|------|------|
| 3-1 | <i>Paraclostridium bifermentans</i> | GCF_000452245.2 | 17,063 | 40 | 9.28% | 0.19 | 0.67 | 1.00 |
| 3-1 | <i>Acinetobacter junii</i> | GCF_001941805.1 | 15,196 | 53 | 12.80% | 0.24 | 0.08 | 1.00 |
| 3-1 | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 15,744 | 53 | 15.72% | 0.31 | 0.23 | 1.00 |
| 3-1 | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 9,883 | 48 | 6.76% | 0.15 | 0.39 | 0.72 |
| 3-1 | <i>Acidovorax</i> sp. KKS102 | GCF_000302535.1 | 8,836 | 50 | 4.63% | 0.09 | 0.19 | 0.78 |
| 3-S | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 14,253 | 39 | 9.92% | 0.21 | 0.22 | 1.00 |
| 3-S | <i>Acinetobacter junii</i> | GCF_001941805.1 | 7,334 | 38 | 4.30% | 0.08 | 0.42 | 1.00 |
| 3-S | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 4,752 | 39 | 3.02% | 0.06 | 0.37 | 0.81 |
| 3-S | <i>Mycolicibacterium peregrinum</i> | GCF_001403655.1 | 2,464 | 35 | 0.71% | 0.01 | 0.53 | 1.00 |
| 3-S | <i>Cutibacterium acnes</i> | GCF_000008345.1 | 2,284 | 38 | 1.97% | 0.03 | 0.47 | 1.00 |
| 4-1 | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 107,653 | 55 | 53.52% | 2.17 | 0.17 | 1.00 |
| 4-1 | <i>Acinetobacter junii</i> | GCF_001941805.1 | 104,314 | 55 | 51.69% | 1.73 | 0.18 | 1.00 |
| 4-1 | <i>Acidovorax</i> sp. JS42 | GCF_000015545.1 | 61,073 | 57 | 20.53% | 0.76 | 0.70 | 1.00 |
| 4-1 | <i>Acidovorax ebreus</i> | GCF_000022305.1 | 35,538 | 58 | 16.39% | 0.55 | 0.69 | 1.00 |
| 4-1 | <i>Diaphorobacter polyhydroxybutyrativorans</i> | GCF_002214645.1 | 34,576 | 58 | 15.43% | 0.50 | 0.69 | 1.00 |
| 4-S | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 20,090 | 62 | 22.89% | 0.46 | 0.49 | 1.00 |
| 4-S | <i>Acinetobacter junii</i> | GCF_001941805.1 | 17,477 | 61 | 17.60% | 0.32 | 0.16 | 1.00 |
| 4-S | <i>Acidovorax</i> sp. JS42 | GCF_000015545.1 | 5,648 | 64 | 4.50% | 0.08 | 0.55 | 1.00 |
| 4-S | <i>Escherichia coli</i> | GCF_000008865.2 | 1,297 | 48 | 0.66% | 0.01 | 0.34 | 1.00 |
| 4-S | <i>Mycolicibacterium peregrinum</i> | GCF_001403655.1 | 3,255 | 37 | 1.00% | 0.02 | 0.27 | 1.00 |
| 5-1 | <i>Acinetobacter junii</i> | GCF_001941805.1 | 132,133 | 65 | 61.84% | 2.55 | 0.13 | 1.00 |

| | | | | | | | | |
|-----|---|-----------------|---------|----|--------|-------|------|------|
| 5-1 | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 123,164 | 64 | 59.33% | 2.89 | 0.18 | 0.99 |
| 5-1 | <i>Acidovorax</i> sp. JS42 | GCF_000015545.1 | 77,769 | 69 | 26.97% | 1.17 | 0.67 | 1.00 |
| 5-1 | <i>Acidovorax ebreus</i> | GCF_000022305.1 | 45,922 | 70 | 22.44% | 0.85 | 0.64 | 1.00 |
| 5-1 | <i>Diaphorobacter polyhydroxybutyrativorans</i> | GCF_002214645.1 | 44,821 | 70 | 20.87% | 0.78 | 0.66 | 1.00 |
| 5-S | <i>Escherichia coli</i> | GCF_000008865.2 | 2,381 | 57 | 1.56% | 0.02 | 0.29 | 1.00 |
| 5-S | <i>Mycolicibacterium peregrinum</i> | GCF_001403655.1 | 1,917 | 43 | 0.68% | 0.01 | 0.23 | 1.00 |
| 5-S | <i>Cloacibacterium normanense</i> | GCF_900104195.1 | 996 | 46 | 0.97% | 0.02 | 0.26 | 1.00 |
| 5-S | <i>Moraxella osloensis</i> | GCF_001553955.1 | 713 | 50 | 0.80% | 0.01 | 0.33 | 1.00 |
| 5-S | <i>Flavobacterium terrigena</i> | GCF_900108955.1 | 946 | 48 | 0.81% | 0.01 | 0.22 | 0.80 |
| 6-1 | <i>Catellibacoccus marimammalium</i> | GCF_000313915.1 | 84,121 | 79 | 78.42% | 5.21 | 0.17 | 1.00 |
| 6-1 | <i>Lactobacillus fuchuensis</i> | GCF_000615805.1 | 3,161 | 78 | 7.50% | 0.12 | 0.10 | 1.00 |
| 6-1 | <i>Aeromonas salmonicida</i> | GCF_000196395.1 | 1,495 | 62 | 1.14% | 0.02 | 0.57 | 1.00 |
| 6-1 | <i>Allivibrio salmonicida</i> | GCF_000196495.1 | 742 | 77 | 0.83% | 0.01 | 0.28 | 1.00 |
| 6-1 | <i>Carnobacterium maltaromaticum</i> | GCF_000317975.2 | 511 | 77 | 0.76% | 0.01 | 0.19 | 1.00 |
| 6-S | <i>Catellibacoccus marimammalium</i> | GCF_000313915.1 | 607,265 | 75 | 89.94% | 35.52 | 0.22 | 0.93 |
| 6-S | <i>Aeromonas salmonicida</i> | GCF_000196395.1 | 13,807 | 64 | 10.25% | 0.18 | 0.33 | 1.00 |
| 6-S | <i>Lactobacillus fuchuensis</i> | GCF_000615805.1 | 1,374 | 70 | 2.81% | 0.05 | 0.24 | 1.00 |
| 6-S | <i>Clostridium frigidicarnis</i> | GCF_900111985.1 | 886 | 67 | 0.80% | 0.01 | 0.46 | 1.00 |
| 6-S | <i>Carnobacterium maltaromaticum</i> | GCF_000317975.2 | 588 | 66 | 0.65% | 0.01 | 0.24 | 1.00 |

Table S5. 20 most abundant taxa identified with the metazoan mitochondrial search

| Species | 1-I | 1-S | 2-I | 2-S | 3-I | 3-S | 4-I | 4-S | 5-I | 5-S | 6-I | 6-S |
|------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|
| <i>Pinguinus impennis</i> | 0 | 0 | 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2693 | 3479 |
| <i>Synthliboramphus antiquus</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 420 | 547 |
| <i>Gasterosteus aculeatus</i> | 0 | 0 | 538 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Cecropis daurica</i> | 0 | 0 | 0 | 0 | 0 | 248 | 0 | 0 | 0 | 0 | 0 | 0 |
| <i>Limosa lapponica</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 134 | 204 |
| <i>Aethia cristatella</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 111 | 144 |
| <i>Gelochelidon nilotica</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 120 | 144 |
| <i>Larus vegae</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 109 | 134 |
| <i>Sternula albifrons</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 51 | 115 |
| <i>Xenus cinereus</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 87 | 115 |
| <i>Sterna hirundo</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 113 | 98 |
| <i>Malassezia restricta</i> | 51 | 50 | 11 | 0 | 55 | 102 | 55 | 0 | 38 | 32 | 0 | 0 |
| <i>Platalea leucorodia</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 56 | 81 |
| <i>Stercorarius maccoormicki</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 58 | 75 |
| <i>Eudypetes chrysocome</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 41 | 75 |
| <i>Didymella pinodes</i> | 20 | 71 | 0 | 10 | 21 | 14 | 50 | 0 | 10 | 0 | 0 | 0 |
| <i>Ara chloropterus</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 69 |
| <i>Saccharomyces arboricola</i> | 0 | 0 | 0 | 0 | 0 | 0 | 65 | 0 | 0 | 0 | 0 | 0 |
| <i>Tetrastes bonasia</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 42 | 62 |
| <i>Synthliboramphus wumizusume</i> | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 49 |

Chapter 4













A 5700 year-old human genome and oral microbiome from chewed birch pitch

ARTICLE

<https://doi.org/10.1038/s41467-019-13549-9>

OPEN

A 5700 year-old human genome and oral microbiome from chewed birch pitch

Theis Z.T. Jensen ^{1,2,10}, Jonas Niemann^{1,2,10}, Katrine Højholt Iversen ^{3,4,10}, Anna K. Fotakis ¹, Shyam Gopalakrishnan ¹, Åshild J. Vågene¹, Mikkel Winther Pedersen ¹, Mikkel-Holger S. Sinding ¹, Martin R. Ellegaard ¹, Morten E. Allentoft¹, Liam T. Lanigan¹, Alberto J. Taurozzi¹, Sofie Holtzmark Nielsen¹, Michael W. Dee⁵, Martin N. Mortensen ⁶, Mads C. Christensen⁶, Søren A. Sørensen⁷, Matthew J. Collins^{1,8}, M. Thomas P. Gilbert ^{1,9}, Martin Sikora ¹, Simon Rasmussen ⁴ & Hannes Schroeder ^{1*}

The rise of ancient genomics has revolutionised our understanding of human prehistory but this work depends on the availability of suitable samples. Here we present a complete ancient human genome and oral microbiome sequenced from a 5700 year-old piece of chewed birch pitch from Denmark. We sequence the human genome to an average depth of 2.3× and find that the individual who chewed the pitch was female and that she was genetically more closely related to western hunter-gatherers from mainland Europe than hunter-gatherers from central Scandinavia. We also find that she likely had dark skin, dark brown hair and blue eyes. In addition, we identify DNA fragments from several bacterial and viral taxa, including Epstein-Barr virus, as well as animal and plant DNA, which may have derived from a recent meal. The results highlight the potential of chewed birch pitch as a source of ancient DNA.

¹The Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 1353, Denmark. ²BioArch, Department of Archaeology, University of York, York YO10 5DD, UK. ³Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby 2800, Denmark. ⁴Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark. ⁵Centre for Isotope Research, University of Groningen, Groningen 9747 AG, The Netherlands. ⁶The National Museum of Denmark, I.C. Modewegs Vej, Brede, Kongens Lyngby 2800, Denmark. ⁷Museum Lolland-Falster, Frisegade 40, Nykøbing Falster 4800, Denmark. ⁸McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK. ⁹University Museum, NTNU, 7012 Trondheim, Norway. ¹⁰These authors contributed equally: Theis Z. T. Jensen, Jonas Niemann, Katrine Højholt Iversen *email: hschroeder@bio.ku.dk

Birch pitch is a black-brown substance obtained by heating birch bark and has been used as an adhesive and hafting agent as far back as the Middle Pleistocene^{1,2}. Small lumps of this organic material are commonly found on archaeological sites in Scandinavia and beyond, and while their use is still debated, they often show tooth imprints, indicating that they were chewed³. Freshly produced birch pitch hardens on cooling and it has been suggested that chewing was a means to make it pliable again before using it, e.g. for hafting composite stone tools. Medicinal uses have also been suggested, since one of the main constituents of birch pitch, betulin, has antiseptic properties⁴. This is supported by a large body of ethnographic evidence, which suggests that birch pitch was used as a natural antiseptic for preventing and treating dental ailments and other medical conditions³. The oldest examples of chewed pitch found in Europe date back to the Mesolithic period and chemical analysis by Gas Chromatography-Mass Spectrometry (GC-MS) has shown that many of them were made from birch (*Betula pendula*)³.

Recent work by Kashuba et al.⁵ has shown that pieces of chewed birch pitch contain ancient human DNA, which can be used to link the material culture and genetics of ancient populations. In the current study, we analyse a further piece of chewed birch pitch, which was discovered at a Late Mesolithic/Early Neolithic site in southern Denmark (Fig. 1a; Supplementary Note 1) and demonstrate that it does not only contain ancient human DNA, but also microbial DNA that reflects the oral microbiome of the person who chewed the pitch, as well as plant and animal DNA which may have derived from a recent meal. The DNA is so exceptionally well preserved that we were able to recover a complete ancient human genome from the sample (sequenced to an average depth of coverage of 2.3×), which is particularly significant since, so far, no human remains have been

recovered from the site⁶. The results highlight the potential of chewed birch pitch as a source of ancient human and non-human DNA, which can be used to shed light on the population history, health status, and even subsistence strategies of ancient populations.

Results

Radiocarbon dating and chemical analysis. Radiocarbon dating of the specimen yielded a direct date of 5,858–5,661 cal. BP (GrM-13305; $5,007 \pm 11$) (Fig. 1b; Supplementary Note 2), which places it at the onset of the Neolithic period in Denmark. Chemical analysis by Fourier-Transform Infrared (FTIR) spectroscopy produced a spectrum very similar to modern birch pitch (Supplementary Fig. 4) and GC-MS revealed the presence of the triterpenes betulin and lupeol, which are characteristic of birch pitch (Fig. 1c; Supplementary Note 3)³. The GC-MS spectrum also shows a range of dicarboxylic acids and saturated fatty acids, which are all considered intrinsic to birch pitch and thus support its identification⁷.

DNA sequencing. We generated approximately 390 million DNA reads for the sample, nearly a third of which could be uniquely mapped to the human reference genome (hg19) (Supplementary Table 2). The human reads displayed all the features characteristic of ancient DNA, including (i) short average fragment lengths, (ii) an increased occurrence of purines before strand breaks, and (iii) an increased frequency of apparent cytosine (C) to thymine (T) substitutions at 5'-ends of DNA fragments (Supplementary Fig. 6) and the amount of modern human contamination was estimated to be around 1–3% (Supplementary Table 3). In addition to the human reads, we generated around 7.3 Gb of

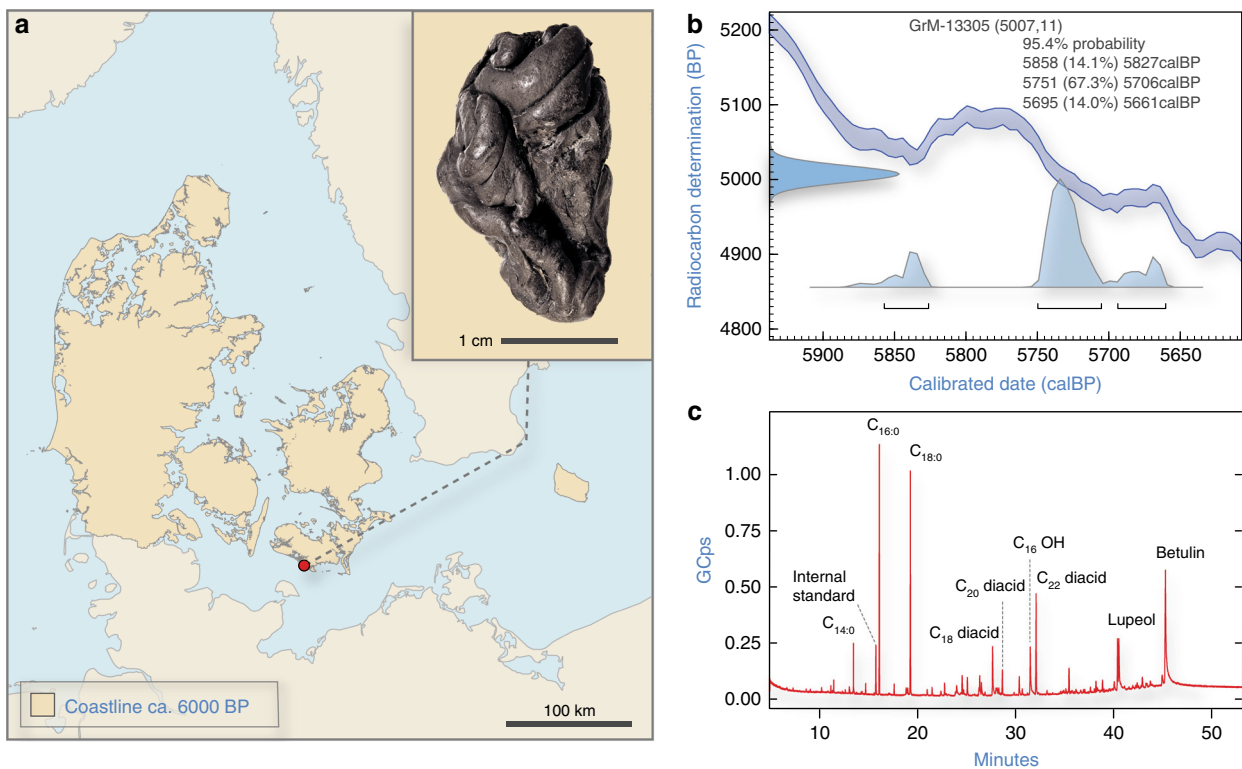


Fig. 1 A chewed piece of birch pitch from southern Denmark. **(a)** Photograph of the Syltholm birch pitch and its find location at the site of Syltholm on the island of Lolland, Denmark (map created using data from Astrup⁷⁸). **(b)** Calibrated date for the Syltholm birch pitch (5,858–5,661 cal. BP; $5,007 \pm 7$). **(c)** GC-MS chromatogram of the Syltholm pitch showing the presence of a series of dicarboxylic acids (C_{xx} diacid) and saturated fatty acids ($C_{xx}:0$) and methyl 16-Hydroxyhexadecanoate ($C_{16}OH$) together with the triterpenes betulin and lupeol, which are characteristic of birch pitch³.

sequence data (68.8%) from the ancient pitch that did not align to the human reference genome.

DNA preservation and genome reconstruction. With over 30%, the human endogenous DNA content in the sample was extremely high and comparable to that found in well-preserved teeth and petrous bones⁸. We used the human reads to reconstruct a complete ancient human genome, sequenced to an effective depth-of-coverage of 2.3×, as well as a high-coverage mitochondrial genome (91×), which was assigned to haplogroup K1e (see Methods). To further investigate the preservation of the human DNA in the sample we calculated a molecular decay rate (*k*, per site per year) and find that it is comparable to that of other ancient human genomes from temperate regions (Supplementary Table 3).

Sex determination and phenotypic traits. Based on the ratio between high-quality reads (MAPQ ≥ 30) mapping to the X and Y chromosomes, respectively⁹, we determined the sex of the individual whose genome we recovered to be female. To predict her hair, eye and skin colour we imputed genotypes for 41 SNPs (Supplementary Data 1) included in the HIrisPlex-S system¹⁰ and find that she likely had dark skin, dark brown hair, and blue eyes (Supplementary Data 2). We also examined the allelic state of two SNPs linked with the primary haplotype associated with lactase persistence in humans and found that she carried the ancestral allele for both (Supplementary Data 1), indicating that she was lactase non-persistent.

Genetic affinities. We called 593,102 single nucleotide polymorphisms (SNPs) in our ancient genome that had previously been genotyped in a dataset of >1000 present-day individuals from a diverse set of Eurasian populations¹¹, as well as >100 previously published ancient genomes (Supplementary Data 3). Figure 2a shows a principal component analysis (PCA) where she clusters with western hunter-gatherers (WHGs). Allele-sharing estimates based on *f_d*-statistics show the same overall affinity to WHGs (Fig. 2b). This is also reflected in the *qpAdm* analysis¹² (see Methods) which demonstrates that a simple one way model assuming 100% WHG ancestry cannot be rejected in favour of

more complex models (Fig. 2c; Supplementary Table 6). To formally test this result we computed two sets of *D*-statistics of the form *D*(Yoruba, EHG/Barcin; test, WHG) and find no evidence for significant levels of EHG or Neolithic farmer gene flow (Supplementary Fig. 7; Supplementary Tables 7, 8).

Metataxonomic profiling of non-human reads. To broadly characterise the taxonomic composition of the non-human reads in the sample, we used MetaPhlan2¹³, a tool specifically designed for the taxonomic profiling of short-read metagenomic shotgun data (see Methods; Supplementary Data 4). Figure 3a shows a principal coordinate analysis where we compare the microbial composition of our sample to that of 689 microbiome profiles from the Human Microbiome Project (HMP)¹⁴. We find that our sample clusters with modern oral microbiome samples in the HMP dataset. This is also reflected in Fig. 3b which shows the order-level microbial composition of our sample compared to two soil samples from the same site and metagenome profiles of healthy human subjects at five major body sites from the HMP¹⁴, visualised using MEGAN6¹⁵.

Oral microbiome characterisation. To further characterise the microbial taxa present in the ancient pitch and to obtain species-specific assignments we used MALT¹⁶, a fast alignment and taxonomic binning tool for metagenomic data that aligns DNA sequencing reads to a user-specified database of reference sequences (see Methods; Supplementary Data 5). As expected, a large number of reads could be assigned to oral taxa, such as *Neisseria subflava* and *Rothia mucilaginosa*, as well as several bacteria included in the red complex (i.e. *Porphyromonas gingivalis*, *Tannerella forsythia*, and *Treponema denticola*) (see Table 1). In addition, we recovered 593 reads that were assigned to Epstein–Barr virus (Human gammaherpesvirus 4). We validated each taxon by examining the edit distances, coverage distributions, and post-mortem DNA damage patterns (see Supplementary Note 5).

Pneumococcal DNA. We also identified several species belonging to the Mitis group of streptococci (Table 1), including

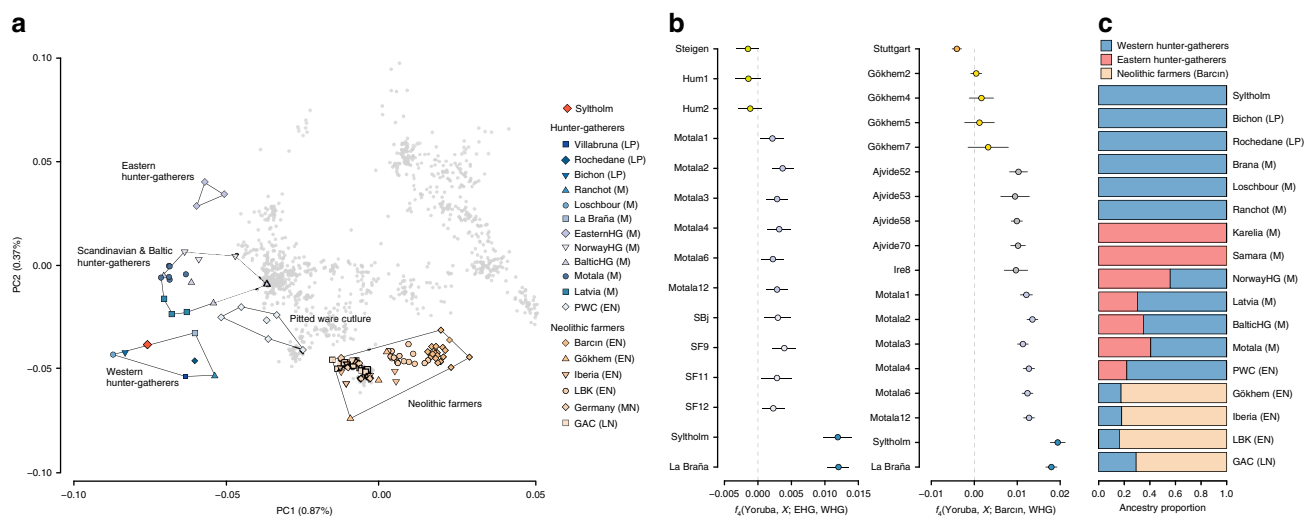


Fig. 2 Genetic affinities of the Sylltholm individual. **a** Principal component analysis of modern Eurasian individuals (in grey) and a selection of over 100 previously published ancient genomes, including the Sylltholm genome. The ancient individuals were projected on the modern variation (see Methods). **b** Allele-sharing estimates between the Sylltholm individual, other Mesolithic and Neolithic individuals, and WHGs versus EHG and Neolithic farmers, respectively, as measured by the statistic *f_d*(Yoruba, X; EHG/Barcin, WHG). **c** Ancestry proportions based on *qpAdm*¹², specifying WHG, EHG, and Neolithic farmers (Barcin) as potential ancestral source populations. PWC Pitted Ware Culture, LBK Linearbandkeramik, GAC Globular Amphora Culture, LP Late Paleolithic, M Mesolithic, EN Early Neolithic, MN Middle Neolithic, LN Late Neolithic. Data are shown in Supplementary Tables 4–6.

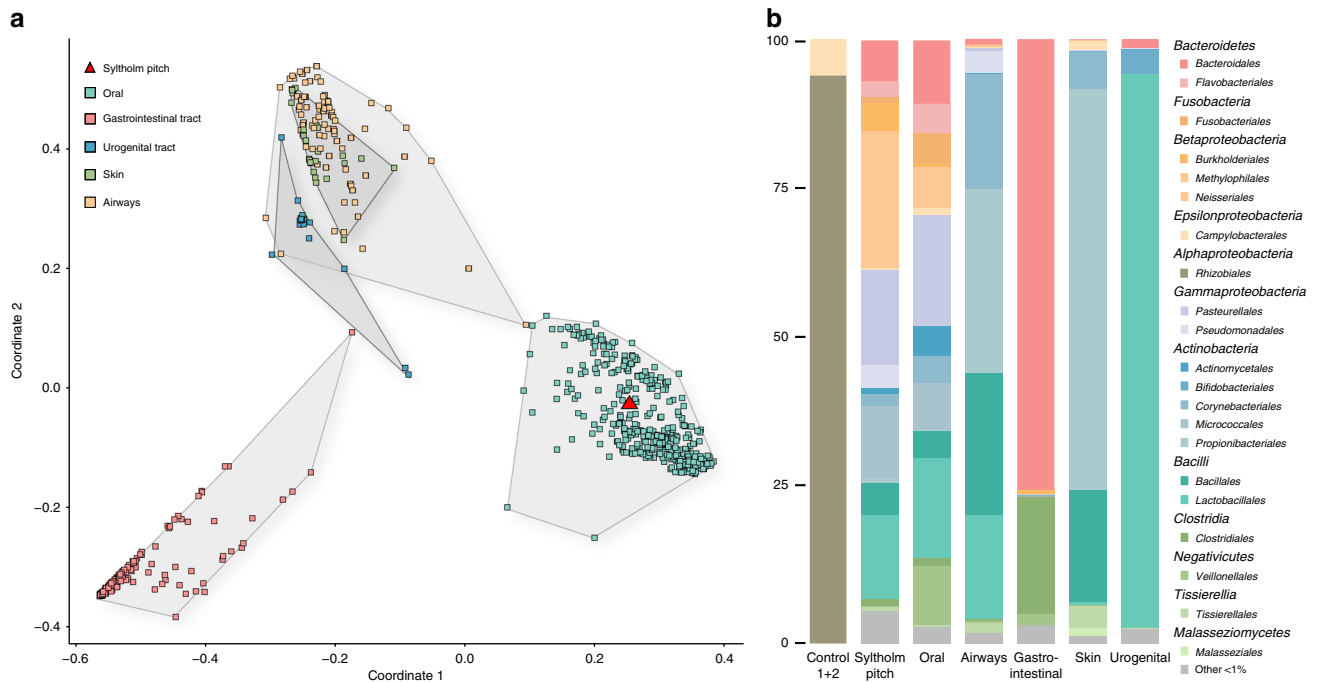


Fig. 3 Metagenomic profile of the Syltholm birch pitch. **a** PCoA with Bray-Curtis at genera level with 689 microbiomes from HMP¹⁴ and the Syltholm sample (see Methods). **b** Order-level microbial composition of the Syltholm sample compared to a control sample (soil) and metagenome profiles of healthy human subjects at five major body sites from the HMP¹⁴, visualised using MEGAN¹⁵.

Streptococcus viridans and *Streptococcus pneumoniae*. We reconstructed a consensus genome from the *S. pneumoniae* reads (Fig. 4) and estimated the number of heterozygous sites (2,597) (see Methods) which indicates the presence of multiple strains. To assess the virulence of the *S. pneumoniae* strains recovered from the ancient pitch, we aligned the contigs against the full Virulence Factor Database¹⁷ in order to identify known *S. pneumoniae* virulence genes (see Methods). We identified 26 *S. pneumoniae* virulence factors within the ancient sample, including capsular polysaccharides (CPS), streptococcal enolase (Eno), and pneumococcal surface antigen A (PsaA) (see Supplementary Data 6).

Plant and animal DNA. Lastly, we used a taxonomic binning pipeline specifically designed for ancient environmental DNA¹⁸ to taxonomically classify the non-human reads in the sample that mapped to other Metazoa (animals) and Viridiplantae (plants). We only parsed taxa with classified reads accounting for >1% of all reads in each of the two kingdoms and a declining edit distance distribution after edit distance 0 (Supplementary Data 7). We then validated each identified taxon as described above (see Supplementary Note 5). Using these criteria, we identified DNA from two plant species in the ancient sample, including birch (*Betula pendula*) and hazelnut (*Corylus avellana*). In addition, we detected over 50,000 reads that were assigned to mallard (*Anas platyrhynchos*).

Discussion

We successfully extracted and sequenced ancient DNA from a 5700-year-old piece of chewed birch pitch from southern Denmark. In addition to a complete ancient human genome (2.3×) and mitogenome (91×), we recovered plant and animal DNA, as well as microbial DNA from several oral taxa. Analysis of the human reads revealed that the individual whose genome we recovered was female and that she likely had dark skin, dark brown hair and blue eyes. This combination of physical traits has

been previously noted in other European hunter-gatherers^{19–22}, suggesting that this phenotype was widespread in Mesolithic Europe and that the adaptive spread of light skin pigmentation in European populations only occurred later in prehistory²³. We also find that she had the alleles associated with lactase non-persistence, which fits with the notion that lactase persistence in adults only evolved fairly recently in Europe, after the introduction of dairy farming with the Neolithic revolution^{24,25}.

From a population genetics point of view, the human genome also offers fresh insights into the early peopling of southern Scandinavia. Recent studies of ancient hunter-gatherer genomes from Sweden and Norway²³ have shown that, following the retreat of the ice sheets around 12–11 ka years ago, Scandinavia was colonised by two separate routes, one from the south (presumably via Denmark) and one from the northeast, along the coast of present-day Norway. This is supported by the fact that hunter-gatherers from central Scandinavia carry different levels of WHG and EHG ancestry, which reached central Scandinavia from the south and northeast, respectively²³. Although we only analysed a single genome, the fact that the Syltholm individual does not carry any EHG ancestry confirms this scenario and suggests that EHG did not reach southern Denmark at this point in prehistory.

The Syltholm genome (5700 years cal. BP) dates to the period immediately following the Mesolithic-Neolithic transition in Denmark. Culturally, this period is marked by the transition from the Late Mesolithic Ertebølle culture (c. 7300–5900 cal. BP) with its flaked stone artefacts and typical T-shaped antler axes, to the early Neolithic Funnel Beaker culture (c. 5900–5300 cal. BP) with its characteristic pottery, polished flint artefacts, and domesticated plants and animals²⁶. In Denmark, the transition from hunting and gathering to farming has often been described as a relatively rapid process, with dramatic shifts in settlement patterns and subsistence strategies²⁷. However, it is still unclear to what extent this transition was driven by the arrival of farming communities as opposed to the local adaptation of farming practices by resident hunter-gatherer populations.

Table 1 List of non-human taxa identified in the Syltholm pitch, including the 40 most abundant oral bacterial taxa, viruses, and eukaryotes. Bacteria in the red complex are denoted with an asterisk. Depth (DoC) and breadth of coverage (>1x) were calculated using BEDTools⁷². Deamination rates at the 5' ends of DNA fragments were estimated using mapDamage 2.0.9⁵⁹. -Δ% refers to the negative difference proportion introduced by Hübler et al⁷⁹. (see Supplementary Note 5).

| Species | Reads | Fragment length (bp) | DoC | SD DoC | >1x (%) | C-T 5' (%) | -Δ% |
|---------------------------------------|---------|----------------------|------|--------|---------|------------|-----|
| Bacteria | | | | | | | |
| <i>Neisseria subflava</i> | 308,732 | 56 | 7.5 | 6.2 | 83.7 | 14.5 | 0.9 |
| <i>Rothia mucilaginosa</i> | 296,610 | 52 | 6.9 | 5.6 | 82.3 | 14.0 | 0.9 |
| <i>Streptococcus pneumoniae</i> | 176,782 | 57 | 4.7 | 6.3 | 65.7 | 13.8 | 0.9 |
| <i>Neisseria cinerea</i> | 153,683 | 58 | 4.9 | 5.1 | 71.7 | 15.1 | 1.0 |
| <i>Lautropia mirabilis</i> | 117,040 | 53 | 2.0 | 1.9 | 71.9 | 13.0 | 1.0 |
| <i>Neisseria meningitidis</i> | 100,540 | 51 | 2.3 | 4.3 | 42.4 | 14.9 | 0.9 |
| <i>Aggregatibacter segnis</i> | 95,670 | 58 | 2.8 | 2.8 | 73.3 | 14.5 | 0.9 |
| <i>Neisseria elongata</i> | 68,407 | 54 | 1.6 | 1.9 | 67.6 | 15.1 | 0.9 |
| <i>Prevotella intermedia</i> | 65,324 | 56 | 1.2 | 1.4 | 55.0 | 16.2 | 0.9 |
| <i>Streptococcus</i> sp. ChDC B345 | 52,614 | 61 | 1.6 | 2.7 | 50.3 | 13.8 | 0.9 |
| <i>Streptococcus</i> sp. 431 | 43,787 | 59 | 1.2 | 1.9 | 47.5 | 13.6 | 0.8 |
| <i>Aggregatibacter aphrophilus</i> | 43,231 | 56 | 1.1 | 1.6 | 50.4 | 15.0 | 0.8 |
| <i>Streptococcus pseudopneumoniae</i> | 38,832 | 61 | 1.1 | 2.4 | 34.9 | 14.4 | 0.9 |
| <i>Capnocytophaga leadbetteri</i> | 36,461 | 59 | 0.9 | 1.1 | 49.8 | 14.0 | 0.8 |
| <i>Corynebacterium matruchotii</i> | 36,070 | 52 | 0.7 | 0.9 | 44.0 | 13.0 | 1.0 |
| <i>Gemella morbillorum</i> | 32,284 | 63 | 1.2 | 1.5 | 56.4 | 16.3 | 1.0 |
| <i>Streptococcus viridans</i> | 27,840 | 60 | 0.8 | 1.5 | 36.5 | 14.5 | 1.0 |
| <i>Neisseria gonorrhoeae</i> | 27,704 | 53 | 0.7 | 2.0 | 21.3 | 15.0 | 1.0 |
| <i>Neisseria sicca</i> | 27,290 | 57 | 0.6 | 1.4 | 22.5 | 13.7 | 0.9 |
| <i>Fusobacterium nucleatum</i> | 26,783 | 64 | 0.8 | 1.1 | 47.8 | 14.1 | 0.9 |
| <i>Prevotella fusca</i> | 26,295 | 57 | 0.5 | 0.7 | 34.6 | 15.7 | 1.0 |
| <i>Kingella kingae</i> | 25,811 | 55 | 0.7 | 1.0 | 44.2 | 14.4 | 1.0 |
| <i>Ottowia</i> sp. 894 | 25,425 | 52 | 0.5 | 0.7 | 34.6 | 14.4 | 1.0 |
| <i>Streptococcus</i> sp. NPS 308 | 24,937 | 59 | 0.8 | 1.4 | 37.5 | 14.3 | 0.8 |
| <i>Actinomyces oris</i> | 24,029 | 52 | 0.4 | 0.7 | 29.8 | 12.7 | 1.0 |
| <i>Streptococcus australis</i> | 23,777 | 60 | 0.7 | 1.3 | 31.5 | 13.8 | 1.0 |
| <i>P. propionicum</i> | 22,864 | 50 | 0.3 | 0.6 | 26.8 | 13.2 | 0.9 |
| <i>Haemophilus</i> sp. O36 | 19,707 | 62 | 0.7 | 1.5 | 28.4 | 14.5 | 1.0 |
| <i>Porphyromonas gingivalis</i> * | 17,651 | 55 | 0.4 | 0.7 | 32.2 | 17.2 | 1.0 |
| <i>Capnocytophaga gingivalis</i> | 16,734 | 58 | 0.3 | 0.6 | 27.1 | 15.0 | 1.0 |
| <i>Neisseria polysaccharea</i> | 14,442 | 57 | 0.4 | 1.4 | 15.0 | 15.8 | 1.0 |
| <i>Tannerella forsythia</i> * | 14,187 | 55 | 0.2 | 0.5 | 19.8 | 15.3 | 1.0 |
| <i>Streptococcus</i> sp. A12 | 13,232 | 59 | 0.4 | 0.9 | 24.9 | 14.6 | 0.9 |
| <i>Capnocytophaga sputigena</i> | 12,587 | 58 | 0.2 | 0.5 | 19.9 | 14.7 | 0.9 |
| <i>Neisseria lactamica</i> | 11,971 | 56 | 0.3 | 1.0 | 14.2 | 14.2 | 0.8 |
| <i>Treponema denticola</i> * | 11,379 | 59 | 0.2 | 0.5 | 19.5 | 14.0 | 0.8 |
| <i>Rothia dentocariosa</i> | 10,944 | 54 | 0.2 | 0.5 | 20.0 | 13.6 | 1.0 |
| <i>Tannerella</i> sp. HOT-286 | 10,397 | 53 | 0.2 | 0.5 | 15.7 | 14.0 | 1.0 |
| <i>Actinomyces meyeri</i> | 10,105 | 51 | 0.3 | 0.5 | 21.3 | 14.0 | 1.0 |
| <i>Filifactor alocis</i> | 9,948 | 61 | 0.3 | 0.6 | 25.6 | 15.0 | 1.0 |
| Viruses | | | | | | | |
| <i>Epstein-Barr virus</i> | 593 | 51 | 0.2 | 0.4 | 13.3 | 17.8 | 1.0 |
| Eukaryotes | | | | | | | |
| <i>Anas platyrhynchos</i> | 55,986 | 51 | <0.1 | 0.05 | 0.2 | 15.6 | 1.0 |
| <i>Corylus avellana</i> | 8,615 | 55 | <0.1 | 0.04 | 0.1 | 19.7 | 1.0 |
| <i>Betula pendula</i> | 3,291 | 54 | <0.1 | 0.02 | <0.1 | 16.1 | 1.0 |

Our analyses have shown that the Syltholm individual does not carry any Neolithic farmer ancestry, suggesting that the genetic impact of Neolithic farming communities in southern Scandinavia might not have been as instant or pervasive as once thought²⁸. While the mtDNA we recovered belongs to haplogroup K1e, which is more commonly associated with early farming communities^{29–31}, there is mounting evidence to suggest that this lineage was already present in Mesolithic Europe^{32–34}. Overall, the lack of Neolithic farmer ancestry is consistent with evidence from elsewhere in Europe, which suggests that genetically distinct hunter-gatherer groups survived for much longer than previously assumed^{35–37}. These WHG “survivors” might have triggered the resurgence of hunter-gatherer ancestry that is proposed to have occurred in central Europe between 7000 and 5000 BP¹².

In addition to the human data, we recovered ancient microbial DNA from the pitch which could be shown to have a human oral microbiome signature. Previous studies^{38–40} have demonstrated that calcified dental plaque (dental calculus) provides a robust biomolecular reservoir that allows direct and detailed investigations of ancient oral microbiomes. However, unlike dental calculus, which represents a long-term reservoir of the oral microbiome built up over many years, the microbiota found in ancient mastics are more likely to give a snapshot of the species active at the time. As such, they provide a useful source of information regarding the evolution of the human oral microbiome that can complement studies of ancient dental calculus.

The majority of the bacterial taxa we identified (Table 1) are classified as non-pathogenic, commensal species that are considered to be part of the normal microflora of the human mouth

studies^{43,44} have demonstrated the great potential of ancient DNA for studying the long-term evolution of blood borne viruses. Formally known as Human gammaherpesvirus 4, EBV is one of the most common human viruses infecting over 90% of the world's adult population⁴⁵. Most EBV infections occur during childhood and in the vast majority of cases they are asymptomatic or they carry symptoms that are indistinguishable from other mild, childhood diseases. However, in some cases EBV can cause infectious mononucleosis (glandular fever)⁴⁶ and it has also been associated with various lymphoproliferative diseases, such as Hodgkin's lymphoma and hemophagocytic lymphohistiocytosis, as well as higher risks of developing certain autoimmune diseases, such as dermatomyositis and multiple sclerosis^{47,48}.

Lastly, we identified several thousand reads that could be confidently assigned to different plant and animal species, including birch (*B. pendula*), hazelnut (*C. avellana*), and mallard (*A. platyrhynchos*). While the presence of birch DNA is easily explained as it is the source of the pitch, we propose that the hazelnut and mallard DNA may derive from a recent meal. This is supported by the faunal evidence from the site, which is dominated by wild taxa, including *Anas* sp. and hazelnuts^{6,49}. In addition, there is evidence from many other Mesolithic and Early Neolithic sites in Scandinavia for hazelnuts being gathered in large quantities for consumption⁵⁰. Together with the faunal evidence, the ancient DNA results support the notion that the people at Syltholm continued to exploit wild resources well into the Neolithic and highlight the potential of ancient DNA analyses of chewed pieces of birch pitch for palaeodietary studies.

In summary, we have shown that pieces of chewed birch pitch are an excellent source of ancient human and non-human DNA. In the process of chewing, the DNA becomes trapped in the pitch where it is preserved due to the aseptic and hydrophobic properties of the pitch which both inhibits microbial and chemical decay. The genomic information preserved in chewed pieces of birch pitch offers a snapshot of people's lives, providing information on genetic ancestry, phenotype, health status, and even subsistence. In addition, the microbial DNA provides information on the composition of our ancestral oral microbiome and the evolution of specific oral microbes and important human pathogens.

Methods

Sample preparation and DNA extraction. We sampled c. 250 mg from the specimen for DNA analysis. Briefly, the sample was washed in 5% bleach solution to remove any surface contamination, rinsed in molecular biology grade water and left to dry. We tested three different extraction methods using between 20–50 mg of starting material: For method (1), 1 ml of lysis buffer containing 0.45 M EDTA (pH 8.0) and 0.25 mg/ml Proteinase K was added to the sample and left to incubate on a rotor at 56 °C. After 12 h the supernatant was removed and concentrated down to ~150 µl using Amicon Ultra centrifugal filters (MWCO 30 kDa), mixed 1:10 with a PB-based binding buffer⁵¹, and purified using MinElute columns, eluting in 30 µl EB. For method (2) the sample was digested and purified as above, but with the addition of a phenol-chloroform clean-up step. Briefly, 1 ml phenol (pH 8.0) was added to the lysis mix, followed by 1 ml chloroform:isoamyl alcohol. The supernatant was concentrated and purified, as described above. For method (3) the sample was dissolved in 1 ml chloroform:isoamylalcohol. The dissolved sample was then resuspended in 1 ml molecular grade water and purified as described above. DNA extracts prepared using a Proteinase K-based lysis buffer followed by a phenol-chloroform based purification step produced the best results in terms of the endogenous human DNA content (see Supplementary Table 1); however, following metagenomic profiling the extracts were found to be contaminated with *Delftia* spp., a known laboratory contaminant⁵². The contaminated libraries were excluded from metagenomic profiling.

Negative controls. We included no template controls (NTC) during the DNA extraction and library preparation steps. The NTCs prepared with the additional phenol-chloroform step were also found to be contaminated with *Delftia* spp., suggesting that the contaminants were introduced during this step. In addition, we included two soil samples from the site, weighing c. 2 g each, as negative controls. DNA was extracted as described above using 3 ml EDTA-based lysis buffer followed by 9 ml 25:24:1 phenol:chloroform:isoamyl alcohol mixture to account for the larger amount of starting material. The sequencing results are reported in Supplementary Table 1.

Library preparation and sequencing. 16 µl of each DNA extract were built into double-stranded libraries using a recently published protocol that was specifically designed for ancient DNA⁵³. One extraction NTC was included, as well as a single library NTC. 10 µl of each library were amplified in 50 µl reactions for between 15 and 28 cycles, using a dual indexing approach⁵⁴. The optimal number of PCR cycles was determined by qPCR (MxPro 3000, Agilent Technologies). The amplified libraries were purified using SPRI-beads and quantified on a 2200 TapeStation (Agilent Technologies) using High Sensitivity tapes. The amplified and indexed libraries were then pooled in equimolar amounts and sequenced on 1/8 of a lane of an Illumina HiSeq 2500 run in SR mode. Following initial screening, additional reads were obtained by pooling libraries #2, #3, and #4 in molar fractions of 0.2, 0.4, and 0.4, respectively and sequencing them on one full lane of an Illumina HiSeq 2500 run in SR mode.

Data processing. Base calling was performed using Illumina's bcl2fastq2 conversion software v2.20.0. Only sequences with correct indexes were retained. FastQ files were processed using PALEOMIX v1.2.12⁵⁵. Adapters and low quality reads (Q < 20) were removed using AdapterRemoval v2.2.0⁵⁶, only retaining reads >25 bp. Trimmed and filtered reads were then mapped to hg19 (build 37.1) using BWA⁵⁷ with seed disabled to allow for better sensitivity⁵⁸, as well as filtering out unmapped reads. Only reads with a mapping quality ≥30 were kept and PCR duplicates were removed. MapDamage 2.0.9⁵⁹ was used to evaluate the authenticity of the retained reads as part of the PALEOMIX pipeline⁵⁵, using a subsample of 100k reads per sample (Supplementary Fig. 6). For the population genomic analyses, we merged the ancient sample with individuals from the Human Origin dataset¹¹ and >100 previously published ancient genomes (Supplementary Data 1). At each SNP in the Human Origin dataset, we sampled the allele with more reads in the ancient sample, resolving ties randomly, resulting in a pseudohaploid ancient sample.

MtDNA analysis and contamination estimates. We used Schmutzi⁶⁰ to determine the endogenous consensus mtDNA sequence and to estimate present-day human contamination. Reads were mapped to the Cambridge reference sequence (rCRS) and filtered for MAPQ ≥ 30. Haploid variants were called using the *endoCaller* program implemented in Schmutzi⁶⁰ and only variants with a posterior probability exceeding 50 on the PHRED scale (probability of error: 1/100,000) were retained. We then used Haplogrep v2.2.6⁶¹ to determine the mtDNA haplogroup, specifying PhyloTree (build 17) as the reference phylogeny⁶². Contamination estimates were obtained using Schmutzi's *mtCont* program and a database of putative modern contaminant mitochondrial DNA sequences.

Genotype imputation. We used ANGS⁶³ to compute genotype likelihoods in 5 Mb windows around 43 SNPs associated with skin, eye, and hair colour¹⁰ and lactase persistence into adulthood (Supplementary Data 2). Missing genotypes were imputed using impute2⁶⁴ and the pre-phased 1000 Genome reference panel⁶⁵, provided as part of the impute2 reference datasets. We used multiple posterior probability thresholds, ranging from 0.95 to 0.50, to filter the imputed genotypes. The imputed genotypes were uploaded to the HIRISplex-S website¹⁰ to obtain the predicted outcomes for the pigmentation phenotypes (Supplementary Data 3).

Principal component analysis. Principal component analysis was performed using smartPCA⁶⁶ by projecting the ancient individuals onto a reference panel including >1000 present-day Eurasian individuals from the HO dataset¹¹ using the option *lsq* project. Prior to performing the PCA the data set was filtered for a minimum allele frequency of at least 5% and a missingness per marker of at most 50%. To mitigate the effect of linkage disequilibrium, the data were pruned in a 50-SNP sliding window, advanced by 10 SNPs, and removing sites with an R² larger than 0, resulting in a final data set of 593,102 SNPs.

D- and f-statistics. *D-* and *f-*statistics were computed using *AdmixTools*⁶⁷. To estimate the amount of shared drift between the Syltholm genome and WHG versus EHG and Neolithic farmers, respectively, we computed two sets of *f₄*-statistics of the form *f₄*(Yoruba, X; EHG/Barcin, WHG) where "X" stands for the test sample. Standard errors were calculated using a weighted block jackknife. To confirm the absence of EHG and Neolithic farmer gene flow in the Syltholm genome and to contrast this result with those obtained for other Mesolithic and Neolithic individuals from Scandinavia, we computed two sets of *D*-statistics of the form *D*(Yoruba, EHG/Barcin; X, WHG) testing whether "X" forms a clade to the exclusion of EHG and Neolithic farmers (represented by Barcin), respectively.

qpAdm. Admixture proportions were modeled using *qpAdm*¹², specifying Mesolithic Western European hunter-gatherers (WHG), Eastern hunter-gatherers (EHG) and early Neolithic Anatolian farmers (Barcin), as possible ancestral source populations. We present the model with the lowest number of source populations that fits the data, as well as the model with all three admixture components (see Supplementary Table 6). When estimating the admixture proportions for WHGs and EHG, the test sample was excluded from their respective reference populations.

MetaPhlan. We used MetaPhlan2¹³ to create a metagenomic profile based on the non-human reads (Supplementary Data 4). The reads were first aligned to the MetaPhlan2 database¹³ using Bowtie2 v2.2.9 aligner⁶⁸. PCR duplicates were removed using PALEOMIX filteruniquebam⁵⁸. For cross-tissue comparisons 689 human microbiome profiles published in the Human Microbiome Project Consortium¹⁴ were initially used, comprising samples from the mouth ($N = 382$), skin ($N = 26$), gastrointestinal tract ($N = 138$), urogenital tract ($N = 56$), airways and nose ($N = 87$). The oral HMP samples consist of attached/keratinised gingiva ($N = 6$), buccal mucosa ($N = 107$), palatine tonsils ($N = 6$), tongue dorsum ($N = 128$), throat ($N = 7$), supragingival plaque ($N = 118$), and subgingival plaque ($N = 7$). Pairwise ecological distances among the profiles were computed at genus and species level using taxon relative abundances and the `vegdist` function from the `vegan` package in R⁶⁹. These were used for principal coordinate analysis (PCoA) of Bray–Curtis distances in R using the `pcor` function included in the `APE` package⁷⁰. Subsequently, we calculated the average relative abundance of each genus for each of the body sites present in the Human Microbiome Project and visualised the abundance of microbial orders of our sample and the HMP body sites with MEGAN6¹⁵.

MALT. To further characterise the metagenomic reads we performed microbial species identification using MALT v. 0.4.1 (Megan ALignment Tool)¹⁶, a rapid sequence-alignment tool specifically designed for the analysis of metagenomic data. All complete bacterial ($n = 12,426$) and viral ($n = 8094$) genomes were downloaded from NCBI RefSeq on 13 November 2018, and all complete archaeal ($n = 280$) genomes were downloaded from NCBI RefSeq on 17 November 2018 to create a custom database. In an effort to exclude genomes that may consist of composite sequences from multiple organisms, the following entries were excluded:

GCF_000922395.1 uncultured crAssphage
GCF_000954235.1 uncultured phage WW-nAnB
GCF_000146025.2 uncultured Termite group 1 bacterium phylotype Rs-D17

The final MALT reference database contained 33,223 genomes and was created using default parameters in `malt-build` (v. 0.4.1). The sequencing data for the ancient pitch sample, two soil control samples and associated extraction and library blanks were de-enriched for human reads by mapping to the human genome (hg19) using BWA aln and excluding all mapping reads. Duplicates were removed with `seqkit v.0.7.1`⁷¹ using the `rmDup` function with the `-by-seq` flag. The remaining reads were processed with `malt-run` (v. 0.4.1) where BlastN mode and SemiGlobal alignment were used. The minimum percent identity (`-minPercentIdentity`) was set to 95, the minimum support (`-minSupport`) parameter was set to 10 and the top percent value (`-topPercent`) was set as 1. Remaining parameters were set to default. MEGAN6¹⁵ was used to visualise the output `.rma6` files and to extract the reads assigned to taxonomic nodes of interest for our sample. A taxon table of the raw MALT output for all samples and blanks, as well as species level read assignments to bacteria, archaea and DNA viruses for the ancient pitch sample are shown in Supplementary Data 5, where reads listed are the sum of all reads assigned to the species node, including reads assigned to specific strains within the species. Reads assigned to RNA viruses were not considered for further analyses, since our dataset consisted of DNA sequences only. Due to the limited number of reads assigned to archaeal species (Supplementary Data 5), we did not consider Archaea in downstream analyses of species identification. To validate the microbial taxa, we aligned the assigned reads to their respective reference genomes and examined the edit distances, coverage distributions, and post-mortem DNA damage patterns (see Supplementary Note 5).

Pneumococcus analysis. We reconstructed a *S. pneumoniae* consensus genome (Fig. 4) by mapping all reads assigned to *S. pneumoniae* by MALT¹⁶ to the *S. pneumoniae* TIGR4 reference genome (NC_003028.3). To investigate the presence of multiple strains we estimated the number of heterozygous sites using `samtools`⁵⁷ `mpileup` function, filtering out transitions, indels, and sites with a depth of coverage below 10. Coverage statistics of the individual alignments ($MQ \geq 30$) were obtained using `Bedtools`⁷² and plotted using `Circos`⁷³ in 100 bp windows. Mappability was estimated using `GEM2`⁷⁴ using a k -mer size of 50 and a read length of 42, which is comparable to the average length of the trimmed and mapped reads in the ancient pitch. Virulence genes were identified by assembling the ancient *S. pneumoniae* MALT extracts into contigs using `megahit`⁷⁵. The contigs were aligned against known *S. pneumoniae* TIGR4 virulence genes in the Virulence Factor Database¹⁷ (downloaded 22/11–2018) using `BLASTn`⁷⁶. Only unique hits with a bitscore >200 , $>20\%$ coverage, and an identity $>80\%$ were considered as shared genes (Supplementary Data 6).

To identify all streptococcus virulence factors in the ancient pitch, we aligned the contigs against the full Virulence Factor Database¹⁷ (downloaded 22/11–2018) using `BLASTn`⁷⁶ and the same filtering criteria as described above (Supplementary Data 6). To validate the approach we repeated the analysis with five modern oral microbiome samples (SRS014468; SRS019120; SRS013942; SRS015055; SRS014692) from the Human Microbiome Project (HMP)¹⁴ using only the forward read (R1) (Supplementary Data 6). We find that the number of virulence genes we recovered directly correlates with sequencing depth (Supplementary Fig. 16).

Holi. For a robust taxonomic assignment of reads aligning to Metazoa (animals) and Viridiplantae (plants), all non-human reads were parsed through the ‘Holi’ pipeline¹⁸, which was specifically developed for the taxonomic profiling of ancient metagenomic shotgun reads. Each read was aligned against the NCBI’s full Nucleotide and Refseq databases (downloaded November 25th 2018), including a newly sequenced full genome of European hazelnut (*Corylus avellana*, downloaded April 10th 2019)⁷⁷. The alignments were then parsed through a naive lowest common ancestor algorithm (ngsLCA) based on the NCBI taxonomic tree. Only taxonomically classified reads for taxa comprising $\geq 1\%$ of all the reads within the two kingdoms and a declining edit distance distribution after edit distance 0 were parsed for taxonomic profiling and further validation. To validate the assignments, we aligned the assigned reads to their respective reference genomes and examined the edit distances, coverage distributions, and post-mortem DNA damage patterns (see Supplementary Note 5; Supplementary Data 7).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequencing reads are available for download from the European Nucleotide Archive under accession number PRJEB30280. All other data are included in the paper or available upon request.

Received: 17 June 2019; Accepted: 15 November 2019;

Published online: 17 December 2019

References

- Mazza, P. P. A. et al. A new Palaeolithic discovery: tar-hafted stone tools in a European Mid-Pleistocene bone-bearing bed. *J. Archaeol. Sci.* **33**, 1310–1318 (2006).
- Kozowyk, P. R. B., Soressi, M., Pomstra, D. & Langejans, G. H. J. Experimental methods for the Palaeolithic dry distillation of birch bark: implications for the origin and development of Neandertal adhesive technology. *Sci. Rep.* **7**, 8033 (2017).
- Aveling, E. M. & Heron, C. Chewing tar in the early Holocene: an archaeological and ethnographic evaluation. *Antiquity* **73**, 579–584 (1999).
- Haque, S. et al. Screening and characterisation of antimicrobial properties of semisynthetic betulin derivatives. *PLoS ONE* **9**, e102696 (2014).
- Kashuba, N. et al. Ancient DNA from mastics solidifies connection between material culture and genetics of mesolithic hunter-gatherers in Scandinavia. *Commun. Biol.* **2**, 185 (2019).
- Sørensen, S. A. Syltholm: Denmark’s largest Stone Age excavation. *Mesolithic Misc.* **24**, 3–10 (2016).
- Aveling, E. M. & Heron, C. Identification of birch bark tar at the Mesolithic site of Star Carr. *Ancient Biomolecules* **2**, 69–80 (1998).
- Gamba, C. et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
- Skoglund, P., Storå, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40**, 4477–4482 (2013).
- Chaitanya, L. et al. The HIRISplex-S system for eye, hair and skin colour prediction from DNA: Introduction and forensic developmental validation. *Forensic Sci. Int. Genet.* **35**, 123–135 (2018).
- Lazaridis, I. et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
- Haak, W. et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
- Truong, D. T. et al. MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
- The Human Microbiome Project Consortium. et al. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
- Huson, D. H. et al. MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* **12**, e1004957 (2016).
- Vågene, Å. J. et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).
- Chen, L. et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–D328 (2005).
- Pedersen, M. W. et al. Postglacial viability and colonization in North America’s ice-free corridor. *Nature* **537**, 45–49 (2016).
- Olalde, I. et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).
- Skoglund, P. et al. Genomic diversity and admixture differs for stone-age Scandinavian foragers and farmers. *Science* **344**, 747–750 (2014).

21. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
22. Brace, S. et al. Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat. Ecol. Evol.* **3**, 765–771 (2019).
23. Günther, T. et al. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* **16**, e2003703 (2018).
24. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659–674 (2017).
25. Ségurel, L. & Bon, C. On the Evolution of Lactase Persistence in Humans. *Annu. Rev. Genomics Hum. Genet.* **18**, 297–319 (2017).
26. Gron, K. J. & Sørensen, L. Cultural and economic negotiation: a new perspective on the Neolithic Transition of Southern Scandinavia. *Antiquity* **92**, 958–974 (2018).
27. Richards, M. P., Price, T. D. & Koch, E. Mesolithic and Neolithic subsistence in Denmark: new stable isotope data. *Curr. Anthropol.* **44**, 288–295 (2003).
28. Becker, C. J. *Mosefundne lerkar fra yngre stenalder: studier over tragtbøgerkulturen i Danmark* (Copenhagen, 1948).
29. Bramanti, B. et al. Genetic Discontinuity Between Local Hunter-Gatherers and Central Europe's First Farmers. *Science* **326**, 137–140 (2009).
30. Brandt, G. et al. Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity. *Science* **342**, 257–261 (2013).
31. Isern, N., Fort, J. & de Rioja, V. L. The ancient cline of haplogroup K implies that the Neolithic transition in Europe was mainly demic. *Sci. Rep.* **7**, 11229 (2017).
32. Hofmanová, Z. et al. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl Acad. Sci. USA* **113**, 6886–6891 (2016).
33. González-Forbes, G. et al. Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. *Curr. Biol.* **27**, 1801–1810.e10 (2017).
34. Mittnik, A. et al. The genetic prehistory of the Baltic Sea region. *Nat. Commun.* **9**, 442 (2018).
35. Bollongino, R. et al. 2000 years of parallel societies in Stone Age Central Europe. *Science* **342**, 479–481 (2013).
36. Lipson, M. et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature* **551**, 368–372 (2017).
37. Jones, E. R. et al. The neolithic transition in the Baltic was not driven by admixture with early European farmers. *Curr. Biol.* **27**, 576–582 (2017).
38. Adler, C. J. et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat. Genet.* **45**, 450–455 (2013).
39. Warinner, C. et al. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).
40. Jersie-Christensen, R. R. et al. Quantitative metaproteomics of medieval dental calculus reveals individual oral health status. *Nat. Commun.* **9**, 4744 (2018).
41. Suzuki, N., Yoneda, M. & Hirofujii, T. Mixed red-complex bacterial infection in periodontitis. *Int. J. Dent.* **2013**, 587279 (2013).
42. Weiser, J. N., Ferreira, D. M. & Paton, J. C. *Streptococcus pneumoniae*: transmission, colonization and invasion. *Nat. Rev. Microbiol.* **16**, 355–367 (2018).
43. Krause-Kyora, B. et al. Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *Elife* **7**, e36666 (2018).
44. Mühlemann, B. et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature* **557**, 418–423 (2018).
45. Williams, H. & Crawford, D. H. Epstein-Barr virus: the impact of scientific advances on clinical practice. *Blood* **107**, 862–869 (2006).
46. Henle, G., Henle, W. & Diehl, V. Relation of Burkitt's tumor-associated herpes-type virus to infectious mononucleosis. *Proc. Natl Acad. Sci. USA* **59**, 94–101 (1968).
47. Toussiro, E. & Roudier, J. Epstein-Barr virus in autoimmune diseases. *Best. Pract. Res. Clin. Rheumatol.* **22**, 883–896 (2008).
48. Rezk, S. A., Zhao, X. & Weiss, L. M. Epstein-Barr virus (EBV)-associated lymphoid proliferations, a 2018 update. *Hum. Pathol.* **79**, 18–41 (2018).
49. Bangsgaard, P. *Report on the faunal remains from MLF 00906-II Syltholm II*. (Zoological Museum, University of Copenhagen, 2015).
50. Regnell, M. Plant subsistence and environment at the Mesolithic site Tägerup, southern Sweden: new insights on the 'Nut Age'. *Veg. Hist. Archaeobot.* **21**, 1–16 (2012).
51. Allentoft, M. E. et al. Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
52. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).
53. Caroe, C. et al. Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* **9**, 410–419 (2018).
54. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2011).
55. Schubert, M. et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056 (2014).
56. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
58. Schubert, M. et al. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).
59. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
60. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
61. Weissensteiner, H. et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
62. van Oven, M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci. Int.: Genet. Suppl. Ser.* **5**, e392–e394 (2015).
63. Korneliusen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: analysis of next generation sequencing data. *BMC Bioinforma.* **15**, 356 (2014).
64. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
65. Consortium, The 1000 Genomes Project. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2011).
66. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
67. Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
69. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
70. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
71. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **11**, e0163962 (2016).
72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
73. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
74. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).
75. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultrafast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
76. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinforma.* **10**, 421 (2009).
77. Rowley, E. R. et al. A Draft Genome and High-Density Genetic Map of European Hazelnut (*Corylus avellana* L.). Preprint at <https://www.biorxiv.org/content/https://doi.org/10.1101/469015v1> (2018).
78. Astrup, P. M. *Sea-level change in Mesolithic southern Scandinavia. Long- and short-term effects on society and the environment*. Jysk Arkæologisk Selskabs Skrifter **106** (Jutland Archaeological Society, 2018).
79. Huebler, R., Key, F. M. M., Warinner, C., Bos, K. I. & Krause, J. HOPS: Automated detection and authentication of pathogen DNA in archaeological remains. Preprint at <https://www.biorxiv.org/content/https://doi.org/10.1101/534198v2> (2018).

Acknowledgements

We thank the Museum Lolland-Falster for access to the sample and the staff at the Danish National High-Throughput Sequencing Center for technical assistance. We also thank Miren Iraeta Orbegozo, Oliver Smith and Kristine Bohmann for their input and helpful discussion. This research was funded by a research grant from VILLUM FONDEN (grant no. 22917) awarded to H.S. T.Z.T.J. and J.N. were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 676154 (ArchSci2020). K.H.I. was supported by the Danish Heart Foundation. M.J.C., A.J.T., and L.T.L. were funded by Danish National Research Foundation (DNRF128). M.D. was supported by a European Research Council grant (ECHOES, 714679). S.R. was supported by the Novo Nordisk Foundation grant NNF14CC0001 and the Jorck Foundation Research Award. H.S. was supported in part by HERA (Humanities in the European Research Area) through the joint research programme "Uses of the Past" and the European Union's Horizon 2020 research and innovation programme under grant agreement no. 649307 (CitiGen).

Author contributions

T.Z.T.J. and H.S. designed and led the study. S.A.S. provided the sample for analysis. M.C.C. and M.N.M. performed the FTIR and GC-MS analyses. M.W.D. performed the radiocarbon dating. T.Z.T.J., M.H.S.S. and M.R.E. generated the genetic data. T.Z.T.J., J.N., K.H.I., A.K.F., S.G., Å.J.V., M.W.P., S.H.N., M.E.A. and H.S. analyzed the genetic data. T.Z.T.J., J.N., K.H.I., A.K.F., S.G., Å.J.V., M.W.P., M.E.A., L.T.L., A.J.T., M.J.C., M.T.P.G., M.S., S.R., and H.S. interpreted the results. T.Z.T.J. and H.S. wrote the manuscript with input from J.N., K.H.I., and the remaining authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-019-13549-9>.

Correspondence and requests for materials should be addressed to H.S.

Peer review information *Nature Communications* thanks Christina Warinner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

Supplementary Information for

A 5,700 year-old human genome and oral microbiome from chewed birch pitch

Theis Z. T. Jensen^{1,2+}, Jonas Niemann^{1,2+}, Katrine Højholt Iversen^{3,4+}, Anna K. Fotakis¹, Shyam Gopalakrishnan¹, Åshild J. Vågene¹, Mikkel Winther Pedersen¹, Mikkel-Holger S. Sinding¹, Martin R. Ellegaard¹, Morten E. Allentoft¹, Liam T. Lanigan¹, Alberto J. Taurozzi¹, Sofie Holtsmark Nielsen¹, Michael W. Dee⁵, Martin N. Mortensen⁶, Mads C. Christensen⁶, Søren A. Sørensen⁷, Matthew J. Collins^{1,8}, M. Thomas P. Gilbert¹, Martin Sikora¹, Simon Rasmussen⁴, Hannes Schroeder^{1*}

¹The Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 1353, Denmark

²BioArch, Department of Archaeology, University of York, York YO10 5DD, UK

³Department of Bio and Health Informatics, Technical University of Denmark, Kongens Lyngby 2800, Denmark

⁴Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark

⁵Centre for Isotope Research, University of Groningen, Groningen 9747 AG, The Netherlands

⁶The National Museum of Denmark, I.C. Modewegs Vej, Brede, Kongens Lyngby 2800, Denmark

⁷Museum Lolland-Falster, Frisegade 40, Nykøbing Falster 4800, Denmark

⁸McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, UK

⁺These authors contributed equally to this work.

*Correspondence and requests for materials should be addressed to: H.S. (email: hschroeder@bio.ku.dk)

| | |
|---|---|
| Supplementary Note 1. Site description | 3 |
| Supplementary Note 2. Radiocarbon dating | 4 |
| Supplementary Note 3. FTIR and GC-MS analysis | 5 |
| Supplementary Note 4. Decay rate estimate | 6 |
| Supplementary Note 5. Analysis of metagenomic reads | 7 |

Supplementary Note 1. Site description

Theis Z. T. Jensen and Søren A. Sørensen

Syltholm is located in the southern part of Lolland near Rødbyhavn in Denmark (Supplementary Fig. 1). The site covers ca. 187 hectares of land, which up until 1872 was open water. After a series of floods, a reclamation project was undertaken to dam up the area, thus preserving the inundated landscape below¹. In 2012, Museum Lolland-Falster initiated large scale geological surveys and subsequent archaeological excavations in the area due to the upcoming establishment of the Fehmarn Belt fixed-link tunnel connecting Denmark to Germany. Several sites were selected for full-scale excavations, based on coring, landscape topography as well as investigative excavations. Excavation of the former fjord was significantly constrained by high groundwater levels, which preclude the initial use of mechanical excavators. This was solved by localised drainage for several months^{2,3}. To date a total of 21 excavations have been completed. They vary in terms of age, finds intensity as well as preservation of organic material. The vast majority of the site spans from the Late Mesolithic Ertebølle to the Middle Neolithic Funnel Beaker periods. During the time of occupation the area would have been a shallow brackish lagoon protected from the open sea to the south by shifting sandy barrier islands. Human activity in this coastal environment in prehistory is reflected primarily by the finding of large numbers of organic and inorganic artefacts and thousands of faunal remains, many of which include cut-marks^{3,4}.

In the area of the site where the chewed birch pitch (Supplementary Fig. 2) was found (MLF906-I-II), the stratigraphy consists of 1) a top layer is a heterogeneous matrix of sand separated by thin sections of gyttja, ca. 1-1,5 m in thickness, which indicates several storm events, 2) a thin (5-10 cm) transgression horizon of coarse drift gyttja containing large amounts of molluscs as well as *ex situ* water rolled artefacts of flint and organic material, indicating an erosive milieu, 3) a layer (10-80 cm) of coarse brown gyttja where large amounts of *in situ* organic and inorganic archaeological artefacts and other material was uncovered, indicating a sheltered environment, and 4) a bottom layer of glacial till, which consists of blue clay. The glacial topography consists of several small depressions, which in the Ertebølle and Neolithic periods would, at certain places, gradually have been filled with organic matter forming gyttja.

Several hundred unpublished AMS dates from MLF906-II, including the ones presented in this manuscript (Supplementary Fig. 3) indicate that the area was frequented by people more or less continuously from the Late Mesolithic until Middle Neolithic. Continued artefact depositions seem to have been carried out at the site, as dates obtained from organic material, such as bone, antler or wood, found in small confined areas span nearly 1,000 years. During the earlier (Mesolithic) phase, the deposits are dominated by wild taxa, including red deer (*Cervus elaphus*), roe deer (*Capreolus capreolus*), and pig (*Sus sp.*), as well as ox (*Vulpes vulpes*), otter (*Lutra lutra*), and wildcat (*Felis silvestris*), although large numbers of domestic dog (*Canis familiaris*) remains have also been found⁵. From around 5,800 BP other domesticated species (e.g. *Bos taurus*) also start to appear, but keep being dominated by wild taxa (see Supplementary Fig. 3).

Supplementary Note 2. Radiocarbon dating

Michael W. Dee

Radiocarbon dating was performed on ca. 10 mg of the birch pitch, following an acid-base-acid pretreatment⁶. First, the sample was treated with 4% HCl (80°C) and then rinsed to neutrality with ultra-pure water. Second, a basic solution 1% NaOH (RT) was applied, and the reaction vessel rinsed again to neutrality. Finally, a further acid step was applied using 4% HCl (80°C) to ensure no atmospheric CO₂ absorbed during the alkaline phase remained in the reaction vessel. After a last rinse to neutrality, the product was thoroughly air dried. An aliquot of ca. 4 mg was then weighed into a tin capsule for combustion in an Elementar IsotopeCube NCS Elemental Analyser (EA). The EA was coupled to an Isoprime 100 Isotope Ratio Mass Spectrometer, which allowed the δ¹³C value of the sample to be measured, as well as a fully automated cryogenic system that trapped the liberated CO₂ into an airtight vessel. The vessel was manually transferred to a vacuum manifold, where a stoichiometric excess of H₂(g) (1: 2.5) was added, and the sample CO₂(g) reduced to graphite over an Fe(s) catalyst. The graphite was pressed into a cathode for radioisotope analysis in an MICADAS IonPlus accelerator mass spectrometer. The MICADAS generated an estimate of the ¹⁴C:¹²C ratio that was close to ±1%, and from this data, and in accordance with all standard operations and conventions, the ¹⁴C date (in yrs BP) was calculated. The calendar date range (years cal BP) was computed using the calibration program OxCal (v 4.3)⁷.

Inaccuracy in ¹⁴C dating largely arises from processes that occur before the sample reaches the laboratory. Misassociation of sample and context, or factors which can make substances ‘appear older’ such as marine/dietary reservoir effects or material reuse, are the most common. Enriching effects, which would cause the age to be too young, are negligible in the natural world. However, human error in the ¹⁴C laboratory can introduce both younger or older contamination. To guard against such sources of inaccuracy, the radiocarbon laboratory in Groningen regularly analyzes reference samples, including full pretreatments and measurements on materials of independently known age.

Supplementary Note 3. FTIR and GC-MS analysis

Martin N. Mortensen and Mads C. Christensen

For the FTIR analysis ca. 1 mg of sample was ground with KBr (Fischer Scientific, IR Grade), pressed into a pellet and measured on a Perkin Elmer Spectrum 1000 FT-IR spectrometer. The FTIR spectra for the Syltholm pitch and a modern birch sample are shown in Supplementary Fig. 4. For the GC-MS analysis, ca. 0.5 mg sample was hydrolysed in methanolic KOH (Merck) and extracted with GC-grade tert-Butyl methyl ether (MTBE) after acidification. The extract was methylated using diazomethane (Sigma-Aldrich)⁸. 1 μl of this solution was injected on a Bruker SCION 456 GC-TQMS system equipped with a Programmable Temperature Vaporizer that was held at 64°C for 0.5 min, raised to 315°C at 200°C min⁻¹ and held at that temperature for 40 min. The split ratio was high during the first 0.5 min and then switched to 5. The GC column was a Restek Rtx-5 capillary column (30 m, 0.25 mm ID, 0.25 μm) and the He flow rate was 1 cm³ min⁻¹. The GC oven temperature was held at 64°C for 0.5 min, then raised to 190°C at 10°C min⁻¹ and then onto 315°C at 4°C min⁻¹ and held at that temperature for 15 min. The EI (electron ionisation) ion source was held at 250°C and the ionisation potential was -70 eV. The mass spectrometer was operated in the full scan mode from m/z 45 to m/z 800. The GC-MS chromatograms for the Syltholm pitch and the betulin and lupeol references are shown in Supplementary Fig. 5.

Supplementary Note 4. Decay rate estimate

Morten E. Allentoft and Hannes Schroeder

To investigate the rate of human DNA degradation in the ancient pitch sample we examined the DNA read length distributions of the mapped reads, using a previously published method⁹. The distribution follows a typical pattern of degraded DNA with an initial increase in the number of reads towards longer DNA fragments, followed by a decline. We observe that the declining part of the distribution follows an exponential decay curve ($R^2=0.99$), as expected if the DNA had been randomly fragmented over time. Deagle et al.¹⁰ showed that the decay constant (λ) in the exponential equation represents the fraction of broken bonds in the DNA strand (the damage fraction) and that $1/\lambda$ is the average theoretical fragment length in the DNA library. By solving the equation, we obtain a DNA damage fraction (λ) of 3.4%, which corresponds to a theoretical average fragment length ($1/\lambda$) of 29 bp (Table S2). We note that this is not directly comparable to the observed average length, which is affected by lab methods and sequencing technology. If the DNA is found in a stable matrix long term DNA fragmentation can be expressed as a rate and the damage fraction (λ , per site) can be converted to a decay rate (k , per site per year), when the age of the sample is known. By applying an estimated age of 5,700 years for the Syltholm pitch, the corresponding DNA decay rate (k) is $5.96 \cdot 10^{-6}$ breaks per bond per year, which corresponds to a molecular half-life of 1,162 years for a 100 bp DNA fragment. This means that after 1,162 years (post cell death), each 100 bp DNA stretch will have experienced one break on average. This estimated rate of DNA decay for the pitch sample seems within the expected age for DNA preserved in a stable matrix in a temperate climate zone. For example, the rate is close to that observed in the La Braña sample¹¹, preserved at similar temperatures as the pitch sample (Supplementary Table 2). By contrast, the DNA decay in human remains from warmer climates is much faster¹². Although these calculations are only based on a single sample, the results suggest that ancient mastics provide remarkable conditions for molecular preservation.

Supplementary Note 5. Analysis of metagenomic reads

Jonas Niemann and Hannes Schroeder

Accurate taxonomic classification of complex metagenomic datasets can be challenging, especially if closely related species are present in the sample or as environmental contaminants¹³. Additionally, robust classification can be complicated if reference databases are incomplete or sequencing effort is insufficient. A further complication is that, in some cases, reference databases contain poor quality genomes with contaminant sequences, which can lead to incorrect assignments^{14,15}. While specific pipelines for the taxonomic classification of ancient metagenomic datasets have been developed^{16,17}, further validation is often necessary to exclude the possibility of false positive (misidentified) assignments. Methods used for validation include confirming the presence of ancient DNA damage patterns, evaluating edit distances, and assessing coverage distributions^{18,19}.

To test the robusticity of our pipelines^{16,17}, we performed two *in silico* experiments using archaeological and environmental samples as controls. First, we ran MetaPhlan2²⁰ and MALT¹⁷ on two soil samples from the site and show that they have a completely different microbial composition from the ancient pitch (Fig. 4; Supplementary Data 4; Supplementary Data 5). We then ran Holi¹⁶ on the same controls and, using the same criteria as for the ancient sample, did not retrieve any reads that could be assigned to the eukaryotic taxa we identified in the ancient pitch (Supplementary Data 7). Second, we ran the Holi pipeline¹⁶ on a previously published dataset¹² generated from an ancient tooth (~33 million reads with an average length of 69 bp) to test whether some of our results might be false positives resulting from reference genomes being contaminated with DNA from other species, especially human DNA. Using the same criteria as the ones we applied in the present study, we did not identify any of the taxa we identified in the ancient pitch.

Independent validation of taxonomic assignments

To validate the taxonomic assignments of the metagenomic reads recovered from the ancient pitch, we aligned the assigned reads to their respective reference genomes and examined the edit distances, coverage distributions, and post-mortem DNA damage patterns^{18,19}. For the bacterial taxa identified by MALT, we chose to further investigate bacterial species with $\geq 10,000$ assigned reads (including strain specific reads). We then aligned the taxon-specific MALT extracts to their respective reference genomes that we obtained from the NCBI assembly database (Supplementary Data 5). The sequences were aligned using *bwa aln*²¹ and PCR duplicates were removed using Picard Tools v.2.13.2²². MapDamage v.2.0.9²³ was used to estimate deamination rates (Supplementary Fig. 8). The breadth and depth of coverage were calculated with *bedtools* v.2.27.1²⁴ and visualised with *Circos* v.0.69-6²⁵ using a window size of 100 bp (Supplementary Fig. 9). Edit distances for all reads and filtered for PMD score ≥ 1 were extracted from the bam files with *samtools view*²¹ and *PMDtools*²⁶ and plotted in R v.3.4.1²⁷ (Supplementary Fig. 10). The negative difference proportion ($-\Delta\%$) was calculated using only reads with PMD score ≥ 1 . This metric was first introduced by Hübner et al.¹⁹ and is a measure of the decline in the edit distance distribution, with a $-\Delta\%$ value of 1

indicating a strictly declining distribution. Correct taxonomic assignments generally result in a continuously declining edit distance distribution, which reflects the fact that most of the aligned reads show no or only few mismatches, mostly resulting from aDNA damage or divergence of the ancient genome from the modern reference. By contrast, mapping to an incorrect reference tends to result in an increased number of mismatches, which is reflected in the edit distance distribution¹⁹. For the microbial taxa, we report species-specific assignments with a $-\Delta\%$ value >0.8 to account for the possibility of cross-alignments due to horizontal gene transfer and the presence of closely related microbial species in the sample.

The Human Oral Microbiome Database (HOMD) was referred to in order to classify bacterial species as belonging to the human oral/respiratory microbiome or as environmental. Of the 64 most abundant bacterial species identified in the ancient pitch (Supplementary Data 5), four are known contaminants originating from lab reagents (*Delftia* spp.), which are also evident in the extraction blanks (Supplementary Data 5), while seven (*Pseudomonas stutzeri*, *Hydrogenophaga* sp. RAC07, *Leptospira alstonii*, *Ramlibacter tataouinensis*, *Thalassolituus oleivorans*, *Achromobacter spanius*, *Pseudomonas aeruginosa*) are likely derived from the environment. None of these 11 species showed the characteristic damage patterns of ancient DNA and were, therefore, not included in further analyses. The remaining 53 bacterial species are predominantly found in the oral cavity and the upper respiratory tract (see Table 1; Supplementary Data 5).

Among the viral species identified we chose to further authenticate reads assigned to the Epstein-Barr virus (*Human gammaherpesvirus 4*) (Supplementary Fig. 11), since it is the only non-bacteriophage viral taxon to which ≥ 200 reads were assigned. Viruses have considerably smaller genomes than bacteria and were therefore subject to a lower threshold of assigned reads.

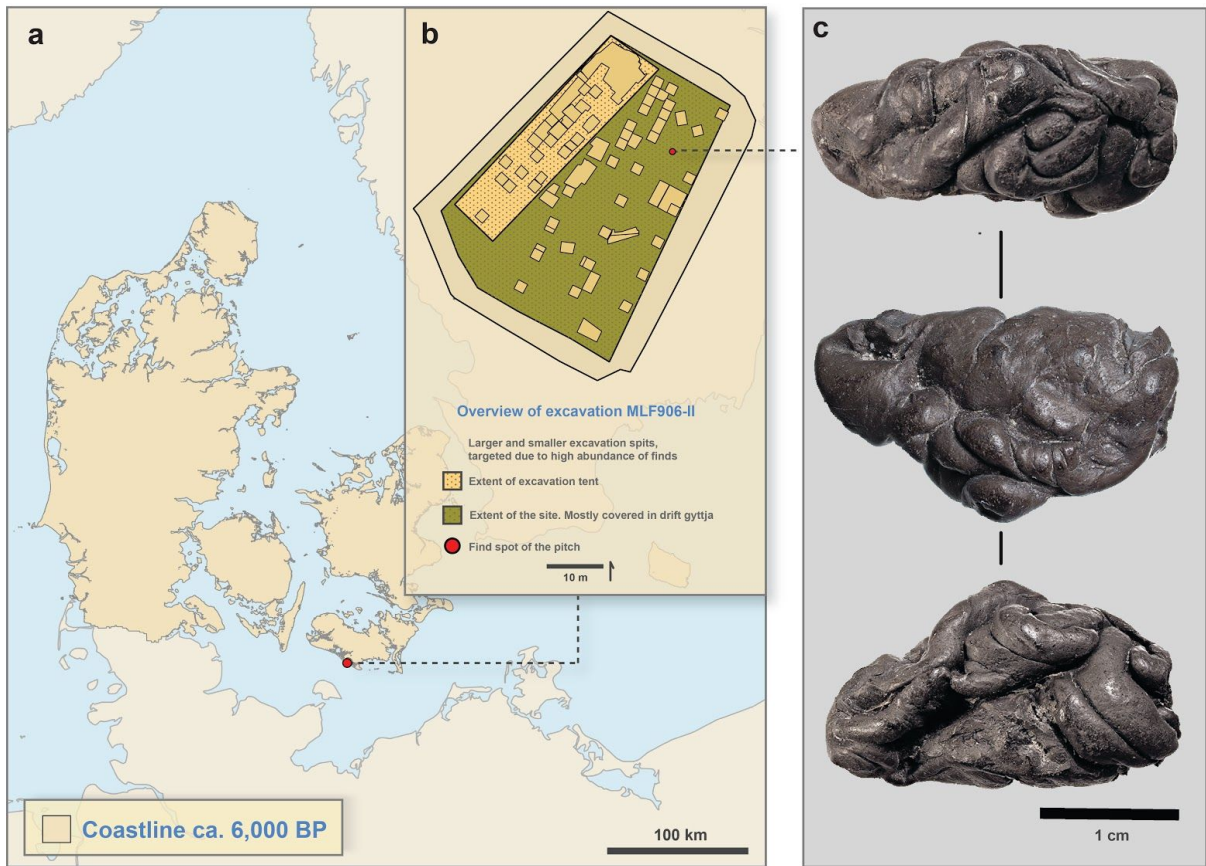
The plant and animal taxa identified by Holi¹⁶ were validated by evaluating sequence identity through edit distance distributions, evenness of coverage, and the presence of post-mortem DNA damage patterns as described above after extracting family level reads for each taxon and aligning them to their respective reference genome. For taxa with low coverage, we used bedtools²⁴ to calculate the proportion of mapped bases with a coverage $>1\times$ as an alternative way of assessing evenness of coverage (Supplementary Data 7). Using these criteria, we identified four taxa (*Anas platyrhynchos*, *Anser cygnoides*, *Betula pendula*, *Corylus avellana*) which showed characteristic ancient DNA damage patterns and a strictly declining edit distance distribution (Supplementary Fig. 12-14 and Supplementary Data 6). However, further analysis using mitochondrial (mtDNA) genomes as reference yielded only 291 reads aligning to the *A. cygnoides* mtDNA in contrast to 2,541 for the *A. platyrhynchos* (mallard) mtDNA, with $>99\%$ of bases covered and nearly $10\times$ average depth of coverage (Supplementary Data 6). Furthermore, the edit distance distribution for the *A. cygnoides* mtDNA (Supplementary Fig. 13) is not declining, suggesting a poor match. We therefore excluded *A. cygnoides* as a likely false positive assignment.

As a further validation step and to assess whether reads from multiple taxa might have been misassigned to a single species, we examined the number of multiallelic sites in the $10\times$ haploid mallard mtDNA. In haploid genomes (i.e. bacterial genomes or mtDNA), the vast

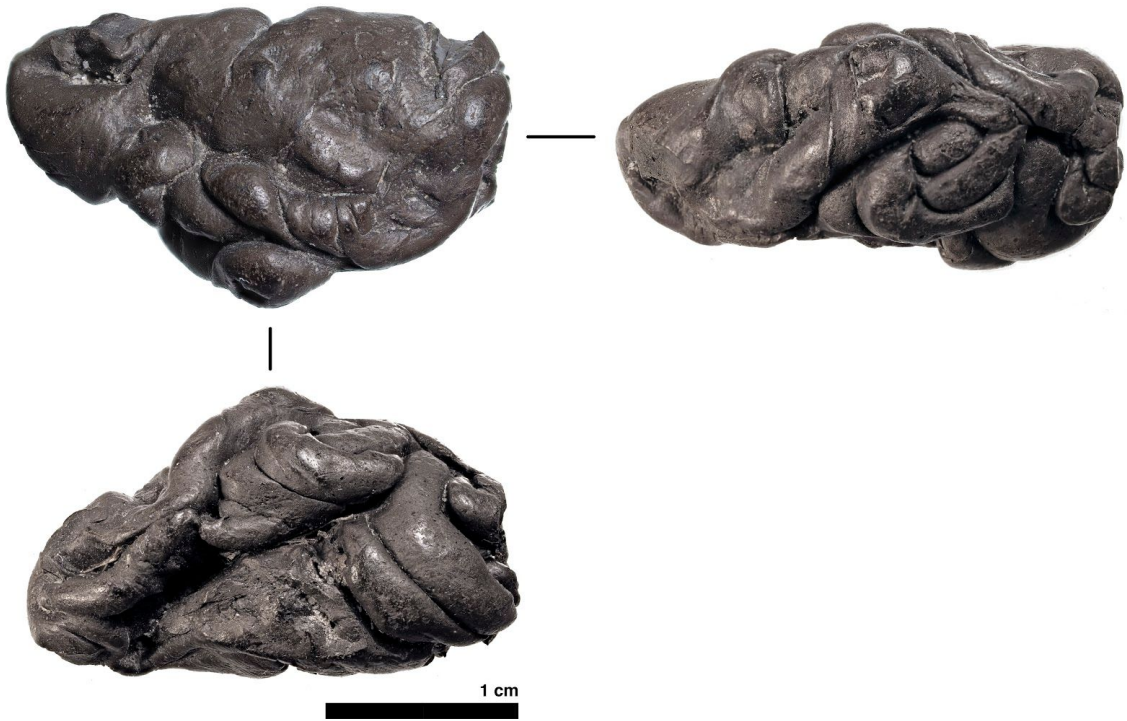
majority of variable sites should be monoallelic, so that a large number of multiallelic sites might be indicative of multiple species or strains being present¹³. To assess the allele frequency distribution for the 10× mallard mtDNA, we rescaled the base qualities of the mallard mtDNA reads according to their likelihood of being damaged using mapDamage v.2.0.9²³ and called variants using samtools²¹ mpileup function using a minimum depth of 10. The allele frequency distribution follows a normal distribution with a mean of ~0.5 indicating the presence of two haplotypes (Supplementary Fig. 15). This was confirmed by visual inspection of the alignment in IGV v.2.3.9²⁸. However, rather than indicating the presence of two different taxa, we believe that this might indicate the presence of two individuals and it is not inconceivable that two or more individuals were consumed. This is supported by the fact that the only other Anatidae species with a significant number of reads identified by Holi¹⁶ was the swan goose (*A. cygnoides*). However, as discussed above, we excluded this taxon based on the poor level of sequence identity with the *A. cygnoides* mtDNA as evident in the edit distances (Supplementary Fig. 13). We were unable to evaluate haploidy for the two plant taxa (*Betula pendula* and *Corylus avellana*) since the depth of coverage of the chloroplast DNA was too low.

We also identified 3,213 reads that could be assigned to the human endoparasite *Spirometra erinaceieuropaei* (tapeworm). However, although the reads appear to be ancient, coverage was not even (>60% of mapped bases >1× despite an average depth of coverage of only 0.000025×) suggesting that they are likely false positive alignments perhaps due to the presence of contaminant (human) sequences in the reference (Supplementary Data 7). Recent studies¹⁵ have shown that public genome assemblies of parasitic worms can be contaminated with DNA from the host species, other species that are commensal in the host, or laboratory contaminants, highlighting the need for curating public reference genome databases²⁹.

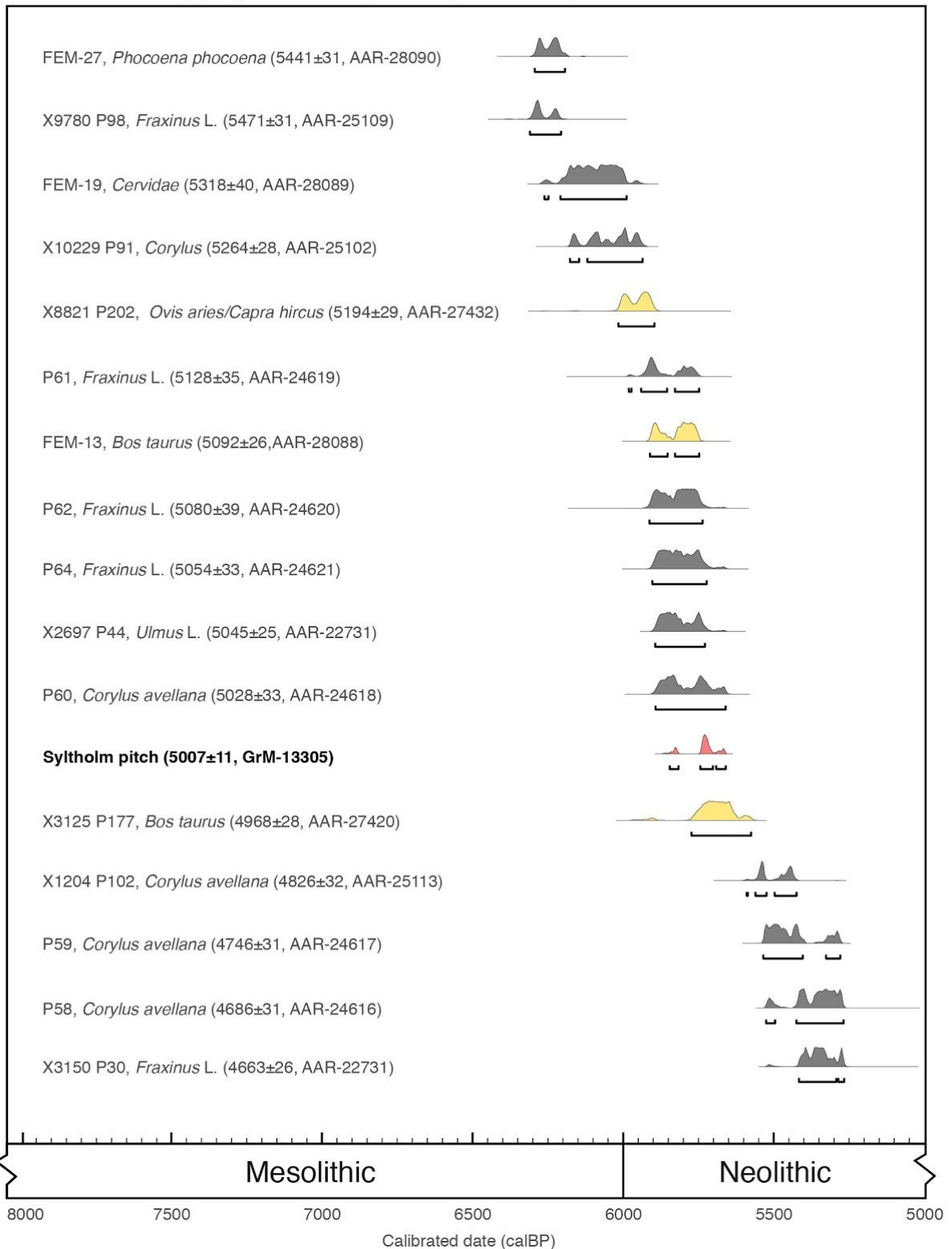
Lastly, we acknowledge that it is possible that some of the eukaryotic taxa we report (e.g. mallard) may have come from the environment as opposed to the diet. However, since the vast majority of the DNA we retrieved from the ancient pitch appears to be endogenous (i.e. either human or from the oral cavity), we find this to be unlikely and we believe that it is more likely that the taxa we report derived from the diet.



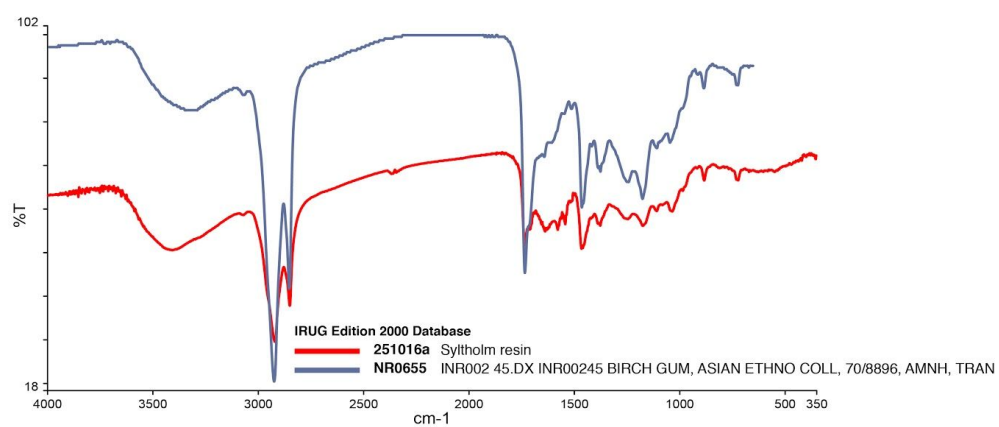
Supplementary Figure 1. a, Map of Denmark showing the location of Syltholm on the island of Lolland (map created using data from Astrup³⁰). **b**, GIS site plan of the excavation and findspot of the birch pitch. **c**, photograph of the birch pitch.



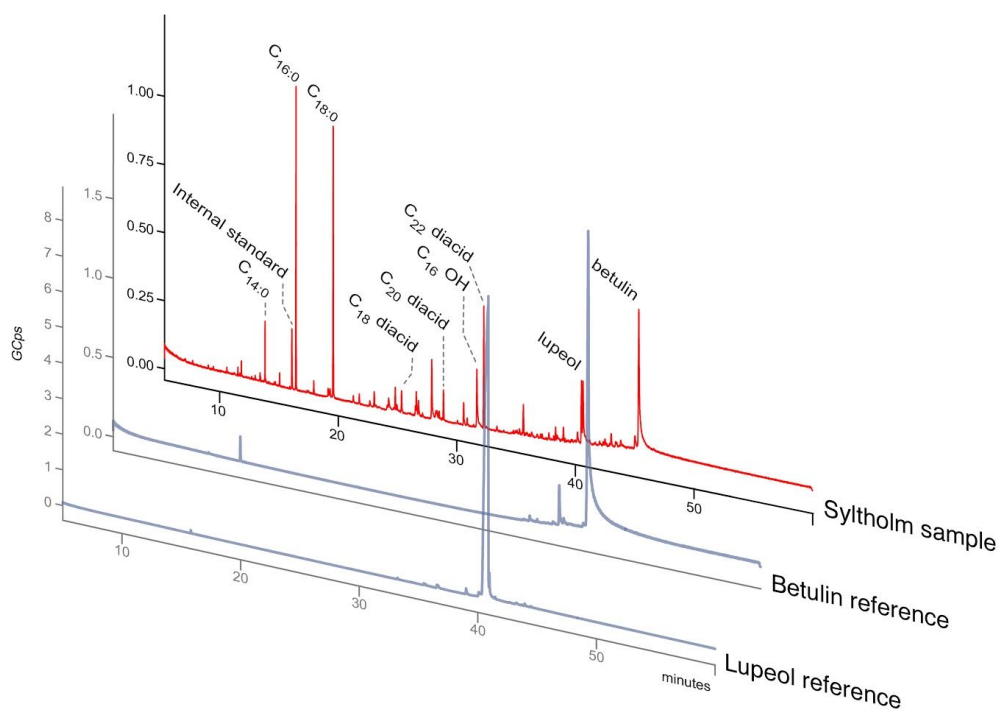
Supplementary Figure 2. Close-up photograph of the Syltholm pitch.



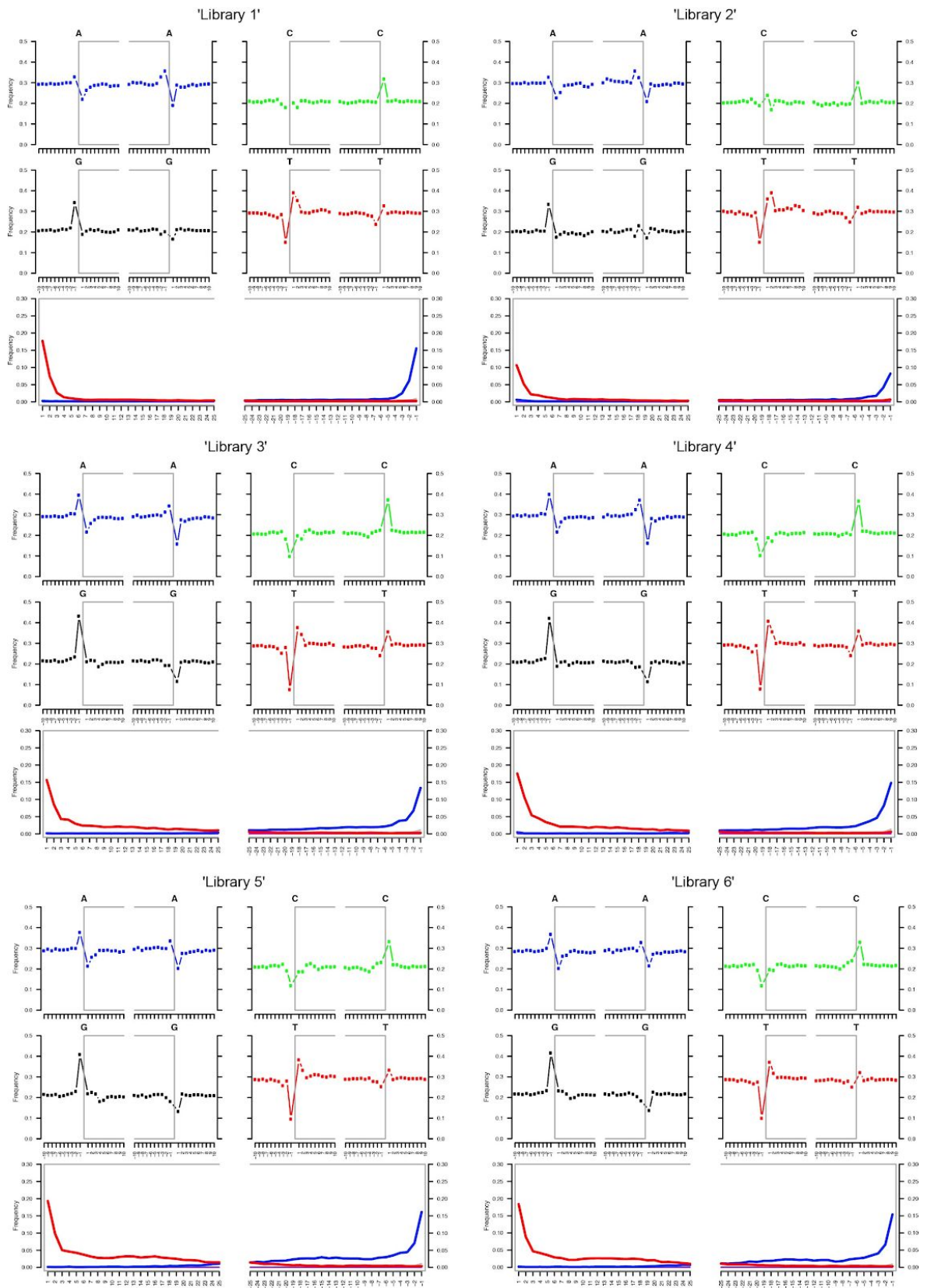
Supplementary Figure 3. Radiocarbon chronology for Syltholm site MLF906-II based on a series of 17 calibrated radiocarbon dates, including the birch pitch (marked in red). Samples from domesticated species are marked in yellow.



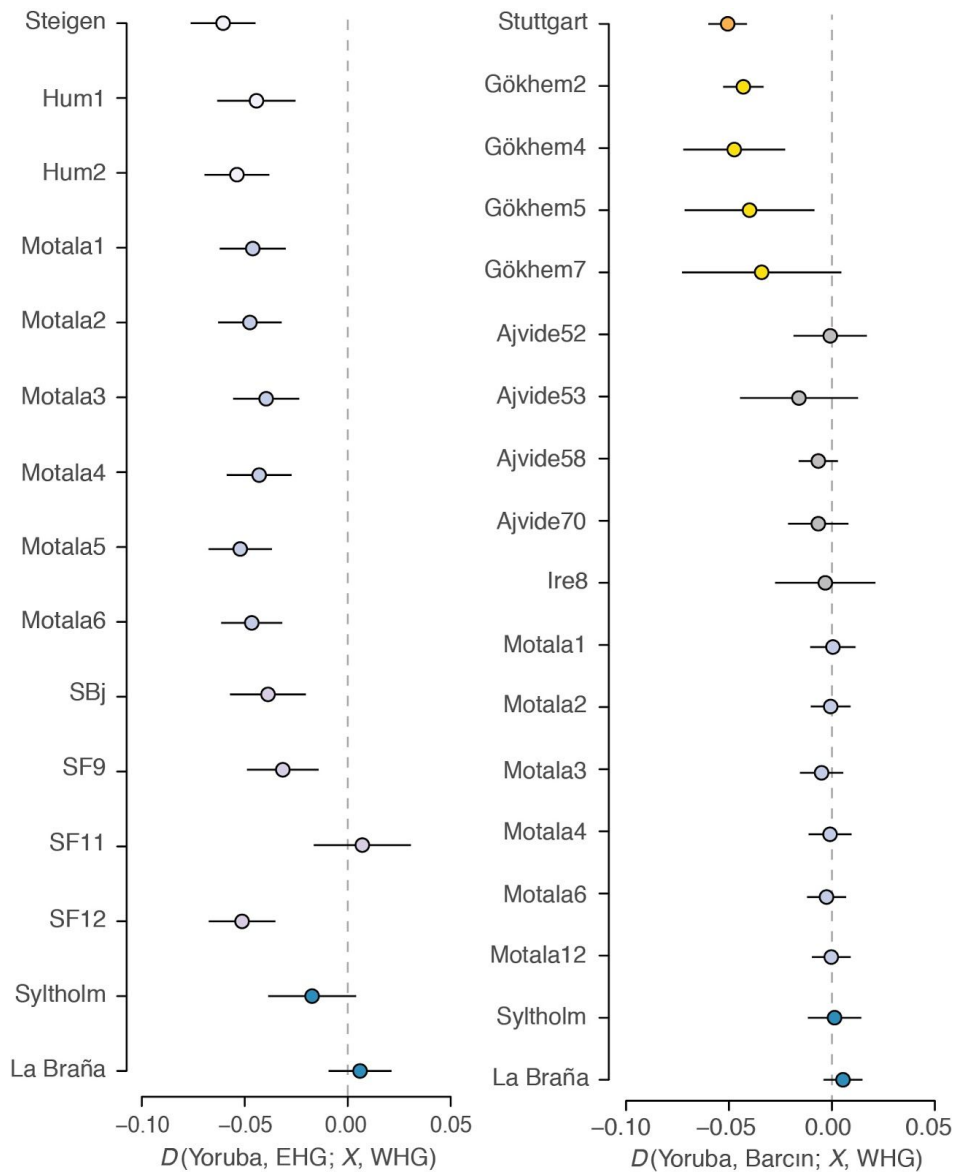
Supplementary Figure 4. FT-IR spectra of the Syltholm pitch and a modern birch sample.



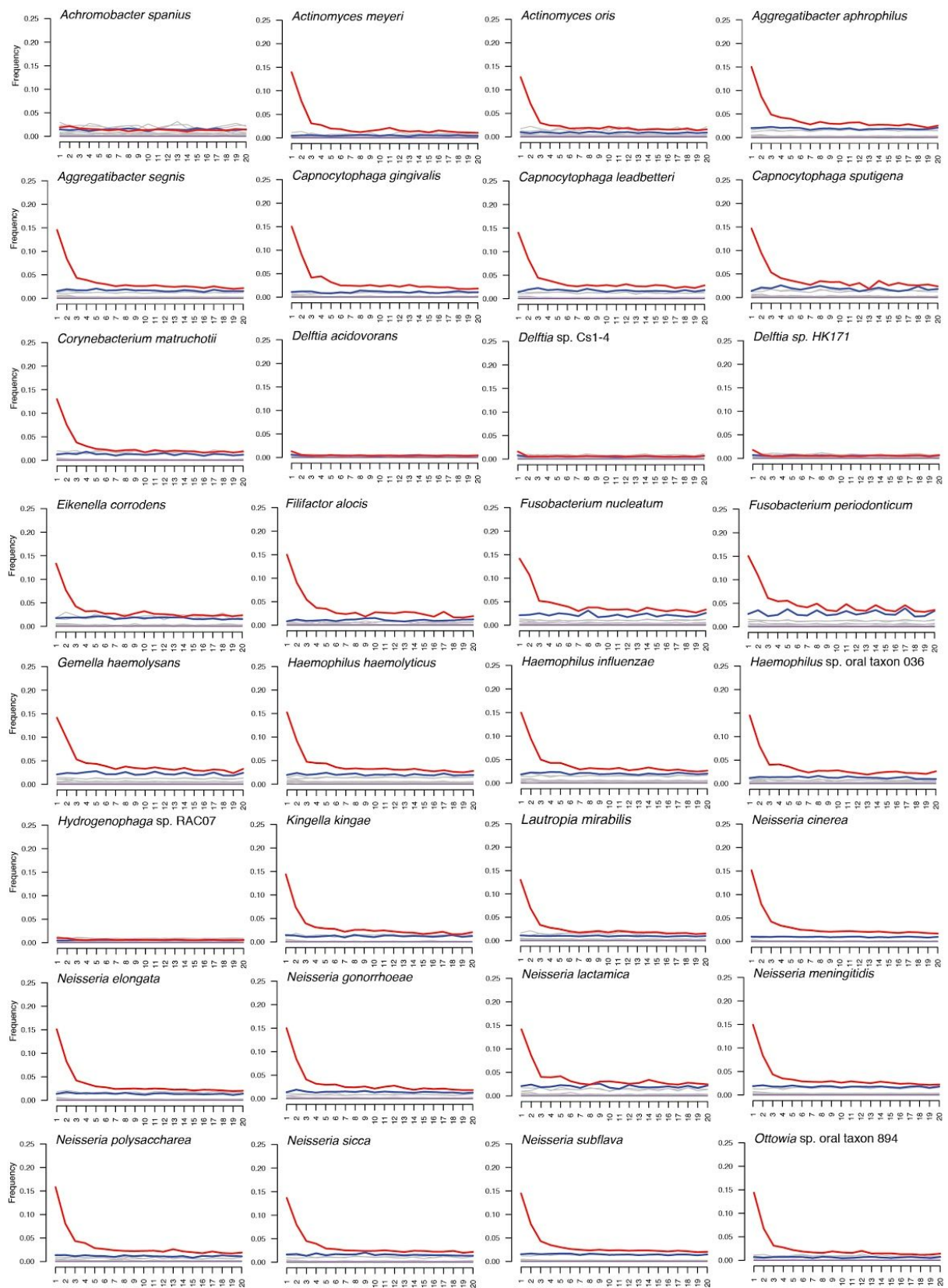
Supplementary Figure 5. GC-MS chromatograms of the Syltholm sample (back), betulin reference (middle) and lupeol reference (front).



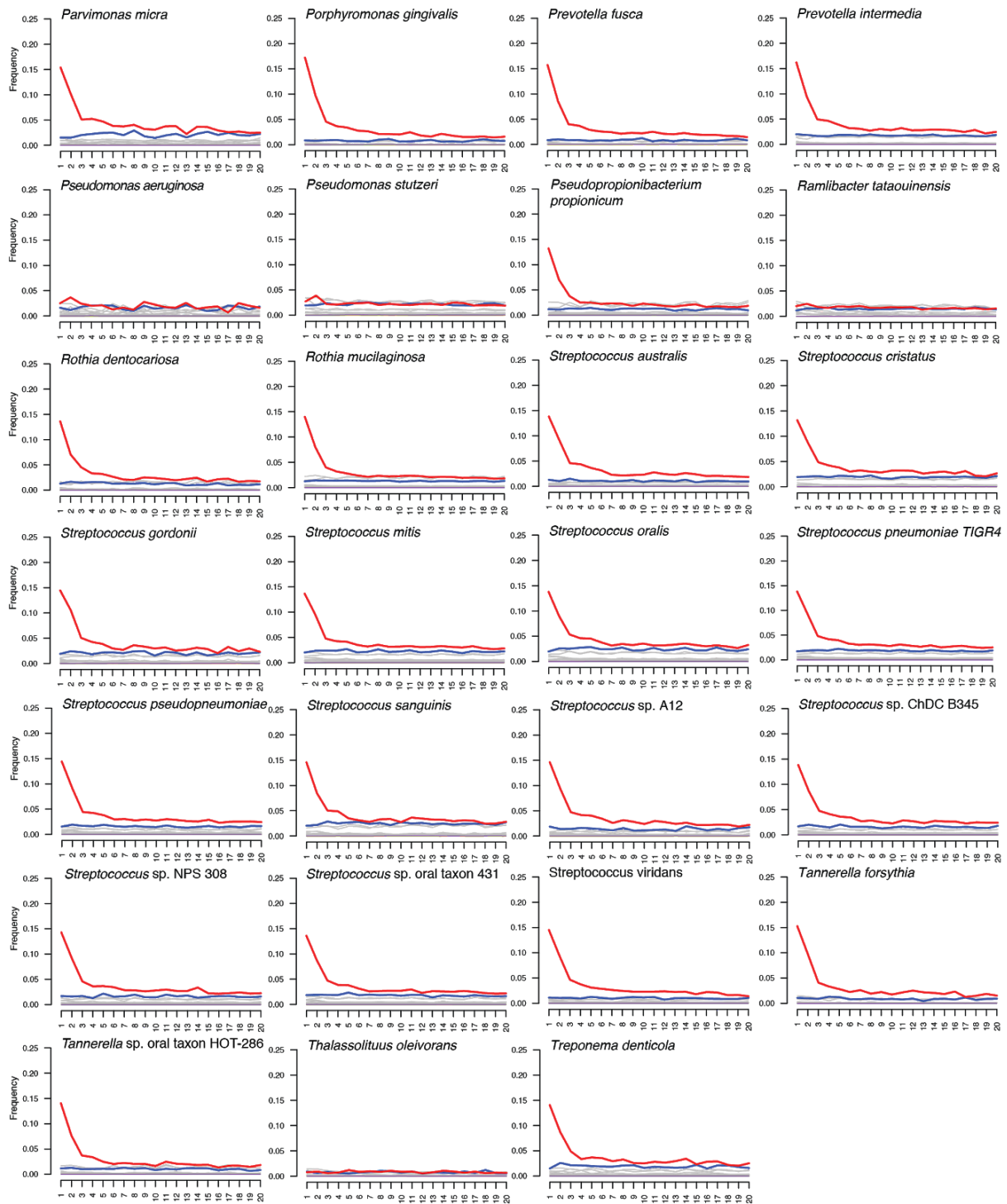
Supplementary Figure 6. MapDamage²³ plots for reads mapping to the human reference genome (hg19), by library.



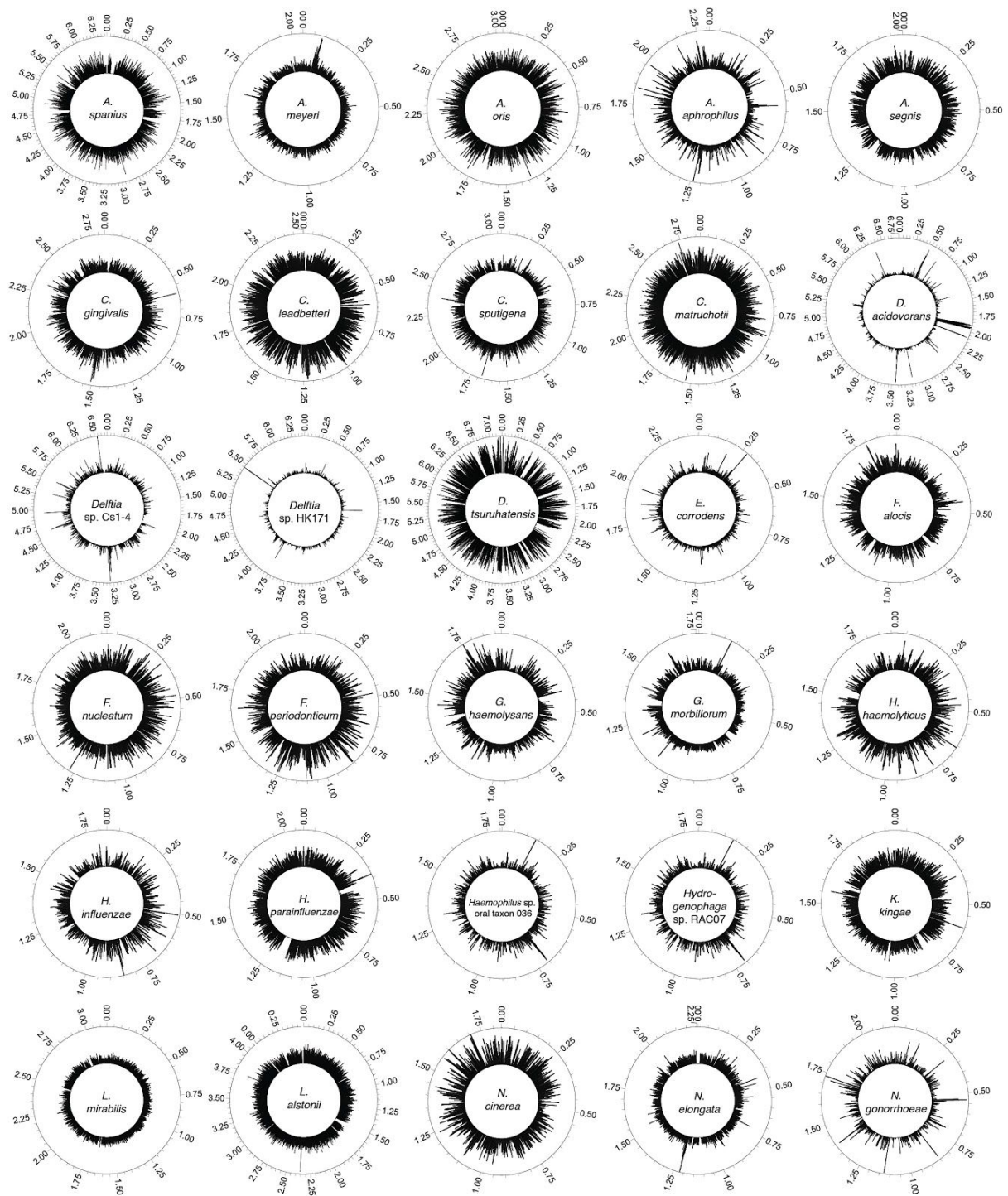
Supplementary Figure 7. D -statistics of the form $D(\text{Yoruba, EHG/Barcin}; X, \text{WHG})$ testing whether “ X ” forms a clade with WHG to the exclusion of EHG and Neolithic farmers (represented by Barcin), respectively. Error bars show three block-jackknife standard errors. Data are shown in Supplementary Tables 7 and 8.



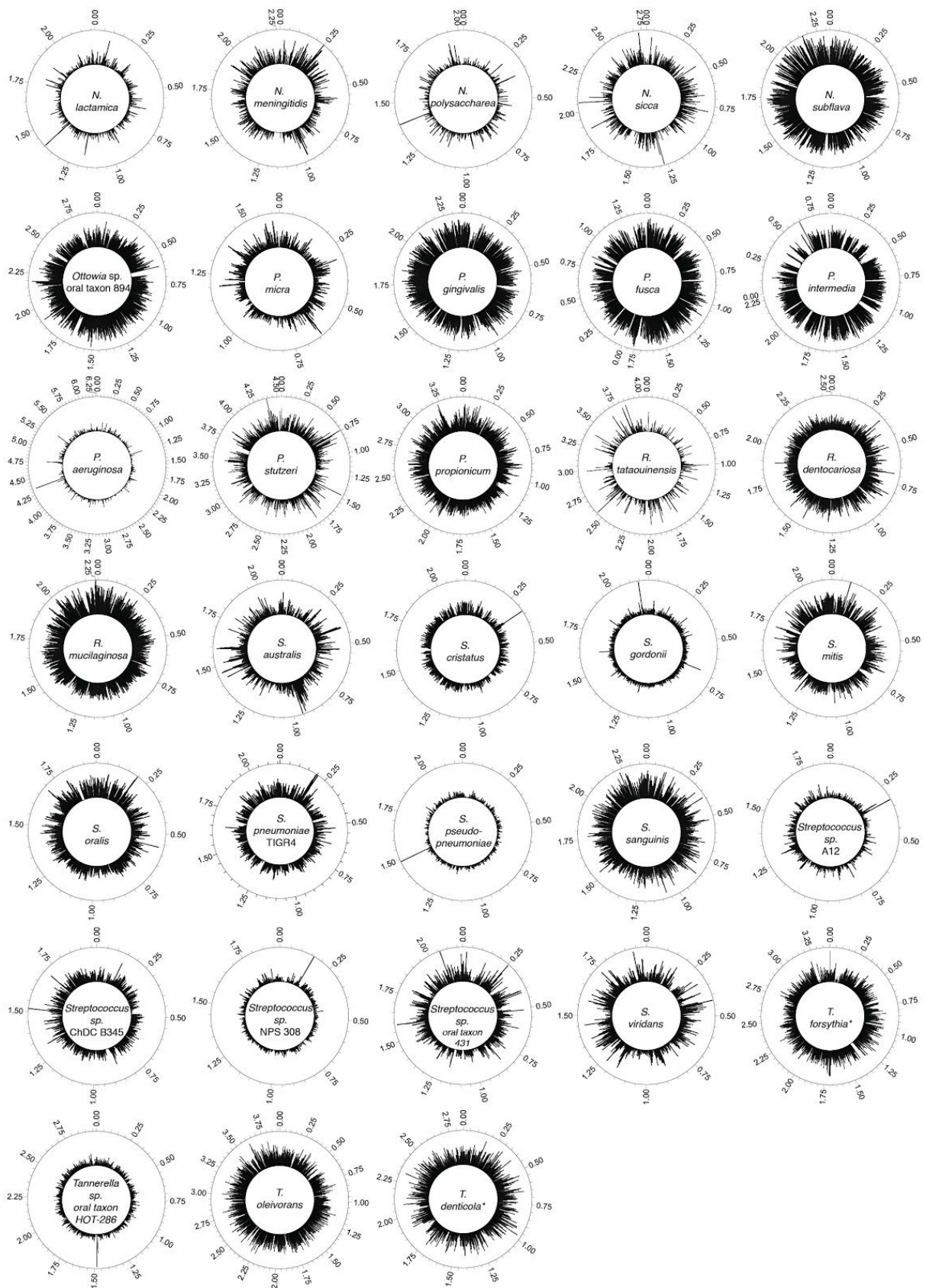
Supplementary Figure 8. MapDamage²³ plots for bacterial taxa with >10,000 assigned reads recovered from the Syltholm pitch.



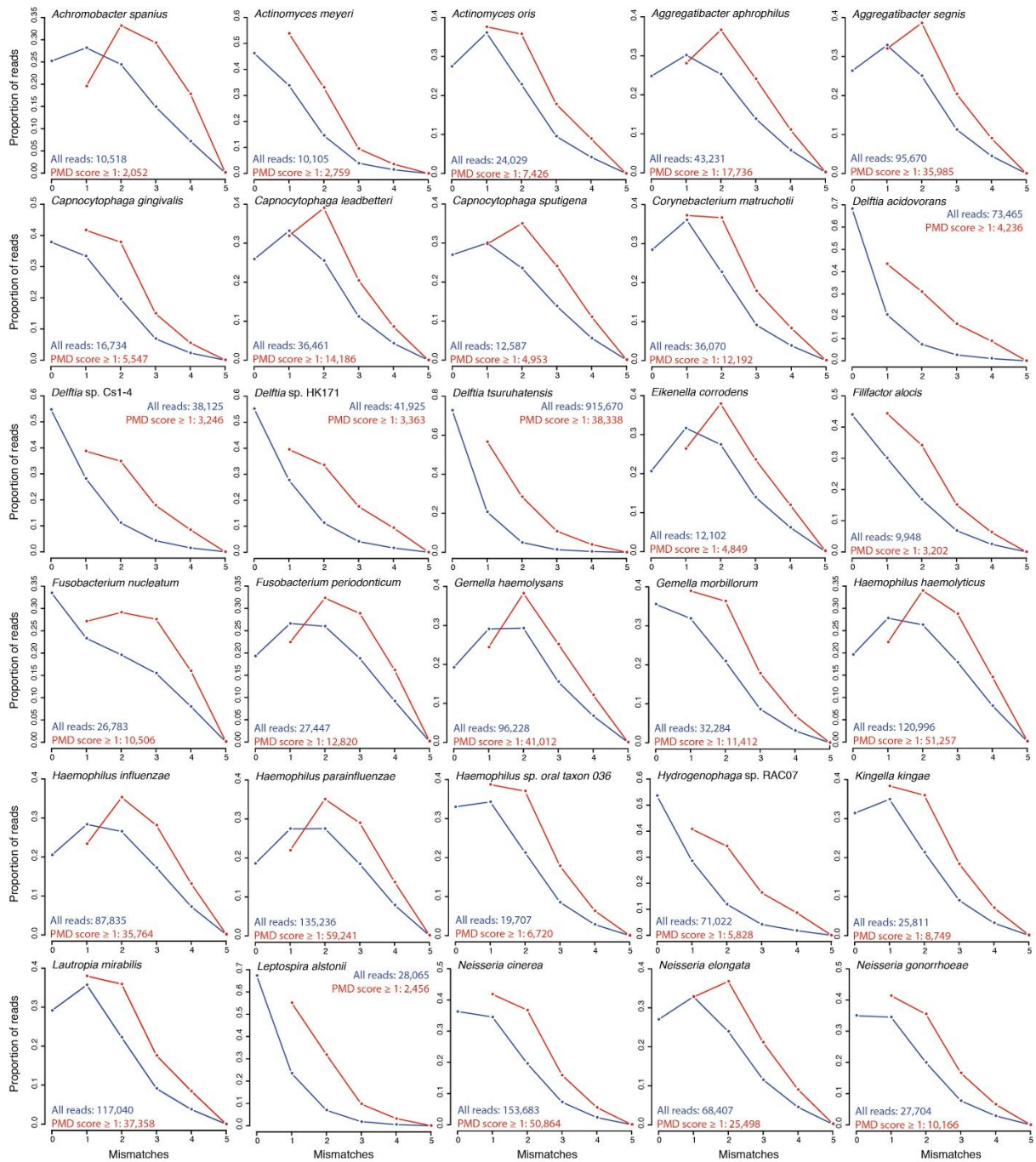
Supplementary Figure 8 ctd. MapDamage²³ plots for bacterial taxa with >10,000 assigned reads recovered from the Syltholm pitch.



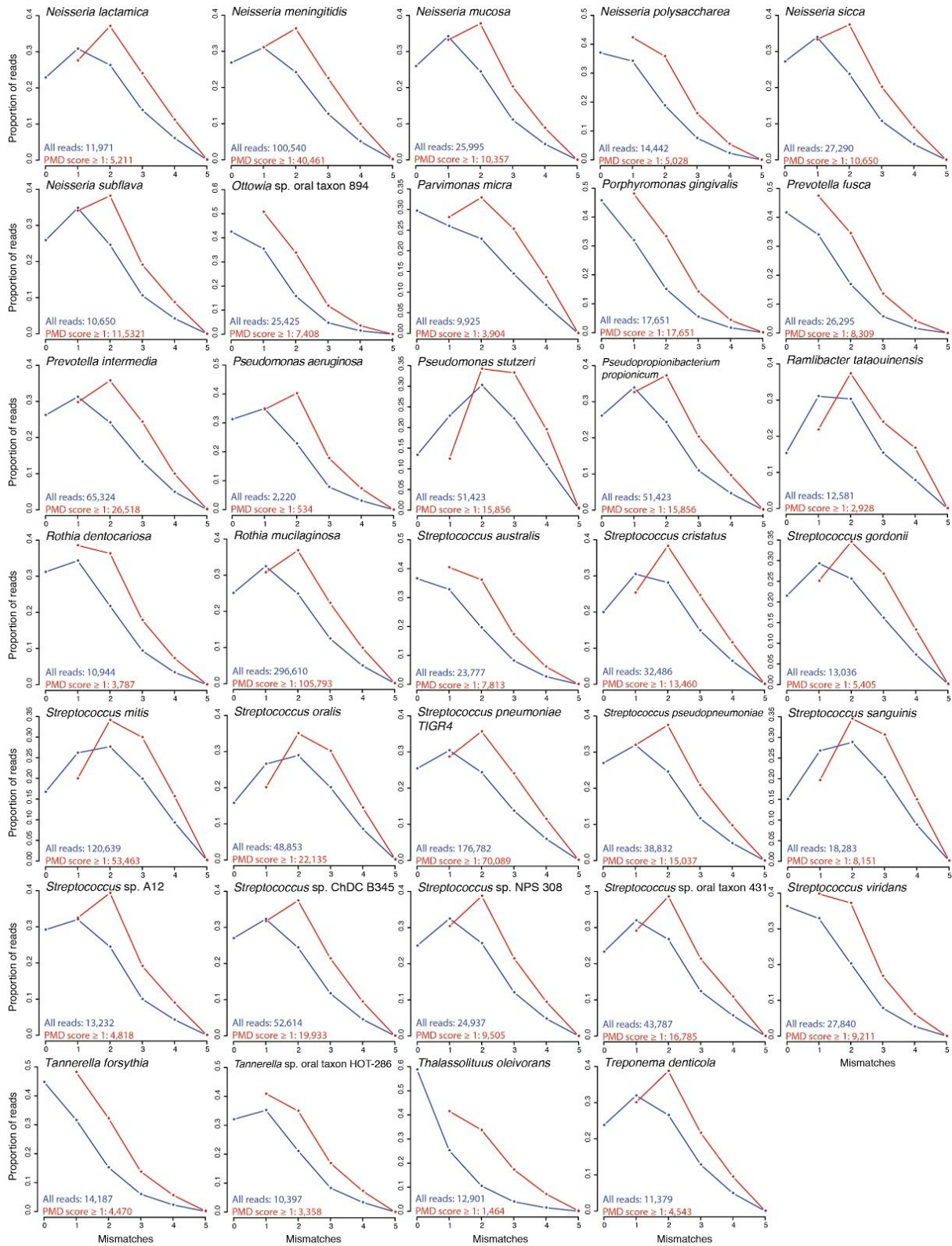
Supplementary Figure 9. Coverage plots for bacterial taxa recovered from the Syltholm pitch.



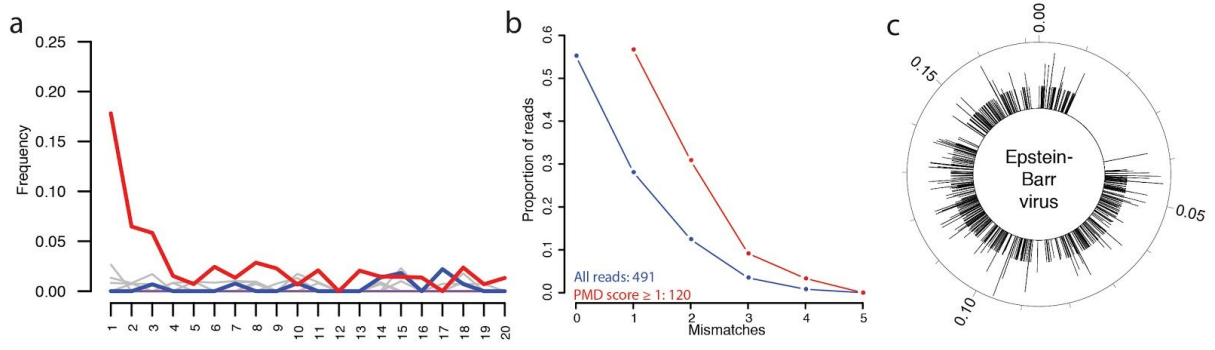
Supplementary Figure 9 ctd. Coverage plots for bacterial taxa recovered from the Syltholm pitch.



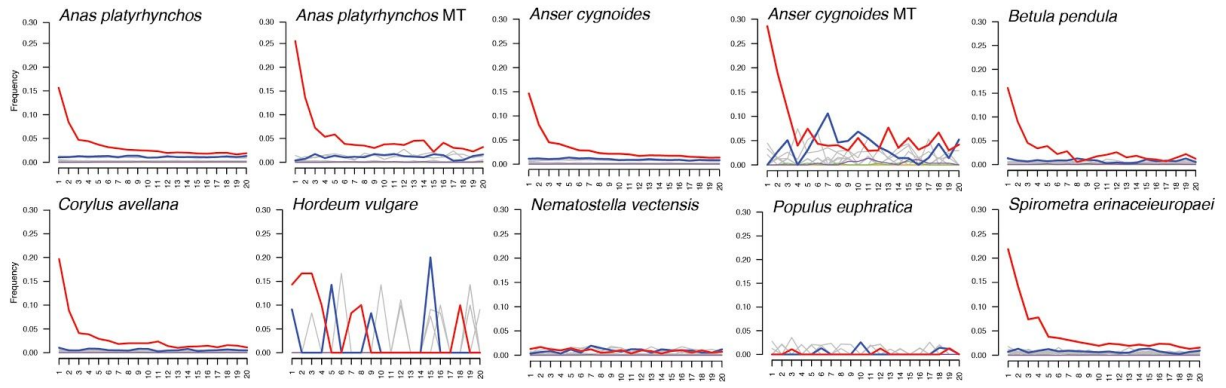
Supplementary Figure 10. Edit distance distributions of all reads (blue) and reads filtered for post-mortem damage (PMD \geq 1) (red) for bacterial taxa with >10,000 assigned reads recovered from the Syltholm pitch.



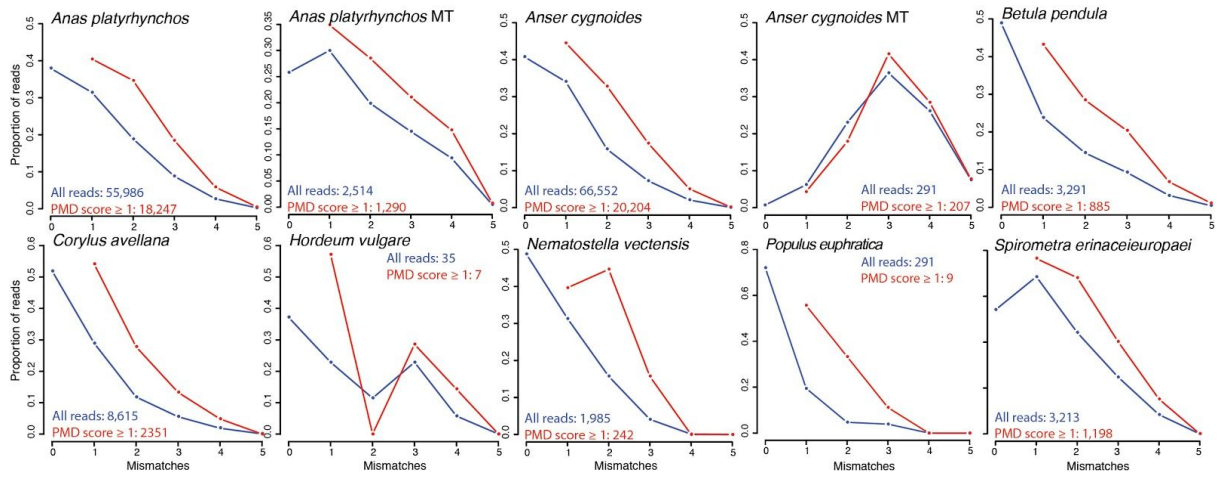
Supplementary Figure 10 ctd. Edit distance distributions of all reads (blue) and reads filtered for post-mortem damage (PMD \geq 1) (red) for bacterial taxa with >10,000 assigned reads recovered from the Syltholm pitch.



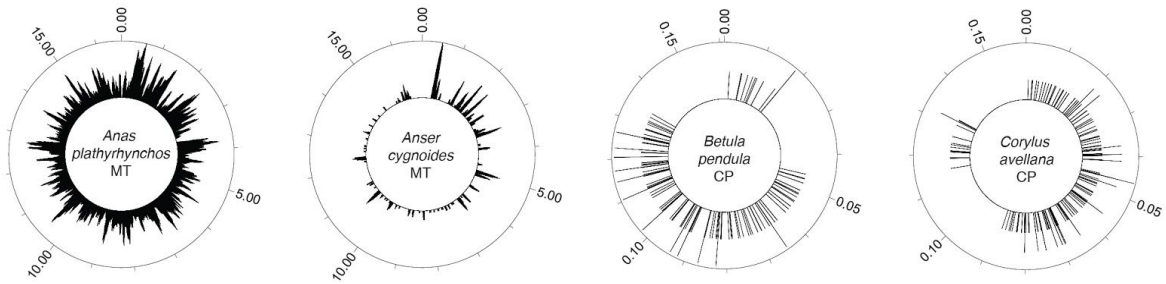
Supplementary Figure 11. MapDamage²³ plot (a), edit distance distribution (b), and coverage plot (c) for reads mapping to Epstein-Barr virus.



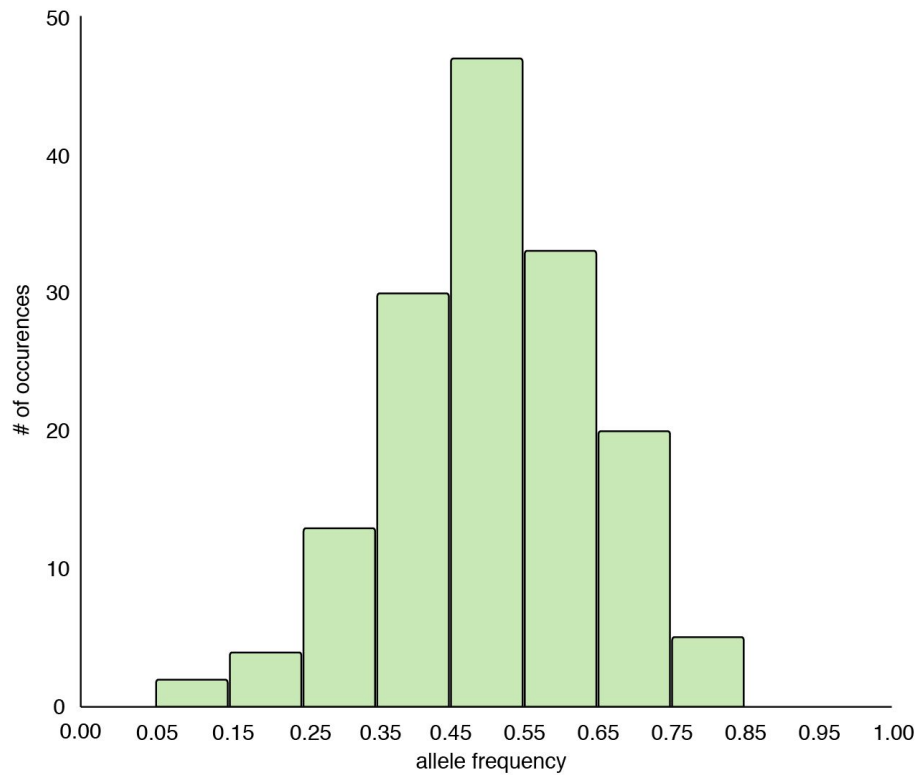
Supplementary Figure 12. MapDamage²³ plots for reads mapping to Metazoa (animals) and Viridiplantae (plants) in the ancient pitch sample. Note the absence of characteristic ancient DNA damage patterns for poplar (*Populus euphratica*), starlet sea anemone (*Nematostella vectensis*) and barley (*Hordeum vulgare*).



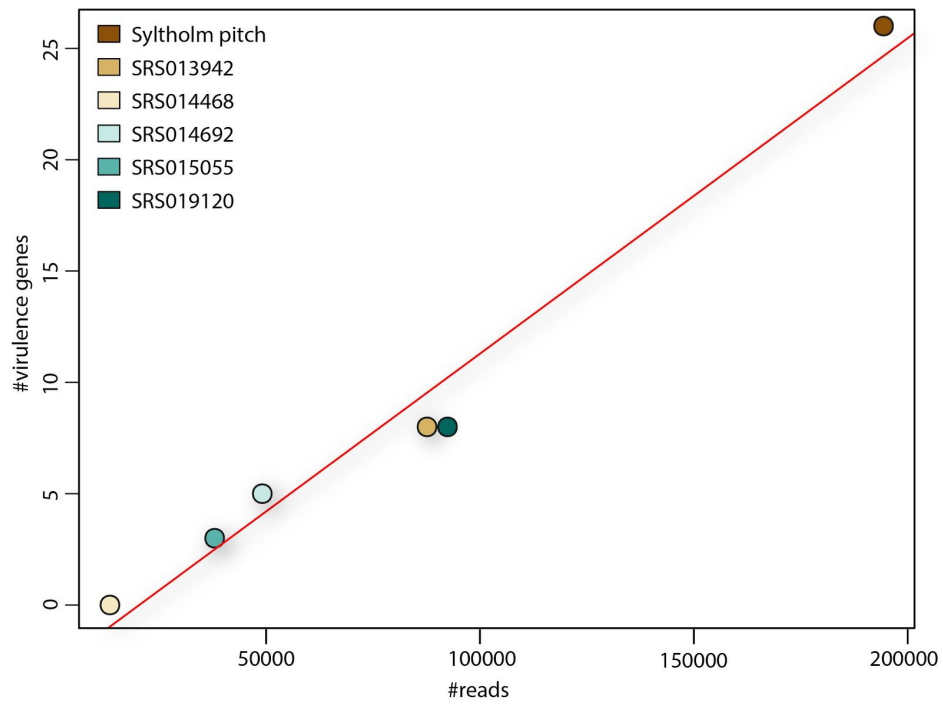
Supplementary Figure 13. Edit distance distributions of all reads from the ancient pitch assigned to Metazoa (animals) and Viridiplantae (plants). Reads filtered for post-mortem damage ($\text{PMD} \geq 1$) are shown in red.



Supplementary Figure 14. Coverage plots for eukaryotic taxa in the ancient pitch sample with more than 100 reads aligning to the chloroplast/mitochondrial genome. For poplar (*Populus euphratica*), starlet sea anemone (*Nematostella vectensis*) and the tapeworm (*Spirometra erinaceieuropaei*) no fragment aligned to its cpDNA or mtDNA, while for barley (*Hordeum vulgare*) only 15 fragments aligned to its cpDNA. The gaps in the chloroplast DNA (cpDNA) represent inverted repeats, which are very similar to each other, although not completely identical.



Supplementary Figure 15. Allele frequency distribution of single nucleotide variants in the 10× mallard (*A. platyrhynchos*) mtDNA genome recovered from the ancient pitch. The symmetric distribution suggests the presence of two haplotypes present in equal abundance.



Supplementary Figure 16. The number of virulence genes identified in the ancient pitch sample and five human oral microbiome samples from the HMP³¹.

Supplementary Table 1. Screening results for six different DNA extracts from the Syltholm pitch, extraction blank, and two soil control samples from the site.

| sample | weight | method | yield (ng)¹ | hg19 reads² | % dupl.³ | % end.⁴ | fragment length⁵ | C-T 5' (%)⁶ |
|---------------|---------------|---------------|-------------------------------|-------------------------------|----------------------------|---------------------------|------------------------------------|-------------------------------|
| 1 | 54 mg | 1 | 4.7 | 449,096 | 9.1 | 3.7 | 56.1 | 17.4 |
| 2 | 52 mg | 1 | 17.3 | 2,189,982 | 44.1 | 24.1 | 55.4 | 10.4 |
| 3 | 44 mg | 2 | 6.8 | 2,754,931 | 6.3 | 56.5 | 59.9 | 15.0 |
| 4 | 48 mg | 2 | 6.1 | 3,895,487 | 9.0 | 55.1 | 59.8 | 17.0 |
| 5 | 32 mg | 3 | 0.3 | 63,390 | 57.8 | 1.6 | 62.3 | 19.3 |
| 6 | 24 mg | 3 | 0.3 | 144,681 | 46.5 | 4.3 | 64.5 | 18.5 |
| Control 1 | ~2 g | 2 | 61.4 | 450 | 46.4 | <0.1 | 50.8 | 18.8 |
| Control 2 | ~2 g | 2 | 58.2 | 401 | 50.2 | <0.1 | 46.0 | 8.3 |
| NTC | N/A | 2 | N/A | 140 | 72.0 | 0.5 | 57.1 | 13.9 |

¹Total DNA yields (ng) measured using the Agilent 4200 TapeStation; ²Number of reads that could be uniquely mapped to the human reference genome (hg19) after removing duplicates; ³Fraction of duplicate reads in the sample (in percent); ⁴Endogenous human DNA content (in percent); ⁵Average fragment length (in bp); ⁶Deamination rate at 5' ends of DNA fragments (in percent)

Supplementary Table 2. Deep-sequencing results for the Syltholm pitch.

| hg19 reads¹ | end. content² | fragment length³ | C-T 5'⁴ | mtDNA contamination⁵ | >1X⁶ | DoC⁷ | mtDNA hg⁸ |
|-------------------------------|---------------------------------|------------------------------------|---------------------------|--|---------------------------|------------------------|-----------------------------|
| 120,585,267 | 31.2% | 59.9 bp | 16.2% | 1-3% | 78.9% | 2.3x | K1e |

¹Number of reads that could be uniquely mapped to the human reference genome (hg19) after removing duplicates and filtering for mapping quality (MAPQ \geq 30); ²Endogenous human DNA content (in percent); ³Average fragment length (in bp); ⁴Deamination rates at 5' ends of DNA fragments (in percent); ⁵MtDNA based contamination estimates determined using Schmutzi³⁴; ⁶Genome coverage (in percent); ⁷Average depth of genome coverage; ⁸Mitochondrial DNA haplogroup.

Supplementary Table 3. Molecular decay rates (k , per site per year) for the Syltholm genome and other previously published ancient genomes from different contexts^{11,12,32,33}.

| Sample | Age (yrs BP) | Temp. (°C) | λ | k | k , 100 bp | half-life (yrs), 100 bp |
|---------------------|---------------------|-------------------|-----------|--------------|--------------|--------------------------------|
| Taino (The Bahamas) | 1,000 | 20 | 0.016 | 1.60^{-05} | 1.60^{-03} | 434 |
| Syltholm (Denmark) | 5,700 | 8.5 | 0.034 | 5.96^{-06} | 5.96^{-04} | 1,162 |
| La Braña (Spain) | 7,500 | 8.1 | 0.033 | 4.40^{-06} | 4.40^{-04} | 1,576 |
| Kennewick (WA, USA) | 9,000 | 12.5 | 0.017 | 1.89^{-06} | 1.89^{-04} | 3,670 |
| Anzick (MT, USA) | 12,785 | 4.8 | 0.018 | 1.41^{-06} | 1.41^{-04} | 4,916 |

Supplementary Table 4. F -statistics of the form $f_4(\text{Yoruba}, X; \text{EHG}, \text{WHG})$ measuring the amount of shared genetic drift between different ancient genomes (X), EHG and WHG.

| Pop2 (X) | f_4-stat | SE | Z | BABA | ABBA | SNPs |
|-----------------|------------------------------|-----------|----------|-------------|-------------|-------------|
| Syltholm | 0.011917 | 0.000698 | 17.063 | 6,281 | 4,903 | 115,687 |
| La Braña | 0.012022 | 0.000525 | 22.894 | 29,695 | 23,219 | 538,716 |
| Hum1 | -0.001431 | 0.000644 | -2.224 | 9,966 | 10,262 | 207,167 |
| Hum2 | -0.001152 | 0.000592 | -1.947 | 26,029 | 26,646 | 536,119 |
| Steigen | -0.001494 | 0.000565 | -2.645 | 20,418 | 21,047 | 421,170 |
| Motala1 | 0.00216 | 0.000575 | 3.755 | 17,719 | 16,950 | 355,954 |
| Motala2 | 0.003681 | 0.000546 | 6.747 | 22,397 | 20,771 | 441,690 |
| Motala3 | 0.002856 | 0.000529 | 5.396 | 13,191 | 12,427 | 267,396 |
| Motala4 | 0.003171 | 0.000578 | 5.484 | 22,361 | 20,955 | 443,456 |
| Motala6 | 0.002229 | 0.000554 | 4.023 | 18,922 | 18,073 | 380,891 |
| Motala12 | 0.002848 | 0.000545 | 5.223 | 25,448 | 24,005 | 506,761 |

Supplementary Table 5. F -statistics of the form $f_4(\text{Yoruba}, X; \text{NEO}, \text{WHG})$ measuring the amount of shared genetic drift between different ancient genomes (X), WHG, and Neolithic farmers (represented by Barcin).

| Pop2 (X) | f_4-stat | SE | Z | BABA | ABBA | SNPs |
|-----------------|------------------------------|-----------|----------|-------------|-------------|-------------|
| Syltholm | 0.019419 | 0.000586 | 33.11 | 7,292 | 4,941 | 121,065 |
| La Braña | 0.017952 | 0.000436 | 41.145 | 34,006 | 23,859 | 565,167 |
| Motala1 | 0.012127 | 0.000476 | 25.488 | 20,388 | 16,010 | 361,017 |
| Motala2 | 0.013533 | 0.000425 | 31.864 | 25,789 | 19,700 | 449,904 |
| Motala3 | 0.011315 | 0.00042 | 26.961 | 14,803 | 11,765 | 268,527 |
| Motala4 | 0.012719 | 0.000444 | 28.662 | 25,548 | 19,828 | 449,704 |
| Motala6 | 0.012387 | 0.000425 | 29.156 | 21,691 | 16,930 | 384,398 |
| Motala12 | 0.012751 | 0.000426 | 29.936 | 29,548 | 22,926 | 519,374 |
| Ajvide52 | 0.010292 | 0.0007 | 14.711 | 3,011 | 2,450 | 54,498 |
| Ajvide53 | 0.009491 | 0.001117 | 8.5 | 920 | 764 | 16,417 |
| Ajvide58 | 0.009906 | 0.000441 | 22.476 | 29,847 | 24,550 | 534,726 |
| Ajvide70 | 0.01018 | 0.00057 | 17.861 | 5,297 | 4,330 | 95,022 |
| Ire8 | 0.009695 | 0.000916 | 10.588 | 1,315 | 1,082 | 23,981 |
| Gökhem2 | 0.000697 | 0.000428 | 1.629 | 21,220 | 20,929 | 418,556 |
| Gökhem4 | 0.001914 | 0.000943 | 2.029 | 1,078 | 1,038 | 20,804 |
| Gökhem5 | 0.001455 | 0.001157 | 1.258 | 704 | 685 | 13,614 |
| Gökhem7 | 0.00348 | 0.001569 | 2.218 | 378 | 351 | 7,621 |
| Stuttgart | -0.004073 | 0.000368 | -11.065 | 26,734 | 29,024 | 562,246 |

Supplementary Table 6. Admixture proportions based on *qpAdm*³⁵ analysis, specifying western hunter-gatherers (WHG), eastern hunter-gatherers (EHG), and Neolithic farmers (Barcin) as ancestral source populations.

| test population | reference population | admixture proportion | n SNPs | chi square | tail prob |
|-----------------|----------------------|----------------------|---------|------------|-----------|
| Bichon (LP) | WHG | 1.000 | 374,266 | 3.52 | 0.74 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| Rochedane (LP) | WHG | 1.000 | 113,744 | 6.72 | 0.35 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| La Braña (M) | WHG | 1.000 | 538,715 | 7.15 | 0.31 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| Loschbour (M) | WHG | 1.000 | 544,933 | 9.79 | 0.13 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| Ranchot (M) | WHG | 1.000 | 200,185 | 4.02 | 0.67 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| Syltholm | WHG | 1.000 | 115,800 | 6.34 | 0.39 |
| | EHG | 0.000 | | | |
| | Barcin | 0.000 | | | |
| Karelia (M) | WHG | 0.000 | 294,370 | 11.15 | 0.08 |
| | EHG | 1.000 | | | |
| | Barcin | 0.000 | | | |
| Samara (M) | WHG | 0.000 | 294,370 | 11.15 | 0.08 |
| | EHG | 0.100 | | | |
| | NF | 0.000 | | | |
| | Barcin | 0.441 | | | |
| NorwayHG (M) | WHG | 0.697 | 558,124 | 3.10 | 0.68 |
| | EHG | 0.559 | | | |
| | Barcin | 0.000 | | | |
| Latvia (M) | WHG | 0.649 | 560,151 | 4.49 | 0.48 |
| | EHG | 0.303 | | | |
| | Barcin | 0.000 | | | |
| BalticHG (M) | WHG | 0.593 | 562,935 | 3.46 | 0.63 |
| | EHG | 0.351 | | | |
| | Barcin | 0.000 | | | |
| Motala (M) | WHG | 0.780 | 545,689 | 4.83 | 0.44 |
| | EHG | 0.407 | | | |
| | Barcin | 0.000 | | | |
| PWC (EN) | WHG | 0.175 | 523,969 | 3.14 | 0.68 |
| | EHG | 0.220 | | | |
| | Barcin | 0.000 | | | |
| Gökhem (EN) | WHG | 0.180 | 407,865 | 2.68 | 0.75 |
| | EHG | 0.000 | | | |
| | Barcin | 0.825 | | | |
| Iberia (EN) | WHG | 0.162 | 557,569 | 2.98 | 0.70 |
| | EHG | 0.000 | | | |
| | Barcin | 0.820 | | | |
| LBK (EN) | WHG | 0.293 | 563,150 | 2.61 | 0.76 |
| | EHG | 0.000 | | | |
| | Barcin | 0.838 | | | |
| GAC (EN) | WHG | 0.293 | 563,197 | 9.72 | 0.08 |
| | EHG | 0.000 | | | |
| | Barcin | 0.707 | | | |

Supplementary Table 7. *D*-statistics of the form $D(\text{Yoruba, EHG}; X, \text{WHG})$ testing whether “*X*” forms a clade with WHG to the exclusion of EHG.

| Pop3 (X) | <i>D</i>-stat | SE | Z | BABA | ABBA | SNPs |
|-----------------|----------------------|-----------|----------|-------------|-------------|-------------|
| Sylthom | -0.0173 | 0.007118 | -2.432 | 4,736 | 4,903 | 115,687 |
| La Braña | 0.006 | 0.005081 | 1.176 | 23,498 | 23,219 | 538,716 |
| Motala1 | -0.0461 | 0.005356 | -8.611 | 15,456 | 16,950 | 355,954 |
| Motala2 | -0.0475 | 0.005144 | -9.232 | 18,888 | 20,771 | 441,690 |
| Motala3 | -0.0396 | 0.005355 | -7.387 | 11,481 | 12,427 | 267,396 |
| Motala4 | -0.043 | 0.005249 | -8.19 | 19,228 | 20,955 | 443,456 |
| Motala6 | -0.0522 | 0.005122 | -10.192 | 16,280 | 18,073 | 380,891 |
| Motala12 | -0.0466 | 0.004941 | -9.427 | 21,868 | 24,005 | 506,761 |
| SBj | -0.0387 | 0.006137 | -6.308 | 7,622 | 8,236 | 174,952 |
| SF9 | -0.0315 | 0.005791 | -5.44 | 12,895 | 13,734 | 293,510 |
| SF11 | 0.0071 | 0.007857 | 0.91 | 3,261 | 3,214 | 69,375 |
| SF12 | -0.0513 | 0.005394 | -9.516 | 24,421 | 27,064 | 561,611 |
| Hum1 | -0.0443 | 0.006341 | -6.99 | 9,391 | 10,262 | 207,167 |
| Hum2 | -0.0538 | 0.005228 | -10.299 | 23,924 | 26,646 | 536,119 |
| Steigen | -0.0605 | 0.00524 | -11.544 | 18,646 | 21,047 | 421,170 |

Supplementary Table 8. *D*-statistics of the form $D(\text{Yoruba}, \text{Barcin}; X, \text{WHG})$ testing whether “*X*” forms a clade with WHG to the exclusion of Neolithic farmers (represented by Barcin).

| Pop3 (X) | <i>D</i>-stat | SE | Z | BABA | ABBA | SNPs |
|-----------------|----------------------|-----------|----------|-------------|-------------|-------------|
| Syltholm | 0.0013 | 0.004313 | 0.307 | 4,954 | 4,941 | 121,065 |
| La Braña | 0.0054 | 0.003145 | 1.716 | 24,118 | 23,859 | 565,167 |
| Ajvide52 | -0.0008 | 0.005938 | -0.128 | 2,447 | 2,450 | 54,498 |
| Ajvide53 | -0.016 | 0.009548 | -1.673 | 740 | 764 | 16,417 |
| Ajvide58 | -0.0066 | 0.003166 | -2.08 | 24,229 | 24,550 | 534,726 |
| Ajvide70 | -0.0066 | 0.004872 | -1.361 | 4,273 | 4,330 | 95,022 |
| Ire8 | -0.0032 | 0.00813 | -0.395 | 1,075 | 1082 | 23,981 |
| Gökhem2 | -0.043 | 0.003261 | -13.183 | 19,203 | 20,929 | 418,556 |
| Gökhem4 | -0.0474 | 0.008241 | -5.749 | 944 | 1,038 | 20,804 |
| Gökhem5 | -0.04 | 0.010509 | -3.809 | 632 | 685 | 13,614 |
| Gökhem7 | -0.0341 | 0.012892 | -2.648 | 328 | 351 | 7,621 |
| Motala1 | 0.0005 | 0.003661 | 0.145 | 16,027 | 16,010 | 361,017 |
| Motala2 | -0.0006 | 0.00322 | -0.19 | 19,676 | 19,700 | 449,904 |
| Motala3 | -0.005 | 0.003485 | -1.445 | 11,647 | 11,765 | 268,527 |
| Motala4 | -0.0009 | 0.003462 | -0.255 | 19,793 | 19,828 | 449,704 |
| Motala6 | -0.0026 | 0.003151 | -0.833 | 16,841 | 16,930 | 384,398 |
| Motala12 | -0.0003 | 0.003127 | -0.093 | 22,912 | 22,926 | 519,374 |
| Stuttgart | -0.0506 | 0.003118 | -16.231 | 26,228 | 29,024 | 562,246 |

References

1. Mortensen, M. F. *et al.* Fortidens spor og fremtidens forbindelse - bevaring og naturvidenskab på Femern Bælt projektet, Danmarks største arkæologiske udgravning. *Nationalmuseets Arbejdsmark* 22–36 (2015).
2. Groß, D. *et al.* People, lakes and seashores: Studies from the Baltic Sea basin and adjacent areas in the early and Mid-Holocene. *Quat. Sci. Rev.* **185**, 27–40 (2018).
3. Jensen, L. E. *et al.* Syltholmudgravningerne - jagten på stenalderens jægere, fiskere og bønder i et druknet landskab. *Aarbøger for nordisk Oldkyndighed og Historie* (Copenhagen, 2018).
4. Sørensen, S. A. Danmarks største stenalderudgravning – Fantastiske fund fra Femernudgravningerne. *FUND & FORTID* **2**, 17–24 (2018).
5. Sørensen, S. A. Syltholm: Denmark’s largest Stone Age excavation. *Mesolithic Miscellany* **24**, 3–10 (2016).
6. Brock, F., Higham, T., Ditchfield, P. & Ramsey, C. B. Current Pretreatment Methods for AMS Radiocarbon Dating at the Oxford Radiocarbon Accelerator Unit (Orau). *Radiocarbon* **52**, 103–112 (2010).
7. Ramsey, C. B. Radiocarbon Calibration and Analysis of Stratigraphy: The OxCal Program. *Radiocarbon* **37**, 425–430 (1995).
8. Mills, J. J. S. & White, R. *The Organic Chemistry of Museum Objects*. (Butterworth-Heinemann, 1987).
9. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proc Biol Sci.* *279(1748):4724-33* (2012).
10. Deagle, B. E., Eveson, J. P. & Jarman, S. N. Quantification of damage in DNA recovered from highly degraded samples—a case study on DNA in faeces. *Front. Zool.* **3**, 11 (2006).
11. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507**, 225–228 (2014).
12. Schroeder, H. *et al.* Origins and genetic legacies of the Caribbean Taino. *Proc. Natl. Acad. Sci. U. S. A.* **22**, 201716839–201716836 (2018).
13. Warinner, C. *et al.* A Robust Framework for Microbial Archaeology. *Annu. Rev. Genomics Hum. Genet.* **18**, 321–356 (2017).
14. Kryukov, K. & Imanishi, T. Human Contamination in Public Genome Assemblies. *PLoS One* **11**, e0162424 (2016).
15. Coghlan, A., Gordon, D. & Berriman, M. Contamination screening of parasitic worm genome assemblies. Preprint at <https://protocolexchange.researchsquare.com/article/nprot-6669/v1>.
16. Pedersen, M. W. *et al.* Postglacial viability and colonization in North America’s ice-free corridor. *Nature* **537**, 45–49 (2016).
17. Vågene, Å. J. *et al.* Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat Ecol Evol* **2**, 520–528 (2018).
18. Key, F. M., Posth, C., Krause, J., Herbig, A. & Bos, K. I. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* **33**, 508–520 (2017).
19. Huebler, R., Key, F. M. M., Warinner, C., Bos, K. I. & Krause, J. HOPS: Automated

- detection and authentication of pathogen DNA in archaeological remains. Preprint at <https://www.biorxiv.org/content/10.1101/534198v2> (2018).
20. Truong, D. T. *et al.* MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
 21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
 22. Picard Tools - By Broad Institute. Available at: <http://broadinstitute.github.io/picard/>. (Accessed: 15th December 2018).
 23. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
 24. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
 25. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
 26. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2229–2234 (2014).
 27. Team, R. C. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing (2014).
 28. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
 29. Lu, J. & Salzberg, S. L. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* **14**, e1006277 (2018).
 30. Astrup, P. M. *Sea-level change in Mesolithic southern Scandinavia. Long- and short-term effects on society and the environment.* Jysk Arkæologisk Selskabs Skrifter **106** (Jutland Archaeological Society, 2018).
 31. The Human Microbiome Project Consortium *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
 32. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506**, 225–229 (2014).
 33. Rasmussen, M. *et al.* The ancestry and affiliations of Kennewick Man. *Nature* **523**, 455–458 (2015).
 34. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
 35. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).

Chapter 5

Conclusions

Scientific conclusions

During the course of my PhD I used a variety of bioinformatic approaches to analyse the DNA from substrates that differ significantly in age and provenance, covering some of the most common themes in current palaeogenomics. The following sections summarise some of the innovative aspects of each of the chapters.

Chapter 2 explores the enigmatic population history of the extinct Honshū wolves. We made the surprising discovery that the Honshū wolf genome closely resembles that of Late-Pleistocene Siberian wolves, a lineage that was believed to have gone extinct around the Pleistocene-Holocene transition. Instead, our results suggest that some Late-Pleistocene Siberian wolves survived as a relict population on the Japanese archipelago and that their descendants only went extinct approximately 100 years ago. We furthermore detected significant dog introgression in the Honshū wolf genome, it is unclear however how representative this is for Honshū wolves before the drastic population decline in the 19th century.

In Chapter 3 we investigated the microbial DNA preservation in fluid-preserved specimens. We were able to reconstruct the microbiome of one historical razorbill specimen, showcasing for the first time the yet untapped potential of analysing historical microbiomes from “wet” collections. Furthermore we identified fish DNA in one of the samples, indicating that it is possible to extract dietary DNA from fluid-preserved specimens. The DNA preservation of the microbiome of most of the specimens proved to be very poor however and the majority of the samples were heavily contaminated with unknown contamination origin, therefore further research is needed to maximise the potential of this substrate.

Chapter 4 combines the population genomic and metagenomic aspects of Chapter 2 and Chapter 3, where we analysed the DNA present in a 5,700 year old chewed birch bark pitch. The DNA proved very well preserved, allowing the identification of numerous oral microbes and eukaryotic taxa, which are in all likelihood derived from the mastic material itself or a recent meal, as well as the reconstruction of an ancient human herpesvirus 4. The human DNA also provided valuable insight into the population of Late Mesolithic/Early Neolithic Denmark,

as our results indicate that the female who chewed the birch bark pitch, who we dubbed Lola, only has Western hunter-gatherer ancestry. This suggests that during this time point of the Mesolithic-Neolithic transition neither Neolithic farmers nor Eastern hunter-gatherers had a significant genetic contribution to her lineage.

In summary, this dissertation has not only advanced the field of palaeogenomics with the new discoveries on Siberian Pleistocene and Honshū wolves, but also demonstrated that two novel materials are viable substrates for metagenomic analyses on the microbiome and diet of past organisms. I furthermore identified some key challenges of ancient and historical metagenomic analyses, such as contaminated reference genomes, laboratory contaminants, and false positive assignments. I therefore developed an authentication framework to validate the metagenomic assignments that can be employed in future studies on ancient microbiomes.

Challenges and future perspectives

Honshū wolf

One unexpected finding of the second chapter was the large contribution of Japanese dogs to the Honshū wolf genome. While this shed some light on the occurrence of hybridisations between the two populations, it also turned out to be highly problematic for the characterization of introgression from the Honshū wolf to the Japanese dog, as any shared genomic regions could either originate from the “pure” Honshū wolf or the dog heritage of the individual. This is not an unusual problem for genomic studies on dogs and wolves, where the distinctions between breeds and populations are especially complicated and poorly defined. Many of the samples in our reference panel were admixed with wolves, dogs, and coyotes from other populations, which complicated the interpretation of the results as it required a more in-depth knowledge of the genetic makeup of the individuals. For the haplotype-aware methods we therefore used the clustering based on the phased genotypes to define populations instead of relying on the location or breed information.

The discovery that the Honshū wolf is very closely related to Late Pleistocene Siberian wolves raises several questions that should be addressed in future studies. One yet unanswered question of Chapter 2 is the exact contribution of the Honshū wolves to modern Japanese dogs. Given the ancient legacy of the Honshū wolf, it would be of great interest to identify traits that Japanese dog breeds, such as the Akita, Shiba Inu, and Kishu, inherited from the unique Honshū wolves and therefore discern them from non-Japanese dog breeds. In order to identify such traits and describe past admixture events between Honshū wolves and Japanese dogs, a much larger reference panel of Honshū wolves, modern, and, if possible, ancient Japanese dogs is required.

Chapter 2 also makes the enigmatic wolf population that was domesticated, and that modern dogs descend from, a subject of discussion. While our results indicate that the Honshū wolves were probably not the ancestors of all modern dogs, another wolf subspecies that was endemic to Japan, the Ezo wolf, could be the answer to the long-standing mystery of the missing wolf population. We generated sequences of the Ezo wolf to test whether it could be ancestral to modern dogs and Eurasian wolves, but the genomic data generated, equivalent to a nuclear coverage of 0.4X, turned out to be too low to obtain conclusive results on its heritage and was therefore not included in the manuscript. However, we recently resequenced the sample,

thereby increasing the coverage significantly, and were furthermore fortunate enough to gain access to several additional Ezo wolf skulls for genomic studies, paving the way to gather exciting new insights into the wolf and dog evolution.

It is an exciting time to analyse ancient DNA datasets. The ground-breaking advances in sequencing technology result in an ever-growing number of publicly available ancient and historical genomes. Until recently, the study of ancient genomes were mostly limited to the sequenced individual itself, but the steadily increasing size of reference panels and therefore refined representation of ancient and historical communities enables a much more robust and in-depth analysis on population level with sometimes surprising outcomes, such as the strong Pleistocene heritage of the Honshū wolf and the pure Western hunter-gatherer ancestry of “Lola” during the period of Neolithization across Europe in this dissertation. The significant improvements of bioinformatic pipelines have been especially pivotal, and it can be expected that new breakthroughs in the coming years will see progress in areas that are still in their infancy, such as the accuracy and breadth of phenotype prediction.

Ancient and historical metagenomics

The field of ancient metagenomics is especially rapidly evolving, and the last years demonstrated the true value of analysing the metagenome of ancient and historical materials, such as the inference on past pandemics and foodstuffs. While dental calculus and coprolites are the most common substrates in ancient metagenomic studies, research on microbial and non-host eukaryotic aDNA extracted from alternative materials is still underdeveloped. This dissertation showcases the potential of using fluid-preserved specimens and birch bark pitch as substrates for ancient and historic metagenomic studies, yielding particular insights that are unobtainable with dental calculus, sediment, and coprolites. Dental calculus for example contains DNA from the oral microbiome and diet that accumulated in the plaque matrix over decades, while the metagenome of birch bark pitch resembles a snapshot of the DNA present in the mouth of the time when the pitch was chewed. Fluid-preserved specimens on the other hand present the unique opportunity to study the historical microbiome and diet from specimens that either do not preserve under any other conditions, such as soft-bodied organisms, or are only present in “wet” collections now as they have gone extinct in the wild.

It will be interesting to see what other materials are an unexpected source of metagenomic data and can thus be used to inform the lifestyle and health of organisms in the past.

Chewed birch bark pitch

Aside from the remarkable preservation of the human DNA, the chewed birch bark pitch also turned out to be a rich source of DNA with microbial and dietary origin. Highly uncommon for ancient metagenomic substrates, we could not authenticate taxa that are clearly derived from the environment and the microbial profile closely resembles those of human oral microbiomes, making it a very promising substrate for future metagenomic studies. Some laboratory reagent contaminants, however, proved to be in high abundance in the sequences, but could be quickly identified as non-endogenous due to the lack of DNA damage patterns. An unexpected consequence of the contamination with the laboratory reagent contaminant *Delftia spp.* was the large number of misassigned sequences to the Tibetan antelope (*Pantholops hodgsonii*) reference genome, which in turn turned out to be heavily contaminated with *Delftia spp.* DNA. In the same vein did we identify several scaffolds of the American bison (*Bison bison*) and the tapeworm *Spirometra erinaceieuropaei* reference genomes that were contaminated with human DNA, leading to thousands of sequences being incorrectly assigned to both species.

Disentangling the assignments of closely related species was a further obstacle we faced while interpreting our results. For instance, our findings paint a picture of a complex mixture of different Streptococci being present in the chewed birch bark pitch. The high homology among Streptococci however led to a large number of misassigned reads within the Streptococcus clade. The most probable cause is the relatively short read length combined with various other factors, for example that fast-evolving taxa are ancestral to multiple species in the database, have gone extinct, or are misassigned due to horizontal gene transfer or considerable changes of the genome with the passage of time. These issues are common for all ancient reference-based studies and will be difficult to overcome. Assembly-based approaches can aid in reconstructing the genomes of past organisms, but are often limited to relatively well preserved DNA data.

The exceptional DNA preservation and low abundance of contaminant taxa suggest that birch bark pitch is an excellent substrate for both human DNA and the ancient oral microbiome. However it remains to be seen whether this is characteristic for the material or is rather a fortunate consequence from external factors, such as the unique attributes of the archaeological site. I am part of further genomic studies on several more birch bark mastics, where we hope to recover well-preserved human as well as microbial and non-human eukaryotic DNA. Aside from conducting the analyses described in this dissertation, we will also introduce a

metaproteomic approach. Not only can this provide additional confirmation for some of the metagenomic assignments, but the identification of peptides can furthermore inform on the tissue origin of the eukaryotic DNA. In the context of our findings from Chapter 4, for instance, a metaproteomic approach could have been used to distinguish between mallard eggs and mallard meat as a source for the mallard DNA. Finally, the analysis of several samples from the same archaeological site furthermore allows kinship analyses as well as providing a more reliable inference on the lifestyle and ancestry of the community.

Fluid-preserved collections - a treasure trove for historical microbiomes?

Despite the fact that the seabirds analysed in Chapter 3 have lived and died thousands of years after “Lola” chewed the birch bark pitch, the DNA preservation transpired to be extremely poor for five out of the six fluid-preserved specimens. This greatly impeded the authentication of the microbial assignments, and the survival of gut microbial DNA could only be confirmed for one of the specimens - a razorbill that was collected in 1916. We were unable to identify the exact reasons for the disparities in DNA preservation, but it seems probable that the preservation practice and especially the use of preservatives has the largest effect on the DNA degradation.

Another marked difference between the results of Chapter 3 and Chapter 4 is the overwhelming dominance of one microbial species, *Catellibacterium marimammalium*, in the gut and stomach of one of the historical razorbills. The reference genome of *Catellibacterium marimammalium* is incomplete, i.e. consists of contigs as opposed to a single continuous sequence. In reference-based metagenomics it is common practice to restrict the database to complete genomes, which, if implemented in this study, would have resulted in millions of *Catellibacterium marimammalium* sequences being misassigned or not being assigned at all, culminating in spurious interpretations of the only specimen with good DNA preservation. While in metagenomic studies it is certainly improbable for all sequences to be aligned correctly and the inclusion of incomplete genomes come with their own challenges, researchers should be aware that they can miss out on critical taxa if they impose too many restrictions on their databases. While we could demonstrate that it is possible to recover the original microbiome of fluid-preserved specimens, more research is necessary to a priori predict the success rate of extracting endogenous microbial DNA from a specimen. This includes but is not limited to characterising the microbiome of preservation liquids, identifying markers that could indicate

the past use of formaldehyde, and preserving specimens with different preservation strategies and observing the microbial DNA yield before and after preservation.

One limitation of the study is that all fluid-preserved specimens that were investigated are seabirds and the results might therefore not translate well to other fluid-preserved samples. These relatively large-bodied animals are especially difficult to preserve due to the long time it takes for the preserving liquid to permeate through all tissues, and it was therefore recommended in the past to inject larger specimens with formaldehyde. Further studies on smaller organisms that were commonly killed by submerging them in the preservative liquid and where the use of formaldehyde was not as prevalent might reveal a much higher success rate for microbial DNA preservation.

As we only succeeded in recovering the microbiome of one specimen, we were unable to obtain conclusive results on whether the original microbiome can be retrieved from different body sites of fluid-preserved samples. A follow-up study that investigates the survival of highly distinct microbial communities, such as the skin, oral, and gut microbiomes, would therefore inform on how the process of fluid-preservation impacts the boundaries between the organism's microbial communities.

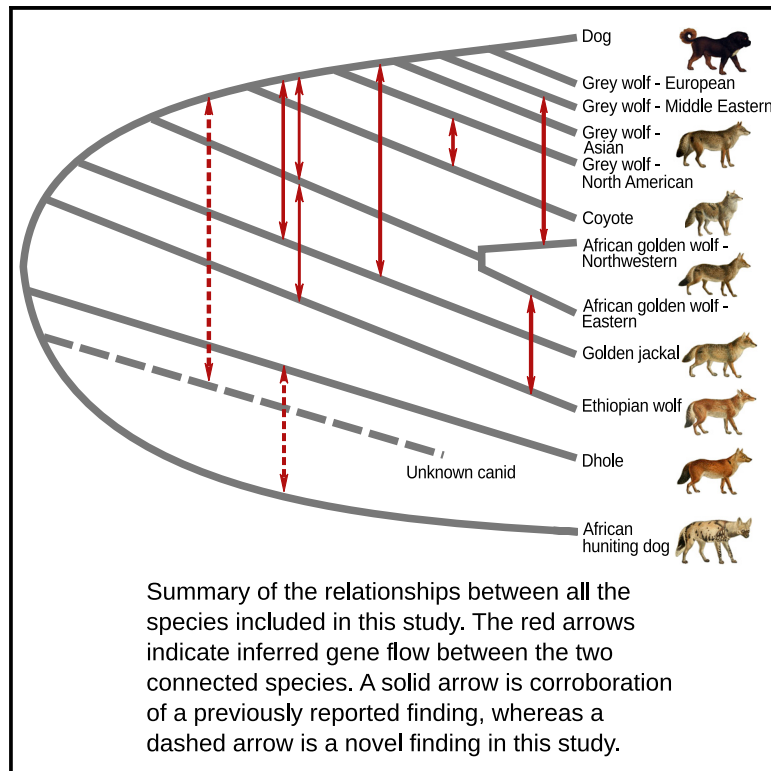
Appendix I

Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*

Current Biology

Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*

Graphical Abstract



Authors

Shyam Gopalakrishnan,
Mikkel-Holger S. Sinding,
Jazmín Ramos-Madrigal, ...,
Øystein Wiig, Anders J. Hansen,
M. Thomas P. Gilbert

Correspondence

shyam@snm.ku.dk

In Brief

Gopalakrishnan et al. present evidence of pervasive gene flow among species of the genus *Canis*. In addition to previously known admixture events, they find evidence of gene flow from a “ghost” canid, related to the dhole, into the ancestor of the gray wolf and coyote. Further, they suggest that the African golden wolf is a species of hybrid origin.

Highlights

- Extensive gene flow in the genus *Canis*, especially among the crown group
- Genetic contribution from an unknown canid into the ancestor of the gray wolf and coyote
- The African golden wolf possibly a hybrid species, from the gray wolf and Ethiopian wolf
- Possible ancient admixture between the dhole and African hunting dog



Interspecific Gene Flow Shaped the Evolution of the Genus *Canis*

Shyam Gopalakrishnan,^{1,23,24,*} Mikkel-Holger S. Sinding,^{1,2,3,4,23} Jazmín Ramos-Madrigal,^{1,23} Jonas Niemann,¹ Jose A. Samaniego Castruita,¹ Filipe G. Vieira,¹ Christian Carøe,¹ Marc de Manuel Montero,⁵ Lukas Kuderna,⁵ Aitor Serres,⁵ Víctor Manuel González-Basallote,⁵ Yan-Hu Liu,⁹ Guo-Dong Wang,¹⁰ Tomas Marques-Bonet,^{5,6,7,8} Siavash Mirarab,¹¹ Carlos Fernandes,¹² Philippe Gaubert,¹³ Klaus-Peter Koepfli,^{14,15} Jane Budd,¹⁶ Eli Knispel Rueness,¹⁷ Claudio Sillero,^{18,19} Mads Peter Heide-Jørgensen,^{1,3} Bent Petersen,^{20,21} Thomas Sicheritz-Ponten,^{20,21} Lutz Bachmann,² Øystein Wiig,² Anders J. Hansen,^{1,3,4} and M. Thomas P. Gilbert^{1,22}

¹Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark

²Natural History Museum, University of Oslo, Oslo, Norway

³The Qimmeq Project, University of Greenland, Nuussuaq, Greenland

⁴University of Greenland, Manuutoq 1, Nuuk, Greenland

⁵Institute of Evolutionary Biology (UPF-CSIC), PRBB, Barcelona, Spain

⁶Catalan Institution of Research and Advanced Studies (ICREA), Passeig de Lluís Companys, 23, 08010, Barcelona, Spain

⁷CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri i Reixac 4, 08028 Barcelona, Spain

⁸Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, Edifici ICTA-ICP, c/ Columnes s/n, 08193 Cerdanyola del Vallès, Barcelona, Spain

⁹State Key Laboratory for Conservation and Utilization of Bio-Resources in Yunnan, Yunnan University, Kunming, Yunnan, China

¹⁰State Key Laboratory of Genetic Resources and Evolution and Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China

¹¹Department of Electrical and Computer Engineering, University of California, San Diego, San Diego, CA, USA

¹²Centre for Ecology, Evolution and Environmental Changes (CE3C), Departamento de Biologia Animal, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal

¹³Institut des Sciences de l'Evolution de Montpellier (ISEM), UM-CNRS-IRD-EPHE, Université de Montpellier, Montpellier, France

¹⁴Smithsonian Conservation Biology Institute, National Zoological Park, 3001 Connecticut Avenue NW, Washington, DC 20008, USA

¹⁵Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, 41A Sredniy Prospekt, St. Petersburg 199034, Russia

¹⁶Breeding Centre for Endangered Arabian Wildlife, Sharjah, United Arab Emirates

¹⁷Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, Oslo, Norway

¹⁸Wildlife Conservation Research Unit, Zoology, University of Oxford, Tubney House, Tubney OX13 5QL, UK

¹⁹IUCN SSC Canid Specialist Group, Oxford, UK

²⁰DTU Bioinformatics, Department of Bio and Health Informatics, Technical University of Denmark, Lyngby, Denmark

²¹Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), Faculty of Applied Sciences, AIMST University, Kedah, Malaysia

²²Norwegian University of Science and Technology, University Museum, Trondheim, Norway

²³These authors contributed equally

²⁴Lead Contact

*Correspondence: shyam@snm.ku.dk

<https://doi.org/10.1016/j.cub.2018.08.041>

SUMMARY

The evolutionary history of the wolf-like canids of the genus *Canis* has been heavily debated, especially regarding the number of distinct species and their relationships at the population and species level [1–6]. We assembled a dataset of 48 resequenced genomes spanning all members of the genus *Canis* except the black-backed and side-striped jackals, encompassing the global diversity of seven extant canid lineages. This includes eight new genomes, including the first resequenced Ethiopian wolf (*Canis simensis*), one dhole (*Cuon alpinus*), two East African hunting dogs (*Lycaon pictus*), two Eurasian golden jackals (*Canis aureus*), and two Middle Eastern gray wolves (*Canis lupus*). The relationships between the Ethiopian wolf, African golden wolf, and golden jackal were

resolved. We highlight the role of interspecific hybridization in the evolution of this charismatic group. Specifically, we find gene flow between the ancestors of the dhole and African hunting dog and admixture between the gray wolf, coyote (*Canis latrans*), golden jackal, and African golden wolf. Additionally, we report gene flow from gray and Ethiopian wolves to the African golden wolf, suggesting that the African golden wolf originated through hybridization between these species. Finally, we hypothesize that coyotes and gray wolves carry genetic material derived from a “ghost” basal canid lineage.

RESULTS AND DISCUSSION

The genome dataset analyzed in this study contains 12 gray wolves and 14 dogs, chosen from regions overlapping the



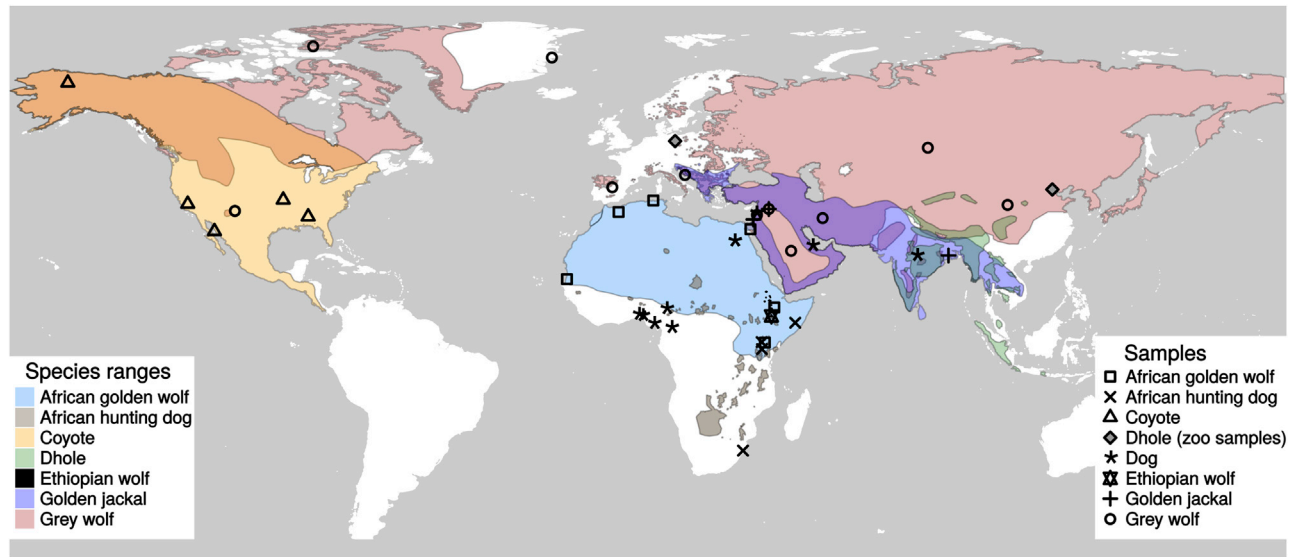


Figure 1. Map Showing the IUCN Ranges, Range Overlaps, and Sampling Locations of the Canids Included in This Study

The overlaps in ranges are shown in blended colors (orange, dark purple, dark olive green, light teal, etc.). Since IUCN does not have range information for African golden wolf, the IUCN range of golden jackal has been split in two; the Eurasian part is shown as the range of golden jackal, and the African part is shown as the range of African golden wolf. Further details on the samples, including their sampling location and source, can be found in [Data S1](#), and their estimated heterozygosities—which are inversely proportional to their population sizes—are shown in [Figure S1](#).

current ranges of the other basal canids included in this study, five coyotes, one Ethiopian wolf, three golden jackals, six African golden wolves (originally *Canis anthus*, but recently reclassified as *Canis lupaster* [1]), two dholes, four African hunting dogs, and one Andean fox (*Lycalopex culpaeus*) (Figure 1). Short-read sequencing of the samples and subsequent alignment to the recently published wolf genome assembly [7] resulted in genome-wide coverages ranging from 0.6–26.6 \times (for details, see [Data S1](#)). The genome-wide heterozygosity estimates (Figure S1) clearly show reduced levels in the Ethiopian wolf, African hunting dog, and dhole, an observation that is consistent with their small population sizes. The reconstructed phylogenetic relationships within this group of canids (Figure 2B) are of considerable relevance in light of extensive prior debate on the relationships between the Ethiopian wolf, golden jackal, and African golden wolf [2–5]. Our results corroborate the recent proposition based on both mitochondrial [2, 3] and nuclear [4, 6] data that the African golden wolf is evolutionarily distinct from the golden jackal (Figure 2C, panel labeled 16), but also that the Ethiopian wolf falls basal to both (Figure 2C, panel labeled 12) [5]. For convenience, we henceforth refer to five canid species, viz. the Ethiopian wolf, African golden wolf, golden jackal, gray wolf, and coyote, as “the crown group” in order to distinguish them from the more basal dholes and African hunting dogs. The placement of the Ethiopian wolf as the basal group in this clade is consistent with tree topologies obtained in previous phylogenetic analyses based on concatenated gene sequences [5] and more recent multispecies coalescent analyses [4] of datasets consisting of a subset of exonic and intronic sequences, but differs from the topology based on concatenated analyses in the latter study. We note that this nuclear-DNA-based phylogeny also places dogs as a sister clade to European gray wolves. However, we caution that this placement has only moderate

support (0.86 mean local posterior probability); moreover, the gene tree quartet frequencies of alternate resolutions within the dog-gray wolf branches are comparable to that recovered in the main tree (Figure 2B, panel labeled 20–22), and thus no conclusion can be drawn about which wolf population gave rise to dogs. Indeed, our findings are not incompatible with previously suggested hypotheses [9] that either (1) the dog was domesticated from a now-extinct wolf population and/or (2) Eurasian gray wolf population genomic diversity has been reduced since the domestication event.

Mitochondrial genomes were *de novo* assembled from all species studied, using MtArchitect [10], which accounts for presence of numts in the reference genome. A maximum-likelihood phylogeny based on these mitochondrial genomes (Figure 2A) is largely consistent with that obtained from the nuclear genome analysis, with one obvious exception—the coyote mitochondrial genomes fall basal to all the other crown canids. This is consistent with Koepfli and colleagues’ [4] results on near-complete mitochondrial genomes and thus contradicts the findings of numerous previous studies that used partial mitochondrial DNA sequences and placed coyotes (1) as sister to gray wolves [11], (2) in an unresolved clade with African golden wolves and Ethiopian wolves [2, 3], (3) as sister to Ethiopian wolves [1, 2, 12, 13], or, finally, (4) as sister to a clade containing Ethiopian wolves and golden jackals [14].

We subsequently explored the degree of interspecific gene flow between the various species. Many publications have reported interspecies gene flow between members of the canid crown group (dog-gray wolf complex, coyotes, Ethiopian wolves, golden jackals, and African golden wolves) [4, 5, 9, 13, 15–19]—something perhaps unsurprising, given the large geographic overlap of many of the populations. Initial analyses of genetic structure among these canids using NGSadmix [20] (Figure S3A)

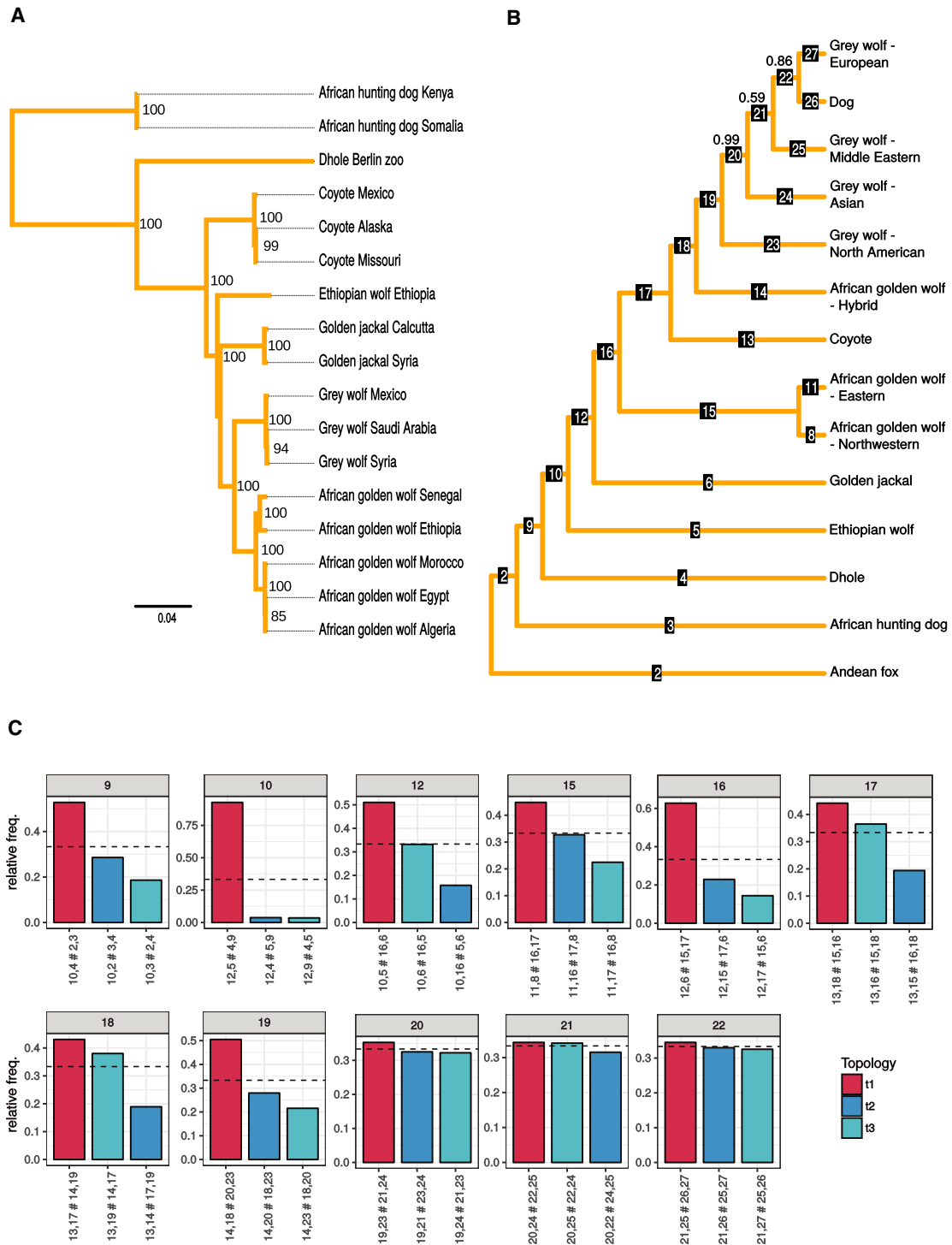


Figure 2. Nuclear and Mitochondrial Phylogeny of Basal Canids

(A) The maximum-likelihood estimate of the mitochondrial phylogeny for a subset of the samples, using *de novo* mitochondrial assemblies obtained with MtArchitect. The node labels show the bootstrap support for the node.

(B) The phylogeny estimated from nuclear DNA by ASTRAL-II, where monophyletic clusters have been collapsed into a single leaf node. The tip labeled “African golden wolf-hybrid” represents a single known hybrid from the Sinai Peninsula—labeled “African golden wolf Egypt” in the mtDNA phylogeny— as described in the main text. The mean local posterior probabilities are shown for branches where this value is less than 1. The full nuclear phylogeny containing the sample relationships, branch supports, branch lengths proportional to divergence times, and estimated split times can be found in [Figures S2A and S2B](#) and [Table S2](#).

(C) For a subset of the internal branches in the nuclear phylogeny, the quartet frequencies of the three possible configurations around each branch in the underlying unrooted tree are shown. The red bar represents the configuration shown in the phylogeny, and the two blue bars represent the two alternative

(legend continued on next page)

revealed that the individuals partition according to expected species structure. However, more details became apparent as the number of estimated clusters (K) was increased. For example, at higher values of K , gray wolves form five principal groups (Mexico, Ellesmere-Greenland, East Asia, the Middle East, and the remaining Eurasia), whereas African golden wolves are split into an Eastern and a Northwestern clade, as previously shown [4, 6, 16]. We note that similar east-west population differentiation is observed for several other African mammalian species [21], thus pointing to a general trend that the African golden wolves follow. The NGSadmixture analyses also suggest the presence of admixture between the different species. For example, we detected not only dog introgression in the gray wolves from Spain and Israel, but also, perhaps of greater interest, gene flow between African golden wolves, golden jackals, and gray wolves. One example is a highly admixed African golden wolf from the Egyptian Sinai Peninsula, whose genome contains contributions from both Middle Eastern gray wolves and dogs (Figure S3A).

Previous studies that have reported admixture between canid species [9] and mitochondrial evidence for overlap of the gray wolf, African golden wolf, and golden jackal in eastern Egypt [4]. This points to the importance of the Sinai Peninsula and the Southwest Levant in canid evolution [4, 9], presumably due to its role as the land bridge between the African and Eurasian continents. We used TreeMix [22], D statistics [23], and admixture graphs [23] to examine signals of admixture between these species. The results confirmed that, in general terms, the level of gene flow between the three species is high, although varying across space in a manner consistent with their natural ranges (Figures 3B and S3A–S3E). For example, gene flow between golden jackals and gray wolves and between African golden wolves and gray wolves is lowest when North American gray wolves are considered, somewhat higher for Asian and European gray wolves, and highest with the gray wolves from the Middle East (e.g., Israel, Syria, and Saudi Arabia) (Figure S3E). Although the latter is not surprising in light of the natural ranges of the species, the evidence of golden jackal ancestry in North American wolves is intriguing. One possible explanation could be that gene flow happened before the divergence of the North American and Eurasian gray wolves. The fact that interspecific gene flow is considerably higher in Middle Eastern than in other gray wolves may also explain the distinctness of this population. The structure between Northwestern and Eastern African golden wolves can be explained using a similar argument—the former have highest levels of golden jackal and gray wolf admixture (Figures 3B, S3A, and S3B), whereas the latter show higher levels of gene flow from Ethiopian wolves. Overall, it is clear that individuals sampled in this land bridge region will be particularly informative for future studies that wish to study canid admixture in greater detail.

Furthermore, D statistics were used to test for gene flow between the dhole and African hunting dog, using members of the crown group as ingroup and the Andean fox as outgroup.

Although no gene flow was detected between species of the crown group and the African hunting dogs, the analyses provided strong evidence of gene flow between the African hunting dog and dhole (Figure S3C). This is a surprising finding, since the ranges of the two species do not overlap. However, it is well documented that the dhole existed as far west as Europe during the Pleistocene [24]. Thus, one possible explanation could be the presence of dholes in the Middle East in the past, from where they could have encountered and mixed with African hunting dogs in North Africa. It must, however, be stressed that given that there has never been any reported evidence of dholes in either the Middle East or North Africa, our hypothesis is purely speculative. The timing and location of this admixture event remain unresolved.

Although there have been several reports of hybridization between dogs and Ethiopian wolves [13, 15], the genetic history of the Ethiopian wolf has not previously been investigated using nuclear genomic data. The D -statistics-based analyses provided evidence for gene flow between Ethiopian wolves and not only African golden wolves, but also golden jackals, gray wolves, and coyotes (Figure S3). The finding of considerable gene flow between the Ethiopian and Eastern African golden wolf lineages is not surprising, given their geographical co-occurrence in Africa. We consistently also observed a Northwestern-Eastern split in the African golden wolves and note that this correlates with our finding that the Ethiopian wolf contributes a higher amount to the Eastern African golden wolves. This suggests that admixture from the Ethiopian wolf may be a key factor contributing to African golden wolf population structure.

The presence of gene flow between the Ethiopian wolf and the other crown canid species is more surprising, given their lack of range overlap. However, this might be explained through the previously reported extensive evidence of admixture between African golden wolves and gray wolves, coyotes, and golden jackals [4, 9]. In short, we hypothesize that the signal of Ethiopian wolf admixture into the other crown canid species is mediated by African golden wolves. A summary of all the admixture events inferred in this study is shown in Figure 3A.

The uncertain placement of the African golden wolf (Figure 2C, panel labeled 17), combined with evidence of gene flow from the Ethiopian wolf, led us to investigate whether the African golden wolf is a species of hybrid origin, derived from a mixture between gray and Ethiopian wolves or close relatives. The current distribution ranges of Ethiopian and gray wolves do not overlap, and indeed, the known historical distribution of Ethiopian wolves is restricted to the Ethiopian highlands [15]. However, extensive gene flow with other canids, combined with the two distinct levels of Ethiopian wolf gene flow into the two distinct populations of African golden wolves, suggests that either Ethiopian wolves or a close (now extinct) relative had, in the past, a much larger range within Africa and thus greater opportunity to admix with other canid species. Additionally, mitochondrial analyses of African golden wolves, in this and previous studies,

configurations. For every quartet, the frequency of the true bipartition has previously been shown to be at least one-third [8], indicated here by a dotted line. Each alternative configuration is labeled by the bipartition it creates, with labels corresponding to those in (A). For example, the second bar of the panel labeled 12 swaps the positions of golden jackal (6) and Ethiopian wolf (5), whereas the third bar puts them as sister to each other. This plot summarizes the gene tree incongruence around examined branches.

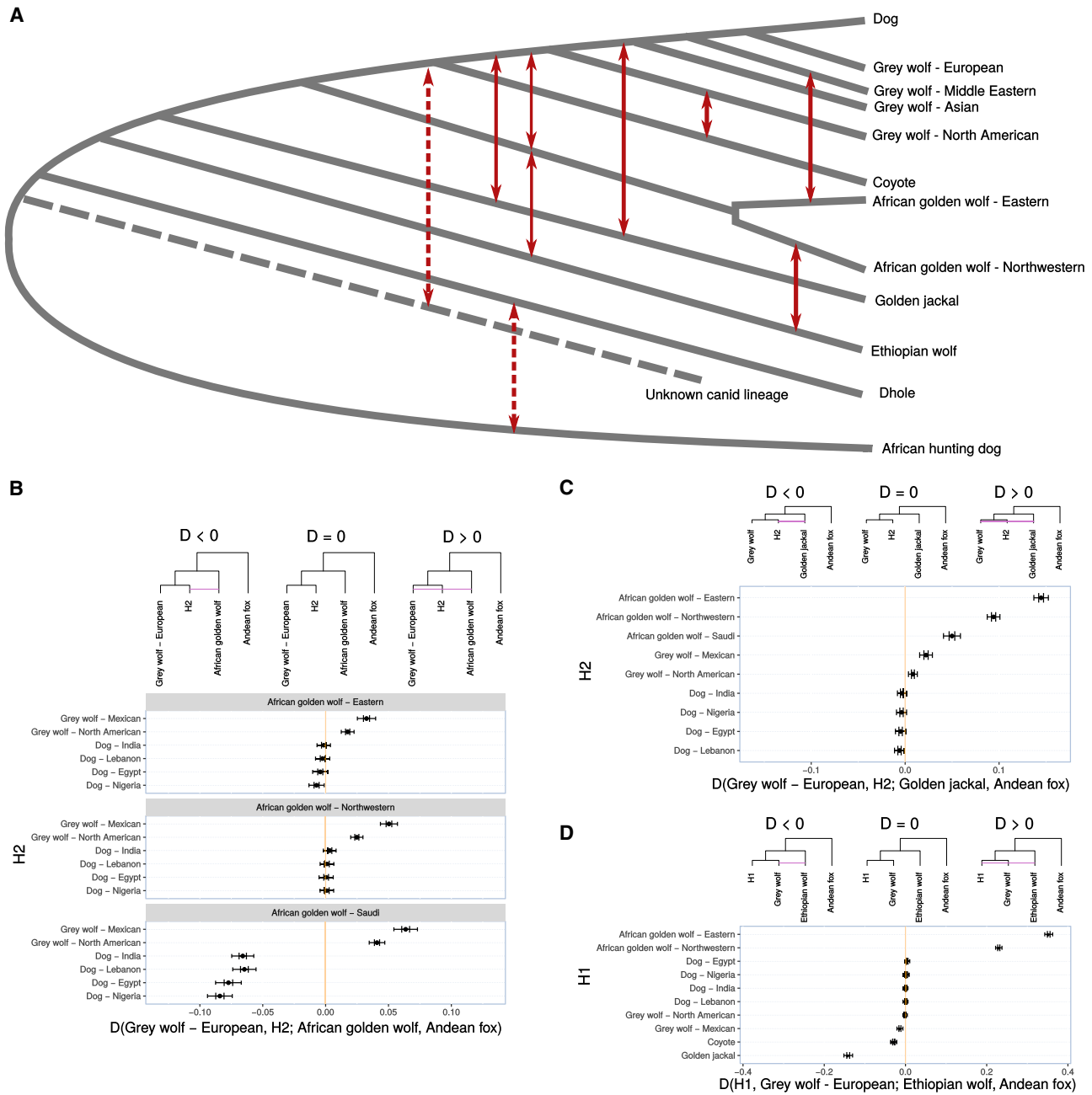


Figure 3. Gene Flow among the Crown Canid Species

(A) This figure summarizes the relationships among the species (phylogeny) and the various gene flow events inferred from the samples included in this study. Gene flow events are indicated with red arrows, and dotted red arrows show possible gene flow events that have been inferred in this study but have not been previously reported.

(B–D) These figures show the gene flow among the different crown canid species using *D* statistics. These *D* statistics show significant gene flow between the gray wolf, African golden wolf, golden jackal, and Ethiopian wolf. One principal new finding is structure within the African golden wolves, splitting into Northwestern and Eastern clades, which show genetic affinity to gray wolves and Ethiopian wolves, respectively. A second principal finding is inferred gene flow from an unknown canid lineage, related to the dhole, into the ancestor of the coyote and the gray wolves. We hypothesize this may explain the unexpected basal placement of the coyote in the mitochondrial tree. Further evidence of gene flow in the crown canids is shown in Figure S3.

find them to be most closely related to gray wolves [2–4, 25]. Further, African golden wolves are a sister clade to gray wolves and coyotes in the nuclear phylogeny, whereas they are a sister group to the Middle Eastern gray wolves in the mitochondrial

phylogeny. We explored the relationships between the golden jackal, Ethiopian wolf, and African golden wolf using G-PhoCS [26] (Table S1), which supported the finding of gene flow into the Ethiopian wolf from the African golden wolf. To further

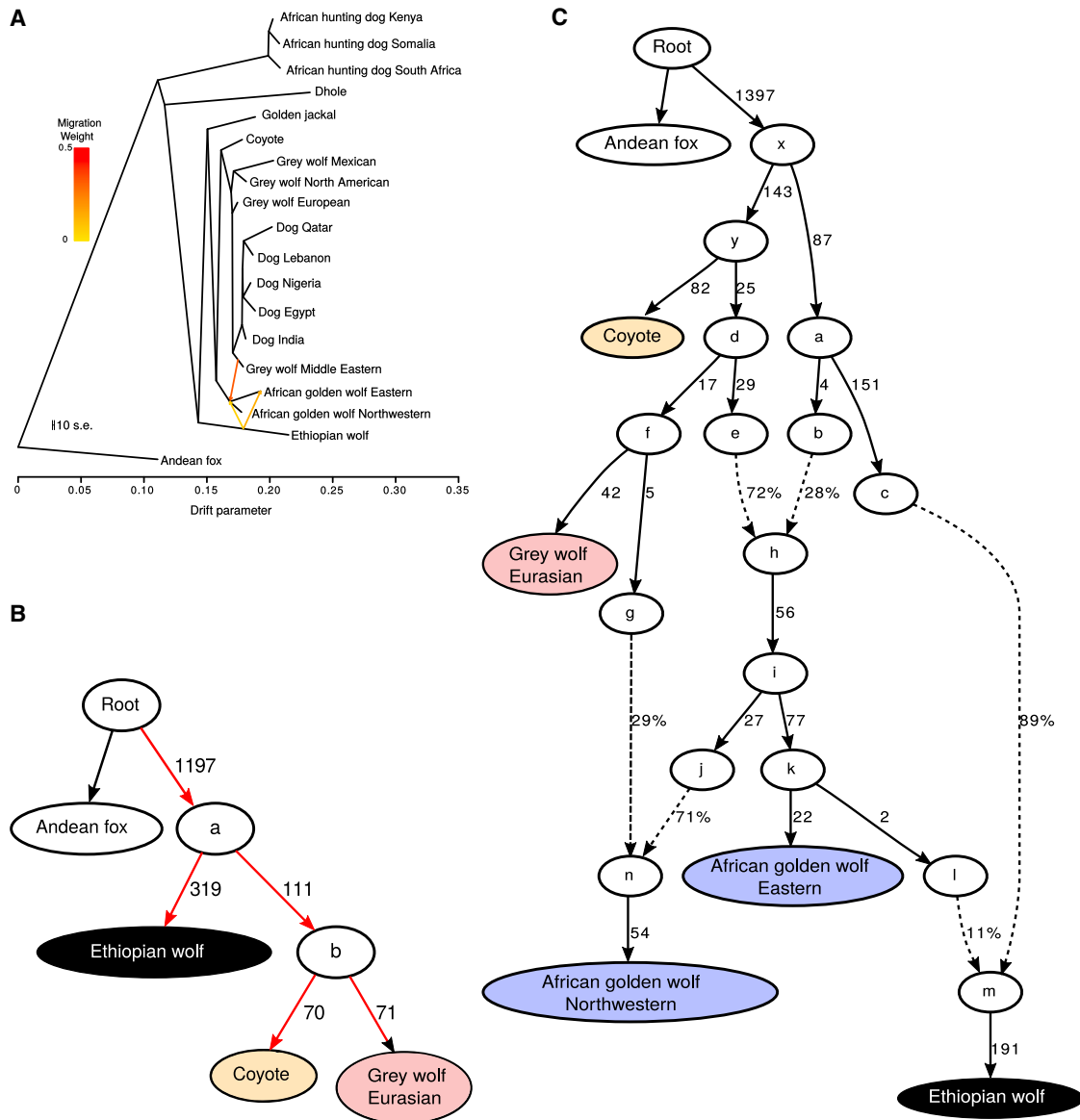


Figure 4. Modeling the Ancestry of African Golden Wolves

(A) TreeMix tree with all samples, estimated using the pairwise correlation of allele frequencies between all groups of samples. This tree is fit with three migration edges. The first three migration edges all indicate extensive gene flow from the gray and Ethiopian wolves into the African golden wolves, suggesting a hybrid origin for this species.

(B and C) The QP graph is an admixture graph estimated using all pairwise *D* statistics between samples. Estimated genetic drift is shown along the solid lines in units of *f*₂ distance (parts per thousand), and estimated mixture proportions are given along the dotted lines. Names of specific modern populations are shown in full, whereas hypothetical ancestral individuals are represented by letters.

(B) This tree shows all the possible placements—highlighted in red—for the Northwestern African golden wolf, chosen due to their low levels of gene flow with the Ethiopian wolf. These were modeled as possible internal and external nodes and as an admixed group from all possible node pairs.

(C) The best fitting graph with a *Z* value closest to 0, modeling the Ethiopian wolf-like and gray wolf-like ancestry of Northwestern and Eastern African golden wolves, as well as gene flow into modern Ethiopian wolves from the Eastern African golden wolves. This admixture graph suggests that the African golden wolves are probably a species of hybrid origin, derived from the gray wolf and Ethiopian wolf as the parental species. Further, Figure S4 shows admixture graphs showing potential gene flow from a “ghost” basal canid lineage into the ancestor of wolves and dogs.

explore the relationship between these species and the gray wolf, we used TreeMix [22] and admixture graphs [23] to obtain trees, which were used to assess whether the African golden wolf is a hybrid species (Figures 4B and 4C). We initially constructed a graph including the coyote, Ethiopian wolf, gray

wolf, and Andean fox and assessed the most likely position for the African golden wolf in this graph. The placement of the two African golden wolf populations in this tree was further investigated by modeling them as sister to all possible nodes and as admixed populations deriving ancestry from two possible nodes.

Finally, the model was extended to account for African golden wolf admixture into the Ethiopian wolf. We found that the common ancestor of the African golden wolf populations is best modeled as admixed between a component related to the Ethiopian wolf (~28%) and another related to the gray wolf (~72%) (worst-fitting f statistic Z value = -1.086 ; Figure 4C). Finally, the northwestern African golden wolf population is more closely related to the gray wolf, which is best explained in our model through admixture from gray wolves.

Lastly, our attention was drawn to the curious result of potential gene flow between the lineage representing the ancestor of the coyote and gray wolves and that representing all other canid species, excluding the African hunting dog (Figure S4), in all D statistics analyses computed with the coyote or gray wolf in the ingroup, namely position H2. Notably, these signals disappeared when the sister clade—H3—was replaced with the African hunting dog, leading us to hypothesize that the coyote and gray wolf genomes may contain a basal ancestral component derived from an as-yet-unknown species that evolved after the divergence of the African hunting dog branch from the other canid species and that the signal of gene flow can be attributed to outgroup attraction of the coyote and gray wolf lineage. Note that such a hypothetical ancient admixture event would also explain the unexpectedly basal position of the coyote mitochondrial genome—the coyote may simply have retained the mitogenome from this unidentified ancestor. We acknowledge that the existence of an unknown ancestral component would be controversial—previous analyses of coyotes and the fossil records from their direct ancestors argue that they have been strictly restricted to North America for over a million years [27, 28]. However, within North America, the coyote has coexisted alongside several now extinct canids, including the American dhole (*Cuon* sp.) and dire wolf (*Canis dirus*) [29]. Although the unknown ancestral component cannot be attributed to any of the known fossil species at this time, future paleogenomic analyses on such materials (if any can be found with surviving DNA) may provide exciting possibilities to test our hypothesis.

In conclusion, our results highlight how interspecific gene flow has played an important role in shaping the species and population structure of gray wolves, coyotes, African golden wolves, golden jackals, and Ethiopian wolves and that African golden wolves, coyotes, and gray wolves may have been greatly affected by hybridization events. In particular, we conclude not only that African golden wolves arose through hybridization between a Ethiopian-wolf-like and gray-wolf-like ancestral population, but that subsequently the resulting northwestern and eastern African golden wolf populations underwent continuous admixture with modern gray and Ethiopian wolves, respectively. We furthermore argue that the common ancestor of gray wolves and coyotes differentiated from the lineage leading to golden jackals, in part by admixing with a dhole-like canid. Finally, the robust signal of gene flow observed between African hunting dogs and dholes testifies to an as-yet-undiscovered prehistoric overlap between the two lineages. This underscores how much remains to be discovered about the history of the wolf-like canids and how paleogenomic approaches may be required to advance our understanding of this group. Lastly, our study adds to the growing evidence for the importance of gene flow and hybridization in the evolution of mammalian species in general [23, 30–32]

and that rather than being isolated entities that evolve along tree-like phylogenies, they are interlinked and evolve through interactions in network-like topologies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
 - Whole-genome sequencing
 - Read mapping
 - Genotype calling
- QUANTIFICATION AND STATISTICAL ANALYSIS
 - Heterozygosity
 - Admixture
 - Nuclear genome phylogeny
 - Species split times
 - Mitochondrial reconstruction using *de novo* assembly
 - D statistics
 - TreeMix
 - qpGraph
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, two tables, and one data file and can be found with this article online at <https://doi.org/10.1016/j.cub.2018.08.041>.

ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of the Danish National High-Throughput Sequencing Centre for assistance in Illumina data generation. We also gratefully acknowledge the Danish National Supercomputer for Life Sciences, Computerome (<https://www.computerome.dk>), for the computational resources to perform the sequence analyses. For making sample material available, we would like to thank Jörns Fickel from Leibniz-Institut für Zoo- und Wildtierforschung and Kristian Gregersen from the Natural History Museum of Denmark. We also acknowledge the following for funding our research: the Qimmeq project funded by The Velux Foundations and Aage og Johanne Louis-Hansens Fond; Carlsbergfondet grant CF14-0995 and Marie Skłodowska-Curie Actions grant 655732-WhereWolf to S.G.; grant 676154-ArchSci2020 to J.N.; NSFC grant 91531303 to G.-D.W.; Danish National Research Foundation grant DNRF94, Lundbeckfonden grant R52-5062, and ERC Consolidator grant 681396-Extinction Genomics to M.T.P.G.; and the Universities of Oslo and Copenhagen for a PhD stipend awarded to M.-H.S.S. T.M.-B. is supported by MINECO/FEDER, UE, grant BFU2017-86471-P, NIMH grant U01 MH106874, a Howard Hughes Medical Institute International Early Career grant, Obra Social “La Caixa,” and Secretaria d’Universitats i Recerca and CERCA Programme del Departament d’Economia i Coneixement de la Generalitat de Catalunya.

AUTHOR CONTRIBUTIONS

S.G., M.-H.S.S., A.J.H., and M.T.P.G. conceived the study. M.-H.S.S. and C.C. did the DNA lab work for high-throughput sequencing. S.G., J.R.-M., J.N., J.A.S.C., F.G.V., M.d.M.M., L.K., A.S., V.M.G.-B., Y.-H.L., and S.M. performed analyses. C.F., P.G., K.-P.K., J.B., E.K.R., C.S., and M.P.H.-J. contributed with sample collection. B.P. and T.S.-P. provided computation expertise and support. L.B., Ø.W., T.M.-B., A.J.H., and M.T.P.G. supervised the work.

S.G., M.-H.S.S., J.R.-M., and M.T.P.G. wrote the manuscript. All authors contributed to the preparation and editing of the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 9, 2017

Revised: April 30, 2018

Accepted: August 16, 2018

Published: October 18, 2018; corrected online: November 8, 2019

REFERENCES

- Viranta, S., Atickem, A., Werdelin, L., and Stenseth, N.C. (2017). Rediscovering a forgotten canid species. *BMC Zoology* 2, 6.
- Rueness, E.K., Asmyhr, M.G., Sillero-Zubiri, C., Macdonald, D.W., Bekele, A., Atickem, A., and Stenseth, N.C. (2011). The cryptic African wolf: *Canis aureus lupaster* is not a golden jackal and is not endemic to Egypt. *PLoS ONE* 6, e16385.
- Gaubert, P., Bloch, C., Benyacoub, S., Abdelhamid, A., Pagani, P., Djagoun, C.A.M.S., Couloux, A., and Dufour, S. (2012). Reviving the African wolf *Canis lupus lupaster* in North and West Africa: a mitochondrial lineage ranging more than 6,000 km wide. *PLoS ONE* 7, e42740.
- Koepfli, K.-P., Pollinger, J., Godinho, R., Robinson, J., Lea, A., Hendricks, S., Schweizer, R.M., Thalmann, O., Silva, P., Fan, Z., et al. (2015). Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. *Curr. Biol.* 25, 2158–2165.
- Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C., et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438, 803–819.
- Fan, Z., Silva, P., Gronau, I., Wang, S., Armero, A.S., Schweizer, R.M., Ramirez, O., Pollinger, J., Galaverni, M., Ortega Del-Vecchyo, D., et al. (2016). Worldwide patterns of genomic variation and admixture in gray wolves. *Genome Res.* 26, 163–173.
- Gopalakrishnan, S., Samaniego Castruita, J.A., Sinding, M.S., Kuderna, L.F.K., Rääkkönen, J., Petersen, B., Sicheritz-Ponten, T., Larson, G., Orlando, L., Marques-Bonet, T., et al. (2017). The wolf reference genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population genomics. *BMC Genomics* 18, 495.
- Allman, E.S., Degnan, J.H., and Rhodes, J.A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.* 62, 833–862.
- Freedman, A.H., Gronau, I., Schweizer, R.M., Ortega-Del Vecchyo, D., Han, E., Silva, P.M., Galaverni, M., Fan, Z., Marx, P., Lorente-Galdos, B., et al. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet.* 10, e1004016.
- Lobon, I., Tucci, S., de Manuel, M., Ghirotto, S., Benazzo, A., Prado-Martinez, J., Lorente-Galdos, B., Nam, K., Dabad, M., Hernandez-Rodriguez, J., et al. (2016). Demographic history of the genus *Pan* inferred from whole mitochondrial genome reconstructions. *Genome Biol. Evol.* 8, 2020–2030.
- Ostrander, E.A., and Wayne, R.K. (2005). The canine genome. *Genome Res.* 15, 1706–1716.
- Wayne, R.K., and Ostrander, E.A. (1999). Origin, genetic diversity, and genome structure of the domestic dog. *BioEssays* 21, 247–257.
- Gottelli, D., Sillero-Zubiri, C., Applebaum, G.D., Roy, M.S., Girman, D.J., Garcia-Moreno, J., Ostrander, E.A., and Wayne, R.K. (1994). Molecular genetics of the most endangered canid: the Ethiopian wolf *Canis simensis*. *Mol. Ecol.* 3, 301–312.
- Yumnam, B., Negi, T., Maldonado, J.E., Fleischer, R.C., and Jhala, Y.V. (2015). Phylogeography of the golden jackal (*Canis aureus*) in India. *PLoS ONE* 10, e0138497.
- Marino, J., and Sillero-Zubiri, C. (2011). *Canis simensis*. The IUCN Red List of Threatened Species 2011: e.T3748A10051312. <https://doi.org/10.2305/IUCN.UK.2011-1.RLTS.T3748A10051312.en>.
- vonHoldt, B.M., Pollinger, J.P., Earl, D.A., Knowles, J.C., Boyko, A.R., Parker, H., Geffen, E., Pilot, M., Jedrzejewski, W., Jedrzejewska, B., et al. (2011). A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. *Genome Res.* 21, 1294–1305.
- Galov, A., Fabbri, E., Caniglia, R., Arbanasić, H., Lapalombella, S., Florijančić, T., Bošković, I., Galaverni, M., and Randi, E. (2015). First evidence of hybridization between golden jackal (*Canis aureus*) and domestic dog (*Canis familiaris*) as revealed by genetic markers. *R. Soc. Open Sci.* 2, 150450.
- vonHoldt, B.M., Cahill, J.A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J.P., Shapiro, B., Wall, J., and Wayne, R.K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Sci. Adv.* 2, e1501714.
- vonHoldt, B.M., Kays, R., Pollinger, J.P., and Wayne, R.K. (2016). Admixture mapping identifies introgressed genomic regions in North American canids. *Mol. Ecol.* 25, 2443–2453.
- Skotte, L., Korneliusen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics* 195, 693–702.
- Lorenzen, E.D., Heller, R., and Siegmund, H.R. (2012). Comparative phylogeography of African savannah ungulates. *Mol. Ecol.* 21, 3656–3670.
- Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967.
- Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093.
- Ripoll, M.P., Morales Pérez, J.V., Sanchis Serra, A., Aura Tortosa, J.E., and Montañana, I.S. (2010). Presence of the genus *Cuon* in upper Pleistocene and initial Holocene sites of the Iberian Peninsula: new remains identified in archaeological contexts of the Mediterranean region. *J. Archaeol. Sci.* 37, 437–450.
- Werhahn, G., Senn, H., Kaden, J., Joshi, J., Bhattarai, S., Kusi, N., Sillero-Zubiri, C., and Macdonald, D.W. (2017). Phylogenetic evidence for the ancient Himalayan wolf: towards a clarification of its taxonomic status based on genetic sampling from western Nepal. *R. Soc. Open Sci.* 4, 170186.
- Gronau, I., Hubisz, M.J., Gulko, B., Danko, C.G., and Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* 43, 1031–1034.
- Kurtén, B. (1974). A history of coyote-like dogs (Canidae, Mammalia). *Acta Zool. Fenn.* 140, 1–38.
- Nowak, R.M. (1979). North American Quaternary Canis (Museum of Natural History, University of Kansas).
- Tedford, R.H., Wang, X., and Taylor, B.E. (2009). Phylogenetic systematics of the North American fossil Caninae (Carnivora: Canidae). *Bull. Am. Mus. Nat. Hist.* 325, 1–218.
- Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., Albrechtsen, A., Petersen, B., Korneliusen, T.S., Vilstrup, J.T., et al. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proc. Natl. Acad. Sci. USA* 111, 18655–18660.
- Figueiró, H.V., Li, G., Trindade, F.J., Assis, J., Pais, F., Fernandes, G., Santos, S.H.D., Hughes, G.M., Komissarov, A., Antunes, A., et al. (2017). Genome-wide signatures of complex introgression and adaptive evolution in the big cats. *Sci. Adv.* 3, e1700299.
- Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M.A., and Janke, A. (2017). The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* 7, 46487.
- Auton, A., Rui Li, Y., Kidd, J., Oliveira, K., Nadel, J., Holloway, J.K., Hayward, J.J., Cohen, P.E., Grealia, J.M., Wang, J., et al. (2013).

- Genetic recombination is targeted towards gene promoter regions in dogs. *PLoS Genet.* 9, e1003984.
34. Campana, M.G., Parker, L.D., Hawkins, M.T.R., Young, H.S., Helgen, K.M., Szykman Gunther, M., Woodroffe, R., Maldonado, J.E., and Fleischer, R.C. (2016). Genome sequence, population history, and pelage genetics of the endangered African wild dog (*Lycaon pictus*). *BMC Genomics* 17, 1013.
 35. Liu, Y.-H., Wang, L., Xu, T., Guo, X., Li, Y., Yin, T.-T., Yang, H.-C., Yang, H., Adeola, A.C., Sanke, J.O., et al. (2017). Whole-genome sequencing of African dogs provides insights into adaptations against tropical parasites. *Mol. Biol. Evol.* 35, 287–298.
 36. Wang, G.-D., Zhai, W., Yang, H.-C., Fan, R.-X., Cao, X., Zhong, L., Wang, L., Liu, F., Wu, H., Cheng, L.-G., et al. (2013). The genomics of selection in dogs and the parallel evolution between dogs and humans. *Nat. Commun.* 4, 1860.
 37. Wang, G.-D., Zhai, W., Yang, H.-C., Wang, L., Zhong, L., Liu, Y.-H., Fan, R.-X., Yin, T.-T., Zhu, C.-L., Poyarkov, A.D., et al. (2016). Out of southern East Asia: the natural history of domestic dogs across the world. *Cell Res.* 26, 21–33.
 38. Meyer, M., and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 2010, t5448.
 39. Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M.D., Fernández, R., Kircher, M., McCue, M., et al. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* 9, 1056–1082.
 40. Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* 9, 88.
 41. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.
 42. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498.
 43. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303.
 44. Korneliusson, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 15, 356.
 45. Mirarab, S., and Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31, i44–i52.
 46. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313.
 47. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973.
 48. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490.
 49. Sayyari, E., Whitfield, J.B., and Mirarab, S. (2017). DiscoVista: interpretable visualizations of gene tree discordance. *arXiv*, arXiv: 1709.09305, <https://arxiv.org/abs/1709.09305>.
 50. Yamada, K.D., Tomii, K., and Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data—re-examination of the usefulness of chained guide trees. *Bioinformatics* 32, 3246–3251.
 51. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191.
 52. Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* 9, 772.
 53. Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321.
 54. Gilbert, M.T.P., Tomsho, L.P., Rendulic, S., Packard, M., Drautz, D.I., Sher, A., Tikhonov, A., Dalén, L., Kuznetsova, T., Kosintsev, P., et al. (2007). Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* 317, 1927–1930.
 55. Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S.S.T., Sinding, M.-H.S., Samaniego, J.A., Wales, N., Sicheritz-Pontén, T., and Gilbert, M.T.P. (2017). Single-tube library preparation for degraded DNA. *Methods Ecol. Evol.* 9, 410–419.
 56. Allentoft, M.E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlström, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. *Nature* 522, 167–172.
 57. Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., and Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. USA* 110, 15758–15763.
 58. Nielsen, R., Paul, J.S., Albrechtsen, A., and Song, Y.S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451.
 59. Sayyari, E., and Mirarab, S. (2016). Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33, 1654–1668.
 60. Schleich, C.M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A.R., Vicente, M., Steyn, M., Soodyall, H., et al. (2017). Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* 358, 652–655.
 61. Busing, F.M.T.A., Meijer, E., and Van Der Leeden, R. (1999). Delete-M jackknife for unequal M. *Stat. Comput.* 9, 3–8.

STAR★METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---------------------|---|
| Biological Samples | | |
| 8 <i>Canid</i> blood or tissue samples | This paper | Data S1 |
| Chemicals, Peptides, and Recombinant Proteins | | |
| Proteinase K | Sigma-Aldrich | Cat# 3115844001 |
| Phenol | Bionordika | Cat# A0447,0500 |
| Chloroform | Sigma-Aldrich | Cat# 288306-1L |
| Critical Commercial Assays | | |
| DNeasy Blood & Tissue Kit | QIAGEN | Cat# 69506 |
| MinElute PCR Purification Kit | QIAGEN | Cat# 28006 |
| NEBNext DNA Sample Prep Master Mix Set 2 | New England Biolabs | Cat# E6070 |
| Deposited Data | | |
| 10 <i>Canid</i> genomes | [33] | Data S1 |
| 2 <i>Canid</i> genomes | [34] | Data S1 |
| 3 <i>Canid</i> genomes | [6] | Data S1 |
| 5 <i>Canid</i> genomes | [9] | Data S1 |
| 1 <i>Canid</i> genomes | [4] | Data S1 |
| 4 <i>Canid</i> genomes | [35] | Data S1 |
| 2 <i>Canid</i> genomes | [18] | Data S1 |
| 1 <i>Canid</i> genomes | [36] | Data S1 |
| 5 <i>Canid</i> genomes | [37] | Data S1 |
| 1 African golden wolf | This article | NCBI SRA sample accession number: SAMN10199001 |
| 2 African hunting dogs | This article | NCBI SRA sample accession numbers: SAMN10180432, SAMN10180433 |
| 3 Coyotes | This article | NCBI SRA sample accession numbers: SAMN10180421, SAMN10180422, SAMN10180423 |
| 1 Dhole | This article | NCBI SRA sample accession number: SAMN10180424 |
| 1 Ethiopian wolf | This article | NCBI SRA sample accession number: SAMN10180425 |
| 2 Golden jackals | This article | NCBI SRA sample accession numbers: SAMN10180426, SAMN10180427 |
| 5 Gray wolves | This article | NCBI SRA sample accession numbers: SAMN10180428, SAMN10180429, SAMN10180430, SAMN10180431, SAMN10180511 |
| Gray wolf reference genome | [7] | N/A |
| Oligonucleotides | | |
| Illumina-compatible adapters | [38] | N/A |
| Software and Algorithms | | |
| PALEOMIX | [39] | https://github.com/MikkelSchubert/paleomix ; RRID:SCR_015057 |
| AdapterRemoval2 | [40] | https://github.com/MikkelSchubert/adaptremoval ; RRID:SCR_011834 |
| bwa v0.7.10 | [41] | http://bio-bwa.sourceforge.net/ ; RRID:SCR_010910 |
| Picard v1.128 | N/A | https://broadinstitute.github.io/picard ; RRID:SCR_006525 |
| GATK v3.3.0 | [42, 43] | https://broadinstitute.github.io/picard ; RRID:SCR_001876 |
| ANGSD | [44] | https://github.com/ANGSD/angsd |

(Continued on next page)

Continued

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|---|
| samtools v1.2 | [41] | http://samtools.sourceforge.net/ ; RRID:SCR_002105 |
| realSFS | [44] | https://github.com/ANGSD/angsd |
| NGSadmix | [20] | http://www.popgen.dk/software/index.php/NgsAdmix/ ; RRID:SCR_003208 |
| ASTRAL-II | [45] | https://github.com/smirarab/ASTRAL |
| RAxML | [46] | https://sco.h-its.org/exelixis/software.html ; RRID:SCR_006086 |
| trimal | [47] | http://trimal.cgenomics.org/ |
| FastTree2 | [48] | http://www.microbesonline.org/fasttree/ ; RRID:SCR_015501 |
| DiscoVista | [49] | https://github.com/esayyari/DiscoVista |
| MtArchitect | [10] | http://biologiaevolutiva.org/tmarques/mtarchitect/ |
| MAFFT | [50] | https://mafft.cbrc.jp/alignment/software/ ; RRID:SCR_011811 |
| Jalview | [51] | http://www.jalview.org/ ; RRID:SCR_006459 |
| jmodeltest2 | [52] | https://github.com/ddarriba/jmodeltest2 ; RRID:SCR_015244 |
| phyML | [53] | http://www.atgc-montpellier.fr/phyml/ ; RRID:SCR_014628 |
| ADMIXTOOLS | [23] | https://github.com/DReichLab/AdmixTools |
| TreeMix | [22] | https://bitbucket.org/nygcreserach/treemix/wiki/Home |

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Shyam Gopalakrishnan (shyam@snm.ku.dk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The current study uses short read sequencing data from the full genomes of 47 canids spanning 8 different species (when the domestic dog is considered a different species from the gray wolf) from Africa, Eurasia and North America, to address questions about the genetic affinities of these species to each other, and the role of interspecific gene flow in shaping the evolution of the genus *Canis*. All known information on the context and sequencing coverage of the samples is provided in [Data S1](#).

METHOD DETAILS**Whole-genome sequencing**

DNA was extracted from 10 modern samples of fresh blood or tissue using the DNeasy Blood & Tissue Kit (QIAGEN, Hilden, Germany) following the manufacturer's protocol. Three samples ('African hunting dog Kenya 1', 'African hunting dog Somalia' and 'Golden jackal Calcutta') are from historical museum hides and were digested in a proteinase K-containing buffer following [54]; these digests were subsequently treated in a phenol chloroform step following [55]. The supernatant was then mixed 1:10 with a binding buffer following [56] in a binding apparatus following [57], including a Minelute column (QIAGEN, Hilden, Germany) that was then washed and DNA was eluted according to the manufacturer's guidelines. All extracts were incorporated into double-stranded DNA libraries build using the NEBNext DNA Sample Prep Master Mix Set 2 (E6070 - New England Biolabs, Beverly, MA, USA) following the manufacturer's protocol and Illumina-compatible adapters [38]. Libraries were sequenced using 50 base pair single (Golden jackal Calcutta, Hunting dog Kenya 1 and Hunting dog Somalia) or 100 base pair paired end (remaining samples) read chemistry on Illumina HiSeq 2000 and 2500 (Illumina, San Diego, CA, USA) platforms.

Read mapping

The short-read data from each sample, including samples from previous publications, was processed using the PALEOMIX pipeline [39]. As the first step of the pipeline, low quality and missing bases were trimmed from the reads, followed by removal of adapters using

AdapterRemoval2 [40]. Additionally, all paired end reads where the two reads overlapped by more than 10 base pairs were merged into a single read. Subsequently, the reads from each sample were mapped to the wolf reference genome [7] using bwa (v0.7.10; aln algorithm) [41]. The mapped reads were filtered for PCR and optical duplicates using Picard (v1.128, <https://broadinstitute.github.io/picard>), and reads that mapped to multiple locations in the genome were excluded. GATK (v3.3.0) [42, 43] was used to perform an indel realignment step to adjust for increased error rates at the end of short reads in the presence of indels. In the absence of a curated dataset of indels in wolves, this step relied on a set of indels identified in the specific sample being processed. After the initial mapping and quality control, the coverages of the samples ranged from 0.6 to 26.6x (for details see [Data S1](#)).

Genotype calling

The samples in this study span a wide range of genomic coverages. To avoid introducing biases in various analyses resulting from genotype calling in low coverage samples [58], the uncertainty in genotypes was instead propagated through to downstream analyses using genotype likelihoods. The genotype likelihoods at variant sites were computed in ANGSD [44] using the mapped reads, with the model for reads used by samtools (v1.2) [41]. Bases with base qualities lower than 20 and reads with mapping quality lower than 20 were discarded. Only sites with data present in at least 46 out of the 48 samples were retained. All sites with minor allele frequencies below 0.1 were excluded.

QUANTIFICATION AND STATISTICAL ANALYSIS

Heterozygosity

The heterozygosity for each sample was calculated using ANGSD, by estimating the per-sample folded site frequency spectrum (SFS) and using the fraction of singletons in the sample as a measure of heterozygosity. The variance of the estimate was obtained by bootstrapping the sites 100 times to obtain 100 bootstrapped estimates of the SFS. Briefly, for each sample, the site allele frequency for every site was estimated (“-doSaf 1 -fold 1”) using the reference genome as ancestral, while keeping all other parameters as above. Afterward, the SFS and their corresponding bootstraps was estimated for each sample using realSFS and, for each case, the fraction of singletons was calculated. The sample heterozygosities are shown in [Figure S1](#).

Admixture

Using the genotype likelihoods obtained from the ANGSD pipeline, the ancestry clusters and admixture proportions for 48 samples representing all species (for details see [Data S1](#)) were estimated using NGSadmix [20] based on 5.7 million SNPs. Admixture analyses were performed using only markers with minor allele frequency greater than 0.1. We used a range of values for the number of clusters (2-15), to explore the structure in the dataset. To avoid convergence to local optima, the admixture analysis was repeated at least 200 times with different random initial parameter values, and the replicate with the highest likelihood was chosen.

Nuclear genome phylogeny

Using 28 individuals representing all species in this study (for details see [Data S1](#)), nuclear genome phylogenetic reconstruction based on coalescence of gene trees was performed using 100 ASTRAL-II trees [45], and an extended majority rule consensus tree was made with RAxML [46] using default parameters. Each tree was based on gene trees inferred from 5000 regions, each roughly 10 kb long sampled from a consensus genome sequence per individuals generated in ANGSD [44] using the “-doFasta 1” option. Regions with missing data were excluded using trimal [47] under the parameters “-gappyout -resoverlap 0.60 -seqoverlap 60.” Each gene tree was generated in FastTree2 [48] using a generalized time-reversible model for sequence evolution. A cut-off at a minimum of four samples per tree was selected, before generation of individual ASTRAL-II trees. Local posterior probabilities and quartet frequencies for the three possible unrooted resolutions around each internal branch were computed using ASTRAL [59] and visualized using DiscoVista [49]. Two support values are computed on the consensus ASTRAL tree: i) frequency of each branch in the 100 replicates and ii) means of local posterior probability across the 100 replicates. The local posterior probability is computed as the probability that the proportion of gene trees consistent with the bipartition shown in the full phylogeny is greater than 0.33, under a multinomial model with three possible outcomes, each representing a bipartition at the interior branch.

Since the branch lengths in the ASTRAL-II analysis are in terms of coalescent time units, another phylogeny was generated to get branch lengths proportional to evolutionary distances, from 1000 randomly sampled 1 kb regions across the genome using a concatenated analysis in RaxML [46], using a GTR-GAMMA model of sequence evolution.

Species split times

The divergence times between the different species were computed using the two plus two (TT) method [60], which uses a pair of samples, and the distribution of derived alleles at all sites, to compute the split time for a focal population from a contrast population. Specifically, the method uses the counts of sites in the genome where the samples fit into one of 9 configurations, i.e., both samples carry 0 derived alleles, one sample carries 1 derived allele and the other carries 0, and so on, to get an estimate of the time of either sample from the most recent common ancestor of the pair of samples. The method provides two estimates of split times for each pair

of samples, with one sample treated as the focal population and the other as the contrast population. One of the main advantages of this method is that it is not affected by the population size dynamics of the two populations after the split, but it does assume no migration and constant population size in the ancestor of the two populations (before the split).

In order to reduce the number of comparisons in this model, we chose one representative of each population for this analysis, viz., dhole – Beijing Zoo, African hunting dog – Kenya 1, golden jackal – Syria, African golden wolf Northwestern – Morocco, African golden wolf Eastern – Kenya, Ethiopian wolf – Ethiopia, coyote – California, gray wolf European – Spain, gray wolf Asian – Altai, gray wolf American – Greenland and Mexico 1, dog – India 1 and Qatar 2. The TT statistic was computed for each pair of samples, using only scaffolds longer than 1 Mb (705 in all), excluding sites with less than 5x coverage in either sample. The bootstrap estimate of the statistic and its variance was obtained treating each scaffold as a single block [61].

Mitochondrial reconstruction using *de novo* assembly

We used MtArchitect [10] to reconstruct *de novo* the mitochondrial genomes for 17 canids representing all species (for details see Data S1). The genomes were aligned using MAFFT [50] and curated with Jalview [51]. MtArchitect is designed to deal with the presence of numts, by aligning the reads to the mitochondrial and nuclear genome separately, and including only read pairs (or single end reads), where both reads of the pair map unambiguously and with high mapping quality to the mitochondria. We tested a total of 56 phylogenetic models with jmodeltest2 [52] and chose HKY85 with gamma-distributed variation in the substitution rate and a fixed proportion of invariable sites as the most suitable model, which finally was used to construct maximum-likelihood tree using phyML [53]. We generally observed a small amount of undetermined sites, but the two African hunting dogs analyzed displayed poorer alignments and smaller genomes. This is most likely due to the reconstruction biases associated with using a distant reference and a lack of paired-end data to exploit the maximum potential of MtArchitect. Alignment visualization and tree inspection of the reconstructions confirmed that the phylogenetic clustering complied with previously reported data [4]. We observed, however, that the D-loop was particularly enriched in undetermined sites, and aligned notably worse than the remaining sequence. Given its potentially confounding nature and its small contribution to the phylogeny reconstruction when the rest of the sequence is well resolved [10], the D-loop, as well as minor positions containing the majority of the gaps, were manually discarded, resulting in a final 15.435 bp alignment.

D statistics

We used allele frequency-based *D* statistics as implemented in ADMIXTOOLS [23] to evaluate possible gene flow between the different lineages. *D* statistics are based on the observation that, if the given topology (((H1,H2), H3), Outgroup) is correct, then under the null hypothesis of no gene flow between any of the two lineages in the ingroup (H1, H2) and the lineage H3, the number of sites across the genome where the segregation patterns ABBA and BABA occur should be equal in number, as they can arise solely due to incomplete lineage sorting. But the presence of gene flow between H1 and H3 would lead to an increase in the number of BABA sites (H1 and H3 share the same allele B), while gene flow between H2 and H3 would lead to an increase in the number of ABBA sites (H2 and H3 share the same allele B). The *D* statistic measures the disparity between the number of ABBA and BABA sites across the genome to infer gene flow.

To account for the varying depth of coverage of the samples, we used a randomly sampled allele per site instead of called genotypes. Reads with mapping quality lower than 30, bases with quality lower than 20 and sites with coverage lower than 3 were discarded from the analysis. The significance of each test was estimated using a weighted block jackknife procedure over 1 Mb blocks. Deviations from $D = 0$ were presumed significant when the observed *Z*-score was above or below 3.3 ($|Z| > 3.3$). To avoid inflating significance of the tests, only scaffolds 1 Mb or longer (~70% of the genome) were used in the analysis. Tests were performed with combinations of samples as individuals and samples were grouped into categories representing the main genetic clusters (for details see Data S1).

TreeMix

TreeMix [22] was used to infer potential admixture edges in the phylogeny. TreeMix models the correlation of allele frequencies at variable positions across the genome. The correlations that do not fit well under the modeled tree are then corrected for using migration events. We used a randomly sampled allele for each sample and a similar filtering approach as the one described for the *D* statistics tests. Tests were with combinations of samples as individuals and samples grouped into categories representing the main genetic clusters (for details see Data S1). Sites with at least one individual with coverage per group were kept. The final dataset consisted of a total of 834,537 segregating sites. We ran TreeMix on the final dataset assuming 0 to 4 migration edges ($m = 0-4$). For each value of *m*, we ran 100 replicates starting in different seed values and evaluated the replicate with the highest likelihood. Figure S3B shows the best replicate obtained for the graph modeled with four migration edges.

qpGraph

We used qpGraph from the ADMIXTOOLS package [23] to evaluate the relationships between the different species in our samples. In particular, we addressed the question of whether the African golden wolf can be modeled as a hybrid species. qpGraph uses the correlation on all possible *f* statistic tests in a given admixture graph to evaluate its overall fit. The same dataset and filtering parameters used for the *D* statistics tests were used in this analysis. Samples were grouped into clusters representing the main lineages in the admixture graph as indicated in Data S1. First, we started with a tree including the coyote, Ethiopian wolf, gray

wolf and Andean fox and evaluated the most likely branching point for the African golden wolf. Then, we modeled the African golden wolf as a sister clade to all possible internal and external nodes and as an admixed group from all possible node pairs. Finally, we extended our model with an admixture event to account for African golden wolf admixture in the Ethiopian wolf ([Figure 4](#)).

DATA AND SOFTWARE AVAILABILITY

The BioProject accession number for the short read sequences used in this paper is available at the NCBI short read archive under the accession PRJNA494815.

Appendix II

Genomics of Extinction

Genomics of Extinction



Johanna von Seth, Jonas Niemann, and Love Dalén

Abstract Many species went extinct during the Late Pleistocene, including a large proportion of the Earth's megafauna. Recent research on Pleistocene extinctions has started to reveal that species responded individually to environmental fluctuations and human interference. Through paleogenomics, it is now possible to study the extinction process in more detail, which could help disentangle why some species went extinct while others did not. Several species seem to have gone through a sudden decline right before extinction, whereas others reached the point of extinction via a gradual decline. In addition, some species experienced an initial severe bottleneck but survived for thousands of years more at reduced numbers before their final extinction. The use of temporally spaced complete genomes allows for a more direct examination of changes in genomic parameters through time, such as declines in standing genetic variation and accumulation of deleterious mutations, as a consequence of these pre-extinction processes. Additionally, the increasing access to complete ancient genomes will in the future allow researchers to investigate whether species were capable of adapting to environmental changes as well as the small population size that they were subject to prior to the extinction.

Keywords Ancient DNA · Demography · Extinction · Genetic drift · Paleogenomics

J. von Seth

Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden

Division of Systematics and Evolution, Department of Zoology, Stockholm University, Stockholm, Sweden

J. Niemann

Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark

L. Dalén (✉)

Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, Sweden

e-mail: love.dalen@nrm.se

Charlotte Lindqvist and Om P. Rajora (eds.), *Paleogenomics*, Population Genomics [Om P. Rajora (Editor-in-Chief)],

https://doi.org/10.1007/13836_2018_53,

© Springer International Publishing AG, part of Springer Nature 2018

1 Introduction

For any single species, extinction represents the end of evolutionary change. In a short time perspective, extinction represents the disappearance of unique genetic variation and also that an ecological niche is vacated. Put in a wider perspective, however, extinctions are merely fundamental biological processes. Through the history of life on Earth, extinctions have continuously battered millions of species while in parallel having been balanced by a continuous formation of new species.

During the Late Pleistocene (~110–11.7 thousand calendar years before present (cal kyr BP)), extraordinarily many species went extinct, not least a large portion of the megafauna (Cooper et al. 2015). Moreover, several species that survived until present day also went through dramatic population declines in the Late Pleistocene (e.g. Gordon et al. 2016; Johnson et al. 2018). The numerous Late Pleistocene extinctions have often been attributed to climate change or human interference (both directly through hunting and indirectly as the human population expanded and outvalled other species in the competition for resources) (Barnosky et al. 2004; Lorenzen et al. 2011; Cooper et al. 2015; Saltre et al. 2016). However, recent research on Pleistocene extinctions has started to reveal a more complex story, suggesting that one factor alone cannot explain the high number of extinctions. Rather, the emerging pattern is that species responded individually to environmental fluctuations (Lorenzen et al. 2011; Cooper et al. 2015). Finding the causes of extinctions becomes even more challenging with the addition of components such as human interference, interspecific competition and unstable population dynamics.

Using the fossil record to track extinctions, as well as formation of new species, has often been successful. On the other hand, little can be said about the causes behind the extinctions through fossil records alone. Even with the arrival of ancient DNA (aDNA) analyses, it has proved challenging to capture a comprehensive depiction of those last moments before extinction. However, the aDNA research field keeps developing and is no longer dependent on short mitochondrial and nuclear DNA sequences. Instead it is now possible to make use of complete genomes for tracking past biological events in extinct species, and thus the prospects of finding out why some species became extinct while others did not have improved.

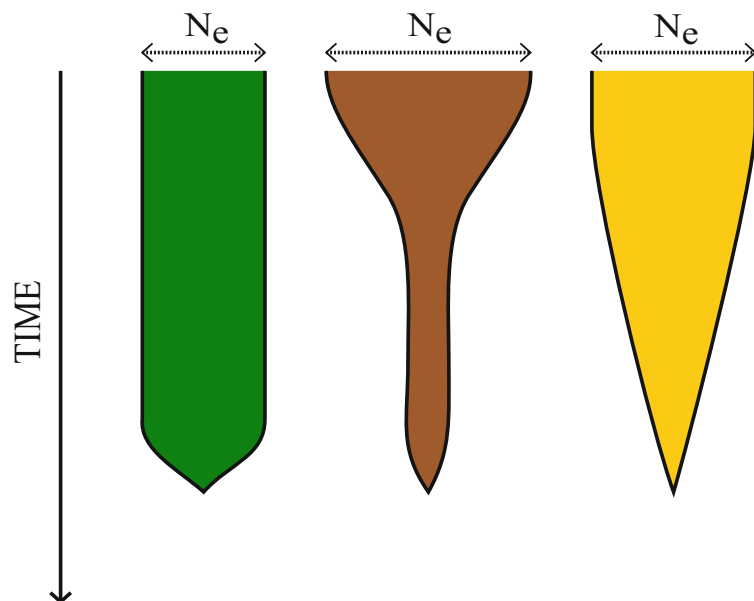
In a time of climate change and numerous species facing extinction, understanding the underlying mechanisms that are pushing some species towards extinction is crucial. Through paleogenomics, it is now possible to study past extinction processes in more detail and to fill in gaps that cannot be filled by morphological data and modern genomics alone (Orlando and Cooper 2014). Furthermore, paleogenomics can be used to add a more complete story of currently threatened species by studying their long-term population histories, as a complement to the snapshot of their present-day genetic status provided by modern DNA.

2 Modes of Decline

All species headed towards extinction first go through a substantial demographic population decline. However, the mode of the decline may vary and is related to the life history of the species as well as the external factors causing the decline (Fig. 1) (Purvis et al. 2000). Some species seem to go extinct without any immediately apparent reasons for the extinction. Others reach the point of extinction in a much slower rate, making it easier to track the decline in the fossil records.

Using the coalescent theory as a starting point, it is possible to investigate the demographic history of a population, since past changes in effective population size can be traced within a population's DNA (Fisher 1930; Wright 1931; Kingman 1982a, b; Kimura 1983). Several demographic history modelling methods have been developed over the past decades, such as the Bayesian skyline plot and the Pairwise Sequentially Markovian Coalescent (PSMC) model, which make use of the coalescent theory within a Bayesian statistical framework (Pybus et al. 2000; Strimmer and Pybus 2001; Drummond et al. 2005; Opgen-Rhein et al. 2005; Heled and Drummond 2008; Minin et al. 2008; Li and Durbin 2011). Shared between the methods is the testing of the hypothesis that a population has been of constant size through time by using the relationship between coalescent time and effective population size (N_e) (Kingman 1982a; Emerson et al. 2001). The relationship states that two randomly chosen DNA sequences in a small population have a higher likelihood of sharing a more recent ancestor (corresponding to fewer substitutional differences between the sequences) than two sequences randomly drawn from a large population (Kingman 1982a). Thus, changes in population size over time will leave signatures in terms of differential substitutions between sequences, where simply put a population decline corresponds to fewer substitutions and vice versa. It is also possible to infer from these methods whether a demographic event happened recently or at a more ancient time point (Emerson et al. 2001).

Fig. 1 Conceptual figure depicting three modes of decline before a species goes extinct; sudden decline (green), terminal refugium decline (brown) and gradual decline (yellow)



It has been argued that these models are sensitive to datasets that are too small or sparsely sampled, population structure, as well as choice of genetic loci and that only certain modes of extinction can be detected using these models (Chang and Shapiro 2016). Still, implementing these methods on high-quality ancient genomes of extinct species can enable an approximation of the mode of population decline before their extinction.

2.1 Sudden Decline

Sudden declines occur when a previously stable population collapses within a few generations, typically due to dramatic changes in its environment. Endemic island populations are particularly susceptible to this type of pre-extinction declines due to their relatively small population sizes, low genetic diversity and adaptation to an environment that is often highly distinct from the mainland (Frankham 1997). Changes to the island environment – notably the effects of human colonization – can have catastrophic consequences. Extensive overhunting and the introduction of predators and diseases led to the demise of the moa (Perry et al. 2014), thylacine (Prowse et al. 2013; Feigin et al. 2017) and dodo (Millberg and Tyrberg 1993) among many others. Even though island species only represent a fraction of all species, 75% of the species that have gone extinct in the past 400 years were endemic to islands (Frankham 1998; Sax and Gaines 2008). Extinctions that are preceded by a sudden decline are however not limited to small island populations. Before the passenger pigeon (*Ectopistes migratorius*) rapidly went extinct at the beginning of the twentieth century, it was a highly abundant species endemic to North America, potentially comprising up to 40% of the continent’s avian population (Schorger 1955; Bucher 1992). In the early and mid-1800s, the population was reported to consist of billions of individuals, constantly migrating between suitable habitats in the search for food and breeding locations while hugely impacting the ecosystems along their path (Schorger 1955; Bucher 1992). The species was however suffering from habitat loss due to human deforestation. Additionally, the large numbers of birds made people associate the species to a pest, and the seemingly never-ending source of cheap meat triggered overhunting once European settlements started taking place in the region (Fulton et al. 2012). Thus, conservation legislation was largely ignored and in just a few decades the species went extinct, with the last individual dying in captivity in 1914 (Schorger 1955; Fulton et al. 2012).

The numerous reports of human overexploitation indicated that this was the main driver of the passenger pigeon extinction, but a study by Hung et al. (2014) revealed that the species regularly went through large population fluctuations in the past. Based on PSMC analyses using three ancient passenger pigeon genomes with a 13- to 20-fold average coverage, the study reported a significant decrease in N_e that started in the last interglacial period (LIG) and reached its lowest number at the last glacial maximum (LGM), before the population once again recovered. They also noted a surprisingly low N_e of the population in comparison with their large census

population size (N_c). The authors thus reasoned that there must have been large fluctuations in N_c that lowered the N_e , following a previously proposed hypothesis, which stated that fluctuations in population size is one of the most important factors explaining variations in the N_e/N_c ratio (Wright 1938; Frankham 1995; Vucetich et al. 1997). To further test the hypothesis of a highly fluctuating N_c , Hung et al. (2014) analysed past fluctuations in various food and habitat resources and argued that these fluctuations would have been large enough to affect the ecosystem's carrying capacity for the passenger pigeon population. Taken together, the researchers suggested that the extinction of the passenger pigeon was a matter of bad timing. The intense hunting coincided with a low N_c during its natural cycle of fluctuations and thereby prevented the population from recovering (Hung et al. 2014).

On the other hand, in a recent study by Murray et al. (2017), analyses of 41 mitochondrial and 4 high-coverage (13- to 51-fold median coverage) nuclear passenger pigeon genomes revealed a stable population size during the approximately 20 kyr prior to the extinction and that the size of the population remained stable even when food and habitat availability was limited. In this study the researchers found indications of strong selection on diversity at linked loci, which could have led to misleading results when estimating population history using PSMC analyses (Murray et al. 2017). Both studies did however agree that human interference may have caused disruptions in the population dynamics that were strong enough to drive the species towards extinction.

2.2 *Gradual Decline*

Other species go through slower, more gradual declines that are often easier to detect than a sudden decline. Stiller et al. (2010) used mitochondrial DNA (mtDNA) to compare the demographic histories of the extinct cave bear (*Ursus spelaeus*) and the extant brown bear (*Ursus arctos*). These two species are especially good to compare since they were closely related, are thought to have had similar life history strategies, and shared habitats when the cave bear was still extant. In the study, by analysing mitochondrial D-loop sequences from 59 temporally spaced cave bears and 40 temporally spaced brown bears, they used the Bayesian coalescent approach to infer the demographic histories of the two species (Drummond et al. 2005; Stiller et al. 2010). From that, they could report that while the extant brown bears appeared to have had a constant and stable demographic history through time, the cave bear population started to decrease some 50 cal kyr BP and then continued to decrease up until their extinction approximately 24 cal kyr BP. Thus, something in the environment appears to have been affecting the cave bears negatively while leaving the brown bears undisturbed. It has been suggested that since cave bears were predominantly herbivorous (Bocherens et al. 1994; Nelson et al. 1998), they were more sensitive to climate changes causing vegetation shifts than were brown bears (Pacher and Stuart 2009). However, the onset of the decline in the cave bear population did not coincide

with extreme changes in vegetation, since the population started to decline long before the onset of the cooling of the climate (Stiller et al. 2010). Another potential difference between the two species was their hibernation strategies. Reports on the relative higher amount of cave bear remains in caves in comparison with brown bear remains imply that cave bears were more reliant on caves for hibernation than were brown bears (Kurtén 1976; Stiller et al. 2010). This might have triggered a competition for access to caves between cave bears, anatomically modern humans and Neanderthals upon the latter two species' arrival to the area that forced cave bears out of the caves where they had to search for other, potentially less favourable, hibernation locations (Grayson and Delpech 2003). Furthermore, a study by Fortes et al. (2016) demonstrated that cave bears might have had a higher tendency to return to their hibernation sites year after year while brown bears did not, which would have intensified the competition between cave bears and human species even more (Fortes et al. 2016).

Taken together, these results suggest that competition of resources between the two bear species, or some other unknown environmental factor, affected the cave bears negatively while leaving the brown bears more or less unaffected long before human arrival and the initiation of extreme climate change. If the subsequent human arrival then forced cave bears out of their caves and the cooling of the climate had started, it is not unlikely that the cave bear population was struggling to remain viable (Stiller et al. 2010). In either case, the cave bear population gradually declined until it went locally extirpated and was later on replaced by another cave bear population. However, this population too could not manage to survive, and the species went globally extinct only a few thousands of years later, at approximately 24 cal kyr BP (Pacher and Stuart 2009; Stiller et al. 2010).

2.3 *Terminal Refugium Decline*

In a third mode of decline, species go through severe population bottlenecks, leaving just a portion of the original population behind, but still survive for thousands of years more. The well-studied woolly mammoth (*Mammuthus primigenius*) seems to have had a quite stable population size during the Late Pleistocene (Palkopoulou et al. 2013, 2015). However, the last surviving mainland mammoth population disappeared approximately 11 cal kyr BP (Nikolskiy et al. 2011). Thereafter, the last remaining populations were situated on the remote St Paul Island and Wrangel Island for another approximately 5 and 6 kyr, respectively (Vartanyan et al. 1993; Veltre et al. 2017). Analyses of the demographic history of the last surviving population, the one located on Wrangel Island, have revealed a dramatic population bottleneck some 8 kyr before their actual extinction (around the same time as the elimination of the mainland population) (Palkopoulou et al. 2015). Although the population subsequently survived for several thousand years more in this terminal refugium, several studies have shown that the population suffered from the bottleneck as well as ensuing small population size, in terms of loss of genetic diversity in

both coding and non-coding regions of the genome (Lister and Stuart 2008; Palkopoulou et al. 2015; Pečnerová et al. 2016; Rogers and Slatkin 2017).

Lister and Stuart (2008) pointed out that this time lag between an extreme range contraction into a terminal refugium and the final extinction is similar to what has been termed an ‘extinction lag’. This phenomenon has already been described for areas that have gone through fragmentation in modern times. Populations that appear to have remained viable post fragmentation are when further investigated discovered to be at risk of future extirpation, mainly due to gene flow barriers, increased demographic allee effects (positive density dependence), as well as decreased genetic diversity within each fragment of the population and a decreased carrying capacity of the area (Brooks et al. 1999; Dixo et al. 2009). Thus, the ‘extinction lag’ or ‘extinction debt’ refers to the future ecological and genetic cost of the fragmentation (Tilman et al. 1994; Lister and Stuart 2008). So while the last woolly mammoth population survived for some thousands years more, the fact that the Wrangel Island population was the last extant population with no possibilities for genetic rescue through gene flow into the population, as well as apparent negative genetic effects of the bottleneck, implies that the population may not have been large enough to be viable.

3 Local Population Turnovers

One of the most significant insights in paleoecology obtained through aDNA analyses is the identification of temporal population discontinuity within specific geographic regions. Such lack of continuity has either been through partial replacement of resident populations (Skoglund et al. 2012) or through extinctions followed by recolonization from genetically different source populations (Barnes et al. 2002). The latter type of population turnovers, extinctions/recolonizations, seem to have been common during the Late Pleistocene and have been described for a wide variety of wild animals as well as humans (e.g. Hofreiter et al. 2007; Leonard et al. 2007; Campos et al. 2010; Posth et al. 2016). The most pronounced example of extinctions/recolonizations comes from the collared lemming (*Dicrostonyx torquatus*), which was a keystone small herbivore that inhabited the Late Pleistocene Eurasian steppe tundra. Analyses of mtDNA sampled across a broad geographical scale and covering the last 50 kyr have indicated that the collared lemming went through a series of population extinctions throughout western Eurasia, with subsequent and repeated recolonizations from further east (Brace et al. 2012; Palkopoulou et al. 2016). These extinctions imply an unexpected instability of the Late Pleistocene ecosystem during the last Ice Age, likely caused by brief warm periods (Dansgaard-Oeschger events).

Most previous paleogenetic studies that have identified local extinctions have been based on analyses of mtDNA. However, mtDNA has limited power since it only provides information on a single gene tree, which may deviate from the species phylogeny due to its maternal inheritance, lineage sorting and introgression.

Moreover, the absence of recombination in mtDNA makes it sensitive to hitchhiking selection (Galtier et al. 2009). Because of this, future paleogenetic studies will likely use genome-wide data to revisit earlier mtDNA-based studies to re-examine the existence and timing of local extinctions. This has recently been done for Neanderthals, where analyses of multiple genomes (Hajdinjak et al. 2018) led to support for an earlier hypothesis that Neanderthals in western Europe went through a population turnover (Dalén et al. 2012). Moreover, a recent study on Paleolithic humans using genome-wide data (Fu et al. 2016) indicated that a previously identified mtDNA replacement (Posth et al. 2016) during the Allerød interstadial likely was caused by migration rather than extinction/recolonization.

4 Genomic Consequences of Demographic Declines

Regardless of the mode of decline, the mere decrease in size of a population increases its risk for extinction simply because small populations are more vulnerable to stochastic events, be they demographic, environmental or genetic (Frankham 2005). This increased risk of extinction related to decreased population size is known as the small population paradigm and was first defined by Caughley (1994). In terms of genetics, loss of genetic diversity and the exposure of recessive deleterious alleles are thought to be the most serious threats for such small populations.

In theory, loss of genetic diversity is inversely proportional to the effective population size (Frankham 2005). This is due to genetic drift, i.e. the random fixation of alleles that occurs within all populations but becomes much stronger in small ones. As a population declines, the fixation of alleles and consequently the loss of all other alleles at the corresponding loci increase (Wright 1950). This loss of standing genetic variation may in turn limit the evolutionary potential of the population (Kohn et al. 2006; Willi et al. 2006), thus reducing its capacity to evolve in response to environmental change, competition or disease. At the same time, inbreeding is likely to increase in a declining population even if mating occurs randomly, simply because the number of non-related potential mating partners decreases. While this does not necessarily result in a loss of genetic variation in the population, other than the loss that can be explained by genetic drift, inbreeding does decrease the within-individual genetic variation as more loci are becoming homozygous when individuals are more often inheriting alleles that are identical by descent (Crow 2010).

Both loss of genetic diversity and inbreeding can cause a lowered individual fitness in the population. This can take place either through an increased homozygosity at loci where heterozygote genotypes have an advantage over homozygote genotypes as, for example, in the major histocompatibility complex (MHC) (Carrington et al. 1999; Bernatchez and Landry 2003; Spurgin and Richardson 2010) or through an increased exposure of recessive deleterious alleles in homozygotes (Charlesworth and Charlesworth 1999). Recessive deleterious alleles are seldom exposed to selection in large populations and can therefore remain fairly unnoticed within a population for a relatively long time. However, since individuals

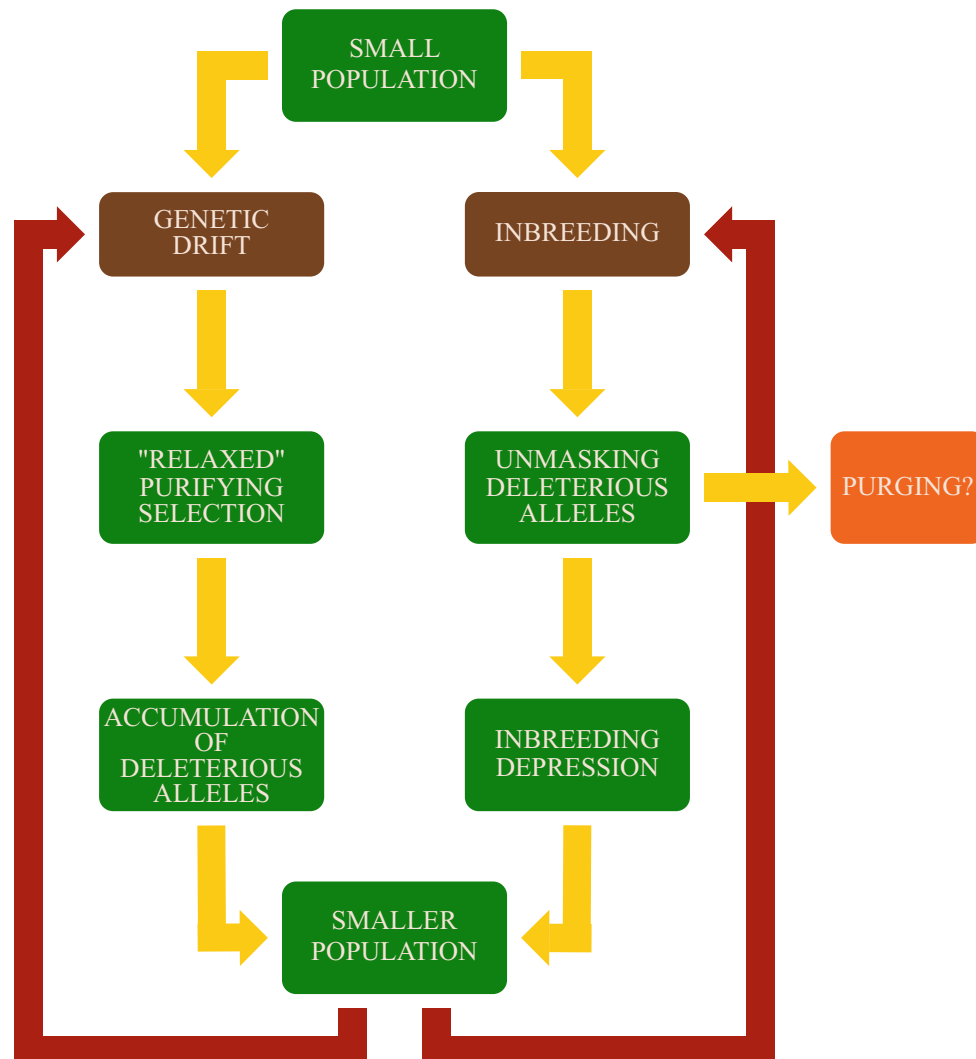


Fig. 2 Genetic processes in small populations

in declining and inbred populations more often become homozygous at loci, including those that are carrying harmful alleles, the individual fitness in small populations is expected to become reduced (Fig. 2). When a population has reached this stage, i.e. several individuals showing clear signs of lowered fitness due to inbreeding, the population is experiencing an inbreeding depression (Charlesworth and Charlesworth 1999). However, as long as variation remains in the population at loci where recessive deleterious alleles are located, these alleles can potentially be purged from the population through purifying selection.

4.1 Purifying Selection

In theory, if a population is maintained at low numbers so that already existing deleterious recessive alleles (and novel ones that originate through mutation)

become exposed, these alleles can be purged from the population through purifying selection (Lynch et al. 1995b; Wang et al. 1999). Since the alleles, when expressed, are expected to cause a lowered fitness of the individuals carrying them, they will be less likely to contribute with genetic material to the next generation in comparison with individuals not carrying the harmful alleles. Thus after a few generations, the population can in theory have a higher individual fitness than it had right before the harmful alleles started to become expressed.

In a study that investigated the effectiveness of purging, it was found that the genetic basis of inbreeding depression greatly affected the outcome of purging (Hedrick 1994). Generally, if the cause of the inbreeding depression was genetic load of lethal alleles rather than slightly deleterious alleles, these lethal alleles could quickly become purged from the population without a highly increased risk of extinction (Hedrick 1994). The opposite was true if the inbreeding depression was caused by slightly deleterious alleles because of the high risk of these alleles becoming fixed via genetic drift (Hedrick 1994). Additionally, concern has been raised regarding whether purging could decrease the standing genetic variation of a population by simultaneously allowing for a decrease in genetic variation at other, non-lethal, loci as a consequence of the maintained small effective population size, thereby decreasing the population's evolutionary potential (Hedrick and Miller 1992; Hedrick 1994).

Several studies dedicated to investigating the efficiency of purging in small and inbred populations have presented contradictive results (e.g. Bryant et al. 1990; Kalinowski et al. 2000). To summarize, it seems that the effectiveness of purging is highly relative, dependent on how purging is measured, and the measurements are sensitive to confounding factors such as temporal environmental changes (Bryant et al. 1990; Barrett and Charlesworth 1991; Hedrick and Kalinowski 2000; Kalinowski et al. 2000).

4.2 The Theory of Mutational Meltdown

Genetic drift can become so strong in small populations that instead of purifying selection removing new detrimental mutations that appear in the population, these mutations become fixed (Fig. 2) (Lynch and Gabriel 1990; Hedrick 1994; Lynch et al. 1995a). Once fixed within a reproductively isolated population, they are bound to be carried onto the following generations unless new mutations appear. As more harmful mutations are accumulating for each generation, the population size is likely to decrease even further (Lynch and Gabriel 1990; Wang et al. 1999; Hedrick and Kalinowski 2000). This decline in population size will in turn lead to further increased strength of genetic drift and additional fixation of detrimental mutations, thus resulting in a negative feedback loop for the population (Lynch and Gabriel 1990; Lynch et al. 1995a; Gaggiotti 2003; Charlesworth and Willis 2009). This phenomenon, where the increasing strength of genetic drift causes a negative feedback loop in

population size, has been termed the population mutational meltdown by Lynch and Gabriel (1990).

4.3 *Fragmentation of Populations*

In many extinct and endangered species, demographic declines also lead to population fragmentation, which in turn can lead to increased genetic drift and inbreeding within each subpopulation (Brooks et al. 1999; Dixo et al. 2009; Frankham et al. 2017). While fragmentation can have natural causes, e.g. as rising sea levels create isolated islands with small isolated populations, human-caused fragmentation is one of the main anthropogenic threats for species and population survival in modern times (Haddad et al. 2015). The split of one population into several small populations, e.g. due to loss of suitable habitats or newly introduced barriers, at best only limits gene flow and at worst eliminates any possibilities for gene flow between the populations (Goossens et al. 2005). Regardless, the smaller population sizes caused by fragmentation increases the vulnerability to and effects of stochastic events, including genetic drift and inbreeding (Dixo et al. 2009; Pečnerová et al. 2016).

5 **Paleogenomics to Study Effects of Decline**

5.1 *Genetic Parameters*

One of the most important aspects of assessing the genomic consequences of a demographic decline is to determine the pre-decline status of important genomic erosion parameters, such as genome-wide diversity, inbreeding levels, as well as the amount genetic load within a population (Fig. 3). Several recent studies have indicated that there are some discordances in the theoretically acknowledged correlation between population size and the level of heterozygosity (Leffler et al. 2012; Díez-del-Molino et al. 2018). For example, the Sumatran orangutan (*Pongo abelii*) and the bonobo (*Pan paniscus*) are two currently endangered species, the former critically so (Prado-Martinez et al. 2013; IUCN 2016). Still, even though the current population sizes of the two species are similar, the Sumatran orangutan population has approximately three times higher genome-wide heterozygosity than the bonobo (Leffler et al. 2012; Prado-Martinez et al. 2013). Similarly, the giant panda, classified as vulnerable according to the IUCN red list, has significantly higher heterozygosity than humans (Cho et al. 2013; IUCN 2016). It has therefore been suggested that ancient bottlenecks and different life history strategies among species are likely to give rise to varying pre-decline levels of diversity, inbreeding and genetic load. In order to be able to distinguish the genomic effects of pre-extinction declines from the effects of more ancient events and life history traits, analysing pre-decline genomes

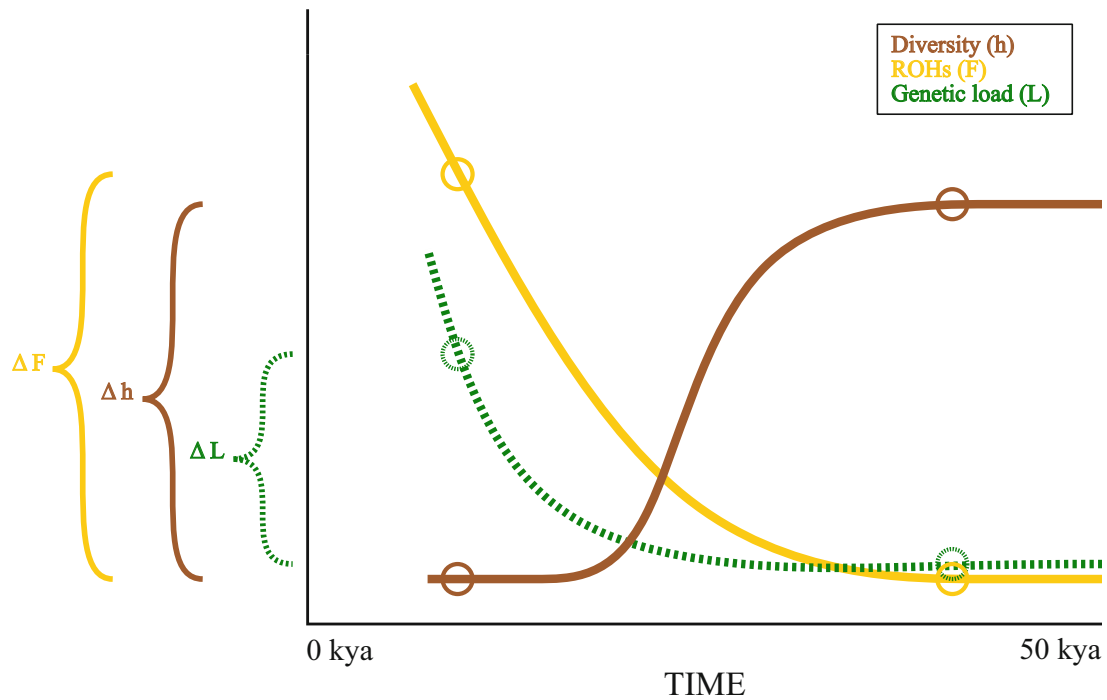


Fig. 3 Conceptual figure showing how pre-decline sampling enables direct estimates of the changes in genomic erosion parameters, such as genome-wide diversity (brown), inbreeding levels measured as amount of runs of homozygosity (ROH) (yellow) and genetic load (green), as a direct consequence of the pre-extinction decline

using, for example, century-old museum specimens can be a valuable approach in conservation genomics (Díez-del-Molino et al. 2018).

5.2 Genomes

When inferring past demographic events and conducting population genetic analyses based on ancient data, the most common DNA marker previously used has been mtDNA, such as the D-loop (Hofreiter et al. 2004; Valdiosera et al. 2007; Lorenzen et al. 2011). There are several benefits of using mtDNA, with one important benefit being the much higher copy number of the mtDNA genome in each cell in comparison with the nuclear genome (Clayton 1982). However, while all mtDNA is inherited from a single parent, the nuclear genome comprises several million independently, biparentally inherited loci and will therefore facilitate greater statistical power for the conduction of population genetic analyses than does mtDNA (Shapiro and Hofreiter 2014). For example, by using genome-wide single nucleotide polymorphism (SNP) sites, it is possible to estimate individual levels of heterozygosity in the population, making estimates of changes in genetic diversity more robust (Park et al. 2015).

Using whole genomes is of high importance when evaluating genetic consequences of declines, since such data in parallel enables analyses of effects on standing genetic variation, effects on fitness by analysing functional regions as well as genome-wide scans for runs of homozygosity (ROHs), i.e. long genomic fragments completely depleted from diversity (Broman and Weber 1999; McQuillan et al. 2008; Kardos et al. 2016). While generating high-coverage genome data is always preferential, there are some unneglectable obstacles for this goal when working with ancient material. First of all, as an organism dies, the natural post-mortem degradation of DNA is initiated, through, for example, enzymatic processes occurring shortly post-mortem, hydrolytic strand cleavage, lesions induced by oxygen-free radicals and cytosine deamination (Pääbo 1989; Pääbo et al. 2004; Wandeler et al. 2007; Skoglund et al. 2014). The rate of DNA degradation is to a large extent dependent on the environment in which the remains are preserved (in general, cold and dry environments can facilitate a slower rate of degradation) (Lindahl 1993). Secondly, as a consequence of this, the quality and the amount of endogenous DNA can vary greatly between different samples, and ancient samples are known to be highly sensitive to modern DNA contamination (Pääbo 1989; Pääbo et al. 2004).

With good aDNA preservation, however, high-coverage genome data can be generated. In this scenario, given the large number of independently inherited loci, only a handful of genomes or so are sufficient for inferring the extent of inbreeding and loss of genomic diversity in a population prior to its extinction (Shapiro and Hofreiter 2014). Quantification of genome-wide diversity as well as the inbreeding levels based on ROHs were recently done in two different studies of the extinct Denisovans and woolly mammoths, respectively (Meyer et al. 2012; Palkopoulou et al. 2015). Here, in-depth analyses such as long-term demographic changes, as well as individual genome-wide heterozygosity and inbreeding estimates, were generated (Meyer et al. 2012; Palkopoulou et al. 2015). Using the software mlRho (Haubold et al. 2010), the two studies reported extremely low to low heterozygosity in one 30-fold coverage Denisovan genome and one 17-fold coverage woolly mammoth genome, respectively. In both cases, the low heterozygosity could not be explained by inbreeding of immediate ancestors since no unusually long ROHs could be detected in the Denisovan genome and in the woolly mammoth genome the lengths of the ROHs were relatively short, a pattern typical for when mating between distant relatives has been taking place for several generations rather than close relatives having mated more recently (Broman and Weber 1999; Gibson et al. 2006; Meyer et al. 2012; Palkopoulou et al. 2015).

With poor aDNA quality on the other hand, only low-coverage genome data can be generated. As a consequence, the analyses will be constrained to population-level analyses. Still, a lot of new insights can come from these types of analyses, like in the case of camel evolutionary history. Through high-coverage mitochondrial genomes from two ancient Yukon *Camelops* specimens and low-coverage nuclear genomes from one of these individuals, results contradicting previous morphology-based phylogenetic analyses of the relationships between different camels species could

be reported (Heintzman et al. 2015). As another example, by using the high-coverage mitochondrial and low-coverage nuclear genome of a wolf sample dated to 35 cal kyr BP, Skoglund et al. (2015) found support for the divergence between wolves and dogs haven taken place $\sim 27\text{--}40$ kyr earlier than previously suggested.

Up until recently, few high-quality genomes (>10 -fold average genome coverage) of extinct species had successfully been generated. In 2008, the first report of an attempt to sequence an extinct mammalian genome was published along with a partial genome sequence covering roughly 70% of the genome, this by sequencing DNA from woolly mammoth hair (Miller et al. 2008). Subsequently in 2015, two complete woolly mammoth genomes were generated with a 17-fold and 11-fold average coverage, respectively (Palkopoulou et al. 2015). The first ancient human genome was sequenced in 2010, with a 20-fold average coverage across 79% of the genome of a Paleo-Eskimo human (Rasmussen et al. 2010). In 2014, the complete genome sequence of one Neanderthal (52-fold average coverage) (Prüfer et al. 2014) and four passenger pigeons (5–20-fold coverage) (Hung et al. 2014) were generated. One year later, Park et al. (2015) managed to sequence the complete genome of the extinct aurochs (*Bos primigenius*) (six-fold average coverage). In 2017, two additional high-coverage passenger pigeon genomes (51- and 41-fold median coverage) (Murray et al. 2017) as well as the complete genome (43-fold average coverage) of the Tasmanian tiger (*Thylacinus cynocephalus*) (Feigin et al. 2017) were published. Thus, the recent advances in sequencing technologies now means that the possibilities to analyse genomes of extinct species have increased immensely and along with it comes the increasing potential for understanding pre-extinction genetic processes.

6 Future Challenges

6.1 Reference Genomes

When working with genomic data generated from extant species, there are either de novo assembled reference genomes already available for mapping the sequencing reads or such de novo genome assemblies can relatively easily be generated for the study species in question (Li et al. 2010). However, since de novo assembly requires high-quality DNA to generate high coverage across the entire genome, this is considered impossible for extinct species. Instead, one has to rely on the most closely related extant species as a reference for mapping sequencing reads. Since the most closely related species can often correspond to a divergence time of millions of years, aligning sequencing reads from an extinct species is not trivial and can often result in gaps in parts of the genome that are non-existing in the genome of the related extant species (Prüfer et al. 2010; Shapiro and Hofreiter 2014; Richmond et al. 2016). It is however still possible to conduct some analyses without a proper reference genome, such as changes in genome-wide diversity, while other important biological questions such as functional genomics may be more difficult to answer.

6.2 *Sequence Analysis*

With whole-genome sequencing comes the generation of massive amounts of data, which has led to the development of bioinformatics softwares that can be applied to such large data sets. Numerous different pipelines and bioinformatics tools are now available for filtering away low-quality sequencing reads (e.g. John 2011; Bolger et al. 2014), mapping high-quality reads to a reference genome (e.g. Li et al. 2009; Li 2013), consensus sequence generation and the conduction of data analyses to statistically test a large range of biological questions (e.g. McKenna et al. 2010; DePristo et al. 2011; Lunter and Goodson 2011). However, analysing whole genomes from ancient samples requires pipelines and bioinformatics tools that can handle data generated from poor-quality DNA and that can distinguish endogenous DNA from contaminant DNA, as well as identify nucleotide changes caused by post-mortem DNA damage. While there are some best practices available (Mourier et al. 2012) and tools applicable to low-quality data are on the rise (e.g. Schubert et al. 2014; Peltzer et al. 2016), a general issue with bioinformatics software development and usage concerns software version updates. As an increasing amount of researchers apply various bioinformatics tools to their specific data sets, new unforeseen issues arise leading to version updates of the tools to correct for these issues. Thus, during the course of a research project, the initial version of a program might have been updated several times making the first analyses irrelevant by the end of the project. Maintenance of previous versions is moreover often abandoned in favour of newer versions, and most research groups include custom-made scripts or programs specifically designed for their data, making it difficult to replicate analyses from previously published studies.

The field of paleogenomics is expanding rapidly, especially due to the increasing possibilities to generate large data sets from degraded DNA, and good practice guidelines for processing and analysing this type of data are desirable.

6.3 *De-extinction*

With the rise of whole-genome sequencing, advanced laboratory techniques that enable in vitro fertilization and cloning, as well as genetic engineering techniques to edit genomes such as CRISPR, the debate about bringing back extinct species has intensified (Jinek et al. 2012). Some argue that it is our moral responsibility to bring back the species we have once driven to extinction. Others suggest that de-extinction could be used to counteract environmental change by bringing back key species to important ecosystems, such as grasslands, where, e.g. woolly mammoths could contribute by halting releases of carbon from soils (Zimov et al. 2012). The advancement of cloning has yielded several successful cloned animals over the past decade, e.g. the successful generation of an afghan dog puppy clone and the creation of a viable mouse clone from a dead mouse donor that had been frozen

at -20°C for 16 years, both of these through somatic cell nuclear transfer (SCNT) (Lee et al. 2005; Wakayama et al. 2008). However, bringing back extinct species is even more complicated. For example, the first trial of bringing back an extinct species through cloning resulted in a Pyrenean ibex clone (*Capra pyrenaica pyrenaica*) that survived for only a few minutes (Folch et al. 2009).

Since it is today possible to sequence more or less complete genomes of extinct species, these genomes could theoretically be synthetically generated and introduced into empty nuclei of egg cells carried by surrogate mothers of closely related species. Since DNA goes through post-mortem degradation, however, the probability of generating high-quality, high-coverage genomes of extinct species decreases with the age of the specimen (Wandeler et al. 2007; Skoglund et al. 2014). In other words, the highest-quality samples are also going to be comparatively close in time to the extinction event and are consequently likely to carry genomes that are depleted of genetic diversity, contain high numbers of ROHs, and most importantly may have an excess of fixed deleterious mutations. With genetic engineering methods, such as CRISPR-Cas9, it is however to some extent possible to circumvent this issue by replacing harmful mutations (Jinek et al. 2012).

The typical read length of degraded DNA poses another challenge, as it is not possible to align fragmented ancient DNA sequences to the regions of the genome that are highly repetitive or duplicated, as this requires much longer sequencing reads than what can be retrieved from ancient samples (Treangen and Salzberg 2012). It is therefore hopeless to retrieve the complete genetic information of historical and ancient individuals, even if the DNA is relatively well preserved. Apart from the previously mentioned technical challenges, it is thus impossible to recreate a perfect clone of long-extinct individuals, as we have no knowledge of a significant part of the genome. Most de-extinction efforts therefore focus on creating hybrids that retain some key phenotypes from the extinct species. In the case of the proposed woolly mammoth hybrid, only 45 genes have been modified so far to carry woolly mammoth alleles in the Asian elephant genome (Campbell and Whittle 2017). It is unclear how mammoth-like such a hybrid would be in appearance and behaviour and whether the outcome justifies the immense efforts.

Besides molecular and genetics issues that need to be addressed before de-extinction can be realized, there are other ecological and behavioural aspects that might affect the outcome. Would the closest living relative be able to teach a newborn the way of life of another species? Are there any remaining suitable habitats for an extinct species in modern times? Are the external factors that originally contributed to the extinction gone?

Instead, perhaps the idea of de-extinction would be best applied to currently threatened species, by using genetic engineering to bring back genetic diversity and ancient, healthier allele variants (Shapiro 2017). The endangered Tasmanian devil (*Sarcophilus harrisii*), for example, is suffering from low genetic diversity not least in the MHC complex and is severely affected by a transmittable cancer (Siddle et al. 2010). Here, the use of genomic data from healthy, long-dead individuals and the CRISPR-Cas9 technique could potentially provide an alternative way to obtain a genetic rescue effect (Tallmon et al. 2004), for example, by adding diversity to the

immune system within the population, thereby increasing its resilience towards the cancer and in the long-term extinction (Jinek et al. 2012).

6.4 Adaptation During the Extinction Process

While the majority of studies on extinct species focus on the cause of extinction and population demography, not much is known about whether populations are capable of adapting to the additional challenges of inbreeding depression or accumulation of deleterious alleles prior to extinction. Species such as the cheetah, channel island fox and the wandering albatross went through severe bottlenecks that led to a very low genetic heterozygosity in the present-day population, yet these species have persisted in relatively stable populations for thousands of years (Milot et al. 2007; Dobrynin et al. 2015; Robinson et al. 2016). It is not clear whether these species currently are in terminal refugia or whether they have escaped from the extinction vortex at the time of the bottleneck due to stochastic factors or species-specific behavioural strategies. Alternatively, the survival of these species could be explained by them having been able to genetically adapt to a small population size during the decline.

This question could in the future be addressed with paleogenomics by comparing the adaptive potential of in-decline populations that went extinct with those populations that persisted after the bottleneck. Genes under positive selection that enabled the population to be less vulnerable to the effects of the extinction vortex, as well as possible decreases in genetic load due to purifying selection, might also be detectable by comparing pre-decline with post-decline individuals. As neither population size nor low heterozygosity is a good proxy for the immediate extinction risk of a species (Díez-del-Molino et al. 2018), the additional information of an estimated adaptive potential could be valuable in conservation to prioritize especially vulnerable populations.

The ability of small populations to adapt to changes in the environment is especially relevant today, given the ongoing changes in climate that is likely to put additional stress on endangered species throughout the world. Paleogenomic analyses on species that became extinct in conjunction with the severe changes in climate that took place at the end of the Pleistocene could provide highly valuable information in this context. In particular, knowledge on the extent to which species were able to adapt to prehistoric temperature increases may help conservation biologists to predict how resilient present-day species will be to future climate change.

7 Conclusions and Future Perspectives

Extinction is one of the most fundamental processes in evolution. However, despite its importance to better understand today's biodiversity crisis, little is known about the demographic trajectories that precede extinction as well as how population declines affect genomic parameters. Paleogenomic analyses of taxa that went extinct in the past offer a unique opportunity to investigate how species demographics changed prior to their disappearance. Moreover, serially sampled genomic data can be used to test whether genome erosion in itself can contribute to the extinction process. Although only a few ancient genomes from wild species have been sequenced to date, this is probably going to change in the near future given the continuous decrease in high-throughput DNA sequencing costs and ongoing developments in ancient DNA recovery methods. It therefore seems highly likely that genomes from several additional extinct species will soon be made available. While this will inevitably result in an increased debate about the possibility of resurrecting these species, comparisons of genomes from multiple extinct species with those from their closest living relatives will also help emphasize the importance of having suitable genome assemblies from related extant species to use for reference-based mapping. In the near future, we are also likely to see comprehensive genomic catalogues for several extinct species, comprising multiple genomes sampled through time leading up to the extinction. Such genomic catalogues will enable detailed studies of how changes in the environment and population size have affected microevolutionary processes through time.

References

- Barnes I, Matheus P, Shapiro B, Jensen D, Cooper A. Dynamics of Pleistocene population extinctions in Beringian brown bears. *Science*. 2002;295:2267–70.
- Barnosky AD, Koch PL, Feranec RS, Wing SL, Shabel AB. Assessing the causes of Late Pleistocene extinctions on the continents. *Science*. 2004;306:70–5.
- Barrett SCH, Charlesworth D. Effects of a change in the level of inbreeding on the genetic load. *Nature*. 1991;352:522–4.
- Bernatchez L, Landry C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? *J Evol Biol*. 2003;16:363–77.
- Bocherens H, Fizet M, Mariotti A. Diet, physiology and ecology of fossil mammals as inferred from stable carbon and nitrogen isotope biogeochemistry: implications for Pleistocene bears. *Palaeogeogr Palaeoclimatol Palaeoecol*. 1994;107:213–25.
- Bolger AM, Lohse M, Usade B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
- Brace S, Palkopoulou E, Dalen L, Lister AM, Miller R, Otte M, Germonpré M, Blockley SPE, Stewart JR, Barnes I. Serial population extinctions in a small mammal indicate Late Pleistocene ecosystem instability. *Proc Natl Acad Sci U S A*. 2012;109:20532–6.
- Broman KW, Weber JL. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am J Hum Genet*. 1999;65:1493–500.

- Brooks TM, Pimm SL, Oyugi JO. Time lag between deforestation and bird extinction in tropical forest fragments. *Conserv Biol.* 1999;13:1140–50.
- Bryant EH, Meffert LM, McCommas SA. Fitness rebound in serially bottlenecked populations of the house fly. *Am Nat.* 1990;136:542–9.
- Bucher EH. The causes of extinction of the passenger pigeon. *Curr Ornithol.* 1992;9:1–36.
- Campbell DL, Whittle PM. Three case studies: aurochs, mammoths and passenger pigeons. In: *Resurrecting extinct species.* Cham: Palgrave MacMillan; 2017.
- Campos PF, Willerslev E, Sher A, Orlando L, Axelsson E, Tikhonov A, Aaris-Sorensen K, Greenwood AD, Kahlke RD, Kosintsev P, Krakhmalnaya T, Kuznetsova T, Lemey P, MacPhee R, Norris CA, Shepherd K, Suchard MA, Zazula GD, Shapiro B, Gilbert MTP. Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (*Ovibos moschatus*) population dynamics. *Proc Natl Acad Sci U S A.* 2010;107:5675–80.
- Carrington M, Nelson GW, Martin MP, Kissner T, Vlahov D, Goedert JJ, Kaslow R, Buchbinder S, Hoots K, O'Brien SJ. HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage. *Science.* 1999;283:1748–52.
- Caughley G. Directions in conservation biology. *J Anim Ecol.* 1994;63:215–44.
- Chang D, Shapiro B. Using ancient DNA and coalescent-based methods to infer extinction. *Biol Lett.* 2016;12:20150822.
- Charlesworth B, Charlesworth D. The genetic basis of inbreeding depression. *Genet Res.* 1999;74:329–40.
- Charlesworth D, Willis JH. The genetics of inbreeding depression. *Nat Rev Genet.* 2009;10:783–96.
- Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, Oh S, Kim HM, Jho S, Kim S, Shin YA, Kim BC, Kim H, Kim CU, Luo SJ, Johnson WE, Koepfli KP, Schmidt-Kuntzel A, Turner JA, Marker L, Harper C, Miller SM, Jacobs W, Bertola LD, Kim TH, Lee S, Zhou Q, Jung HJ, Xu X, Gadhvi P, Xu P, Xiong Y, Luo Y, Pan S, Gou C, Chu X, Zhang J, Liu S, He J, Chen Y, Yang L, Yang Y, He J, Liu S, Wang J, Kim CH, Kwak H, Kim JS, Hwang S, Ko J, Kim CB, Kim S, Bayarlkhagva D, Paek WK, Kim SJ, O'Brien SJ, Wang J, Bhak J. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun.* 2013;4:2433.
- Clayton DA. Replication of animal mitochondrial DNA. *Cell.* 1982;28:693–705.
- Cooper A, Turney C, Hughen KA, Brook BW, McDonald HG, Bradshaw CJA. Abrupt warming events drove Late Pleistocene Holarctic megafaunal turnover. *Science.* 2015;349:602–6.
- Crow JF. Wright and Fisher on inbreeding and random drift. *Genetics.* 2010;184:609–11.
- Dalén L, Orlando L, Shapiro B, Brandström-Durling M, Quam R, Gilbert MTP, Fernández-Lomana JCD, Willerslev E, Arsuaga JL, Götherström A. Partial genetic turnover in Neandertals: continuity in the east and population replacement in the west. *Mol Biol Evol.* 2012;29:1893–7.
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43:491–8.
- Díez-del-Molino D, Sánchez-Barreiro F, Barnes I, Gilbert MTP, Dalén L. Quantifying temporal genomic erosion in endangered species. *Trends Ecol Evol.* 2018;33:176–85.
- Dixo M, Metzger JP, Morgante JS, Zamudio KR. Habitat fragmentation reduces genetic diversity and connectivity among toad populations in the Brazilian Atlantic Coastal Forest. *Biol Conserv.* 2009;142:1560–9.
- Dobrynin P, Liu S, Tamazian G, Xiong Z, Yurchenko AA, Krashennikova K, Kliver S, Schmidt-Kuntzel A, Koepfli KP, Johnson W, Kuderna LF, Garcia-Perez R, Manuel M, Godinez R, Komissarov A, Makunin A, Brukhin V, Qiu W, Zhou L, Li F, Yi J, Driscoll C, Antunes A, Oleksyk TK, Eizirik E, Perelman P, Roelke M, Wildt D, Diekhans M, Marques-Bonet T, Marker L, Bhak J, Wang J, Zhang G, O'Brien SJ. Genomic legacy of the African cheetah, *Acinonyx jubatus*. *Genome Biol.* 2015;16:277.
- Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 2005;22:1185–92.

- Emerson BC, Paradis E, Thébaud C. Revealing the demographic histories of species using DNA sequences. *Trends Ecol Evol.* 2001;16:707–16.
- Feigin CY, Newton AH, Doronina L, Schmitz J, Hipsley CA, Mitchell KJ, Gower G, Llamas B, Soubrier J, Heider TN, Menzies BR, Cooper A, O'Neill RJ, Pask AJ. Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. *Nat Ecol Evol.* 2017;2:182–92.
- Fisher RA. *The genetical theory of natural selection.* Oxford: Clarendon Press; 1930.
- Folch J, Cocero MJ, Chesné P, Alabart JL, Dominguez V, Cognie Y, Roche A, Fernandez-Arias A, Marti JI, Sanchez P, Echegoyen E, Beckers JF, Bonastre AS, Vignon X. First birth of an animal from an extinct subspecies (*Capra pyrenaica pyrenaica*) by cloning. *Theriogenology.* 2009;71:1026–34.
- Fortes GG, Grandal-d'Anglade A, Kolbe B, Fernandes D, Meleg IN, Garcia-Vazquez A, Pinto-Llona AC, Constantin S, de Torres TJ, Ortiz JE, Frischauf C, Rabeder G, Hofreiter M, Barlow A. Ancient DNA reveals differences in behaviour and sociality between brown bears and extinct cave bears. *Mol Ecol.* 2016;25:4907–18.
- Frankham R. Effective population size/adult population size in wildlife: a review. *Genet Res.* 1995;66:95–107.
- Frankham R. Do island populations have less genetic variation than mainland populations? *Heredity.* 1997;78:311–27.
- Frankham R. Inbreeding and extinction: island populations. *Conserv Biol.* 1998;12:665–75.
- Frankham R. Genetics and extinction. *Biol Conserv.* 2005;126:131–40.
- Frankham R, Ballou JD, Ralls K, Eldridge M, Dudash MR, Fenster CB, Lacy RC, Sunnucks P. *Genetic management of fragmented animal and plant populations.* Oxford: Oxford University Press; 2017.
- Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwangler 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 AP, Drozdov N, Slavinsky V, Tsybankov A, Cremonesi RG, Mallegni F, Gely B, Vacca E, Morales MR, Straus LG, Neugebauer-Maresch C, Teschler-Nicola M, Constantin S, Moldovan OT, Benazzi S, Peresani 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 MA, Cupillard C, Bocherens H, Conard NJ, Harvati K, Moiseyev V, Drucker DG, Svoboda J, Richards MP, Caramelli D, Pinhasi R, Kelso J, Patterson N, Krause J, Paabo S, Reich D. The genetic history of Ice Age Europe. *Nature.* 2016;534:200–5.
- Fulton TL, Wagner SM, Fisher C, Shapiro B. Nuclear DNA from the extinct Passenger Pigeon (*Ectopistes migratorius*) confirms a single origin of New World pigeons. *Ann Anat.* 2012;194:52–7.
- Gaggiotti OE. Genetic threats to population persistence. *Ann Zool Fenn.* 2003;40:155–68.
- Galtier N, Nabholz B, Glemin S, Hurst GD. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. *Mol Ecol.* 2009;18:4541–50.
- Gibson J, Morton NE, Collins A. Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet.* 2006;15:789–95.
- Goossens B, Chikhi L, Jalil MF, Ancrenaz M, Lackman-Ancrenaz I, Mohamed M, Andau P, Bruford MW. Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. *Mol Ecol.* 2005;14:441–56.
- Gordon D, Huddleston J, Chaisson MJP, Hill CM, Kronenberg ZN, Munson KM, Malig M, Raja A, Fiddes I, Hillier LW, Dunn C, Baker C, Armstrong J, Diekhans M, Paten B, Shendure J, Wilson RK, Haussler D, Chin C-S, Eichler EE. Long-read sequence assembly of the gorilla genome. *Science.* 2016;352:aae0344.
- Grayson DK, Delpech F. Ungulates and the middle-to-upper Paleolithic transition at Grotte XVI (Dordogne, France). *J Archaeol Sci.* 2003;30:1633–48.
- Haddad NM, Brudvig LA, Clobert J, Davies KF, Gonzalez A, Holt RD, Lovejoy TE, Sexton JO, Austin MP, Collins CD, Cook WM, Damschen EI, Ewers RM, Foster BL, Jenkins CN, King AJ,

- Laurance WF, Levey DJ, Margules CR, Melbourne BA, Nicholls AO, Orrock JL, Song D-X, Townshend JR. Habitat fragmentation and its lasting impact on Earth's ecosystems. *Sci Adv*. 2015;1:e1500052.
- Hajdinjak M, Fu Q, Hubner A, Petr M, Mafessoni F, Grote S, Skoglund P, Narasimham V, Rougier H, Crevecoeur I, Semal P, Soressi M, Talamo S, Hublin JJ, Gusic I, Kucan Z, Rudan P, Golovanova LV, Doronichev VB, Posth C, Krause J, Korlevic P, Nagel S, Nickel B, Slatkin M, Patterson N, Reich D, Prufer K, Meyer M, Paabo S, Kelso J. Reconstructing the genetic history of late Neanderthals. *Nature*. 2018;555:652–6.
- Haubold B, Pfaffelhuber P, Lynch M. mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol Ecol*. 2010;19(Suppl 1):277–84.
- Hedrick PW. Purging inbreeding depression and the probability of extinction: full-sib mating. *Heredity*. 1994;73:363–72.
- Hedrick PW, Kalinowski ST. Inbreeding depression in conservation biology. *Annu Rev Ecol Syst*. 2000;31:139–62.
- Hedrick PW, Miller PS. Conservation genetics – techniques and fundamentals. *Ecol Appl*. 1992;2:30–46.
- Heintzman PD, Zazula GD, Cahill JA, Reyes AV, MacPhee RD, Shapiro B. Genomic data from extinct North American *Camelops* revise camel evolutionary history. *Mol Biol Evol*. 2015;32:2433–40.
- Heled J, Drummond AJ. Bayesian inference of population size history from multiple loci. *BMC Evol Biol*. 2008;8:289.
- Hofreiter M, Serre D, Rohland N, Rabeder G, Nagel D, Conard N, Munzel S, Paabo S. Lack of phylogeography in European mammals before the last glaciation. *Proc Natl Acad Sci U S A*. 2004;101:12963–8.
- Hofreiter M, Muenzel S, Conard NJ, Pollack J, Slatkin M, Weiss G, Paabo S. Sudden replacement of cave bear mitochondrial DNA in the Late Pleistocene. *Curr Biol*. 2007;17:R122–3.
- Hung CM, Shaner PJJ, Zink RM, Liu WC, Chu TC, Huang WS, Li SH. Drastic population fluctuations explain the rapid extinction of the passenger pigeon. *Proc Natl Acad Sci U S A*. 2014;111:10636–41.
- IUCN. The IUCN red list of threatened species 2016. 2016.
- Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337:816–21.
- John JS. SeqPrep. 2011. <https://github.com/jstjohn/SeqPrep>.
- Johnson RN, O'Meally D, Chen Z, Etherington GJ, Ho SYW, Nash WJ, Grueber CE, Cheng Y, Whittington CM, Dennison S, Peel E, Haerty W, O'Neill RJ, Colgan D, Russell TL, Alquezar-Planas DE, Attenbrow V, Bragg JG, Brandies PA, Chong AY-Y, Deakin JE, Di Palma F, Duda Z, Eldridge MDB, Ewart KM, Hogg CJ, Frankham GJ, Georges A, Gillett AK, Govendir M, Greenwood AD, Hayakawa T, Helgen KM, Hobbs M, Holleley CE, Heider TN, Jones EA, King A, Madden D, Graves JAM, Morris KM, Neaves LE, Patel HR, Polkinghorne A, Renfree MB, Robin C, Salinas R, Tsangaras K, Waters PD, Waters SA, Wright B, Wilkins MR, Timms P, Belov K. Adaptation and conservation insights from the koala genome. *Nat Genet*. 2018;50:1102–11.
- Kalinowski ST, Hedrick PW, Miller PS. Inbreeding depression in the Speke's gazelle captive breeding program. *Conserv Biol*. 2000;14:1375–84.
- Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW. Genomics advances the study of inbreeding depression in the wild. *Evol Appl*. 2016;9:1205–18.
- Kimura M. The neutral theory of molecular evolution. New York: Cambridge University Press; 1983.
- Kingman JFC. On the genealogy of large populations. *J Appl Probab*. 1982a;19A:27–43.
- Kingman JFC. The coalescent. *Stoch Process Appl*. 1982b;13:235–48.
- Kohn MH, Murphy WJ, Ostrander EA, Wayne RK. Genomics and conservation genetics. *Trends Ecol Evol*. 2006;21:629–37.

- Kurtén B. The cave bear story. New York: Columbia University Press; 1976.
- Lee BC, Kim MK, Jang G, Oh HJ, Yuda F, Kim HJ, Hossein MS, Kim JJ, Kang SK, Schatten G, Hwang WS. Dogs cloned from adult somatic cells. *Nature*. 2005;436:641.
- Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol*. 2012;10:e1001388.
- Leonard JA, Vila C, Fox-Dobbs K, Koch PL, Wayne RK, Van Valkenburgh B. Megafaunal extinctions and the disappearance of a specialized wolf ecomorph. *Curr Biol*. 2007;17:1146–50.
- Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio.GN]. 2013.
- Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475:493–6.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- Li Y, Hu Y, Bolund L, Wang J. State of the art de novo assembly of human genomes from massively parallel sequencing data. *Hum Genomics*. 2010;4:271–7.
- Lindahl T. Instability and decay of the primary structure of DNA. *Nature*. 1993;362:709–15.
- Lister AM, Stuart AJ. The impact of climate change on large mammal distribution and extinction: evidence from the last glacial/interglacial transition. *Compt Rendus Geosci*. 2008;340:615–20.
- Lorenzen ED, Nogues-Bravo D, Orlando L, Weinstock J, Binladen J, Marske KA, Ugan A, Borregaard MK, Gilbert MTP, Nielsen R, Ho SYW, Goebel T, Graf KE, Byers D, Stenderup JT, Rasmussen M, Campos PF, Leonard JA, Koepfli KP, Froese D, Zazula G, Stafford TW, Aaris-Sorensen K, Batra P, Haywood AM, Singarayer JS, Valdes PJ, Boeskorov G, Burns JA, Davydov SP, Haile J, Jenkins DL, Kosintsev P, Kuznetsova T, Lai XL, Martin LD, McDonald HG, Mol D, Meldgaard M, Munch K, Stephan E, Sablin M, Sommer RS, Sipko T, Scott E, Suchard MA, Tikhonov A, Willerslev R, Wayne RK, Cooper A, Hofreiter M, Sher A, Shapiro B, Rahbek C, Willerslev E. Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*. 2011;479:359–U195.
- Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*. 2011;21:936–9.
- Lynch M, Gabriel W. Mutational load and the survival of small populations. *Evolution*. 1990;44:1725–37.
- Lynch M, Conery J, Bürger R. Mutational meltdowns in sexual populations. *Evolution*. 1995a;49:1067–80.
- Lynch M, Conery J, Burger R. Mutation accumulation and the extinction of small populations. *Am Nat*. 1995b;146:489–518.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
- McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, Tenesa A, Macleod AK, Farrington SM, Rudan P, Hayward C, Vitart V, Rudan I, Wild SH, Dunlop MG, Wright AF, Campbell H, Wilson JF. Runs of homozygosity in European populations. *Am J Hum Genet*. 2008;83:359–72.
- Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andrés AM, Eichler EE, Slatkin M, Reich D, Kelso J, Pääbo S. A high-coverage genome sequence from an archaic Denisovan individual. *Science*. 2012;338:222–6.
- Millberg P, Tyrberg T. Naïve birds and noble savages – a review of man-caused prehistoric extinctions of island bird. *Ecography*. 1993;16:229–50.
- Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A, Tikhonov A, Raney B, Patterson N, Lindblad-Toh K, Lander ES, Knight JR, Irzyk GP,

- Fredrikson KM, Harkins TT, Sheridan S, Pringle T, Schuster SC. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*. 2008;456:387–90.
- Milot E, Weimerskirch H, Duchesne P, Bernatchez L. Surviving with low genetic diversity: the case of albatrosses. *Proc Biol Sci*. 2007;274:779–87.
- Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol*. 2008;25:1459–71.
- Mourier T, Ho SY, Gilbert MT, Willerslev E, Orlando L. Statistical guidelines for detecting past population shifts using ancient DNA. *Mol Biol Evol*. 2012;29:2241–51.
- Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill J, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, Gilbert TP, Heintzman PD, Letts B, McIntosh G, O’Connell BL, Peck M, Pipes M-L, Rice ES, Santos KM, Sohrweide AG, Vohr SH, Corbett-Detig RB, Green RE, Shapiro B. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Sci Rep*. 2017;358:951–4.
- Nelson DE, Angerbjorn A, Liden K, Turk I. Stable isotopes and the metabolism of the European cave bear. *Oecologia*. 1998;116:177–81.
- Nikolskiy PA, Sulerzhitsky LD, Pitulko VV. Last straw versus Blitzkrieg overkill: climate-driven changes in the Arctic Siberian mammoth population and the Late Pleistocene extinction problem. *Quat Sci Rev*. 2011;30:2309–28.
- Opgen-Rhein R, Fahrmeir L, Strimmer K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol Biol*. 2005;5:6.
- Orlando L, Cooper A. Using ancient DNA to understand evolutionary and ecological processes. *Annu Rev Ecol Evol Syst*. 2014;45(45):573–98.
- Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc Natl Acad Sci U S A*. 1989;86:1939–43.
- Pääbo S, Poinar H, Serre D, Jaenicke-Despres V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. Genetic analyses from ancient DNA. *Annu Rev Genet*. 2004;38:645–79.
- Pacher M, Stuart AJ. Extinction chronology and palaeobiology of the cave bear (*Ursus spelaeus*). *Boreas*. 2009;38:189–206.
- Palkopoulou E, Dalen L, Lister AM, Vartanyan S, Sablin M, Sher A, Edmark VN, Brandstrom MD, Germonpre M, Barnes I, Thomas JA. Holarctic genetic structure and range dynamics in the woolly mammoth. *Proc R Soc B Biol Sci*. 2013;280:20131910.
- Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Gotherstrom A, Reich D, Dalen L. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr Biol*. 2015;25:1395–400.
- Palkopoulou E, Baca M, Abramson NI, Sablin M, Socha P, Nadachowski A, Prost S, Germonpre M, Kosintsev P, Smirnov NG, Vartanyan S, Ponomarev D, Nystrom J, Nikolskiy P, Jass CN, Litvinov YN, Kalthoff DC, Grigoriev S, Fadeeva T, Douka A, Higham TFG, Ersmark E, Pitulko V, Pavlova E, Stewart JR, Weglenski P, Stankovic A, Dalen L. Synchronous genetic turnovers across Western Eurasia in Late Pleistocene collared lemmings. *Glob Chang Biol*. 2016;22:1710–21.
- Park SDE, Magee DA, McGettigan PA, Teasdale MD, Edwards CJ, Lohan AJ, Murphy A, Braud M, Donoghue MT, Liu Y, Chamberlain AT, Rue-Albrecht K, Schroeder S, Spillane C, Tai SS, Bradley DG, Sonstegard TS, Loftus BJ, MacHugh DE. Genome sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography and evolution of cattle. *Genome Biol*. 2015;16:234.
- Pečnerová P, Díez-del-Molino D, Vartanyan S, Dalén L. Changes in variation at the MHC class II DQA locus during the final demise of the woolly mammoth. *Sci Rep*. 2016;6:25274.
- Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K. EAGER: efficient ancient genome reconstruction. *Genome Biol*. 2016;17:60.
- Perry GLW, Wheeler AB, Wood JR, Wilmshurst JM. A high-precision chronology for the rapid extinction of New Zealand moa (*Aves*, *Dinornithiformes*). *Quat Sci Rev*. 2014;105:126–35.
- Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwangler A, Wissing C, Francken M, Malina M, Bolus M, Lari M, Gigli E, Capecchi G,

- Crevecoeur I, Beauval C, Flas D, Germonpre M, van der Plicht J, Cottiaux R, Gely B, Ronchitelli A, Wehrberger K, Grigorescu D, Svoboda J, Semal P, Caramelli D, Bocherens H, Harvati K, Conard NJ, Haak W, Powell A, Krause J. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. *Curr Biol*. 2016;26:827–33.
- Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prufer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Campubri C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andres AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T. Great ape genetic diversity and population history. *Nature*. 2013;499:471–5.
- Prowse TA, Johnson CN, Lacy RC, Bradshaw CJ, Pollak JP, Watts MJ, Brook BW. No need for disease: testing extinction hypotheses for the thylacine using multi-species metamodels. *J Anim Ecol*. 2013;82:355–64.
- Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. Computational challenges in the analysis of ancient DNA. *Genome Biol*. 2010;11:R47.
- Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, Li H, Mallick S, Dannemann M, Fu Q, Kircher M, Kuhlwilm M, Lachmann M, Meyer M, Ongyerth M, Siebauer M, Theunert C, Tandon A, Moorjani P, Pickrell J, Mullikin JC, Vohr SH, Green RE, Hellmann I, Johnson PL, Blanche H, Cann H, Kitzman JO, Shendure J, Eichler EE, Lein ES, Bakken TE, Golovanova LV, Doronichev VB, Shunkov MV, Derevianko AP, Viola B, Slatkin M, Reich D, Kelso J, Paabo S. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 2014;505:43–9.
- Purvis A, Gittleman JL, Cowlshaw G, Mace GM. Predicting extinction risk in declining species. *Proc Biol Sci*. 2000;267:1947–52.
- Pybus OG, Rambaut A, Harvey PH. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*. 2000;155:1429–37.
- Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Gronnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Ponten T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*. 2010;463:757–62.
- Richmond DJ, Sinding M-HS, Gilbert MTP. The potential and pitfalls of de-extinction. *Zool Scr*. 2016;45:22–36.
- Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, von Holdt BM, Marsden CD, Lohmueller KE, Wayne RK. Genomic flatlining in the endangered island fox. *Curr Biol*. 2016;26:1183–9.
- Rogers RL, Slatkin M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet*. 2017;13:e1006601.
- Salte F, Rodriguez-Rey M, Brook BW, Johnson CN, Turney CS, Alroy J, Cooper A, Beeton N, Bird MI, Fordham DA, Gillespie R, Herrando-Perez S, Jacobs Z, Miller GH, Nogues-Bravo D, Prideaux GJ, Roberts RG, Bradshaw CJ. Climate change not to blame for Late Quaternary megafauna extinctions in Australia. *Nat Commun*. 2016;7:10511.
- Sax DF, Gaines SD. Colloquium paper: species invasions and extinction: the future of native biodiversity on islands. *Proc Natl Acad Sci U S A*. 2008;105(Suppl 1):11490–7.

- Schorger AW. The passenger pigeon: its natural history and extinction. Whitefish: Literary Licensing, LLC; 1955.
- Schubert M, Ermini L, Der Sarkissian C, Jonsson H, Ginolhac A, Schaefer R, Martin MD, Fernandez R, Kircher M, McCue M, Willerslev E, Orlando L. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat Protoc.* 2014;9:1056–82.
- Shapiro B. Pathways to de-extinction: how close can we get to resurrection of an extinct species? *Funct Ecol.* 2017;31:996–1002.
- Shapiro B, Hofreiter M. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science.* 2014;343:1236573.
- Siddle HV, Marzec J, Cheng Y, Jones M, Belov K. MHC gene copy number variation in Tasmanian devils: implications for the spread of a contagious cancer. *Proc Biol Sci.* 2010;277:2001–6.
- Skoglund P, Malmström H, Raghavan M, Storå J, Hall P, Willerslev E, Gilbert TP, Götherström A, Jakobsson M. Origins and genetic legacy of neolithic farmers and hunter-gatherers in Europe. *Science.* 2012;336:466–9.
- Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Paabo S, Krause J, Jakobsson M. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc Natl Acad Sci U S A.* 2014;111:2229–34.
- Skoglund P, Ersmark E, Palkopoulou E, Dalen L. Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Curr Biol.* 2015;25:1515–9.
- Spurgin LG, Richardson DS. How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proc Biol Sci.* 2010;277:979–88.
- Stiller M, Baryshnikov G, Bocherens H, Grandal d'Anglade A, Hilpert B, Munzel SC, Pinhasi R, Rabeder G, Rosendahl W, Trinkaus E, Hofreiter M, Knapp M. Withering away – 25,000 years of genetic decline preceded cave bear extinction. *Mol Biol Evol.* 2010;27:975–8.
- Strimmer K, Pybus OG. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol Biol Evol.* 2001;18:2298–305.
- Tallmon DA, Luikart G, Waples RS. The alluring simplicity and complex reality of genetic rescue. *Trends Ecol Evol.* 2004;19:489–96.
- Tilman D, May RM, Lehman CL, Nowak MA. Habitat destruction and the extinction debt. *Nature.* 1994;371:65.
- Treangen TJ, Salzberg SL. Erratum: repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet.* 2012;13:146.
- Valdiosera CE, Garcia N, Anderung C, Dalen L, Cregut-Bonnoure E, Kahlke RD, Stiller M, Brandstrom M, Thomas MG, Arsuaga JL, Gotherstrom A, Barnes I. Staying out in the cold: glacial refugia and mitochondrial DNA phylogeography in ancient European brown bears. *Mol Ecol.* 2007;16:5140–8.
- Vartanyan S, Garutt VE, Sher AV. Holocene dwarf mammoths from Wrangel Island in the Siberian Arctic. *Nature.* 1993;362:337–40.
- Veltre DW, Yesner DR, Crossen KJ, Graham RW, Coltrain JB. Patterns of faunal extinction and paleoclimatic change from mid-Holocene mammoth and polar bear remains, Pribilof Islands, Alaska. *Quat Res.* 2017;70:40–50.
- Vucetich JA, Waite TA, Nunney L. Fluctuating population size and the ratio of effective to census population size. *Evolution.* 1997;51:2017–21.
- Wakayama S, Ohta H, Hikichi T, Mizutani E, Iwaki T, Kanagawa O, Wakayama T. Production of healthy cloned mice from bodies frozen at -20°C for 16 years. *Proc Natl Acad Sci U S A.* 2008;105:17318–22.
- Wandeler P, Hoeck PE, Keller LF. Back to the future: museum specimens in population genetics. *Trends Ecol Evol.* 2007;22:634–42.
- Wang J, Hill WG, Charlesworth D, Charlesworth B. Dynamics of inbreeding depression due to deleterious mutations in small populations: mutation parameters and inbreeding rate. *Genet Res.* 1999;74:165–78.

- Willi Y, van Buskirk J, Hoffmann AA. Limits to the adaptive potential of small populations. *Annu Rev Ecol Evol Syst.* 2006;37:433–58.
- Wright S. Evolution in Mendelian populations. *Genetics.* 1931;16:139–56.
- Wright S. Size of population and breeding structure in relation to evolution. *Science.* 1938;87:430–1.
- Wright S. Genetical structure of populations. *Nature.* 1950;166:247–9.
- Zimov SA, Zimov NS, Tikhonov AN, Chapin FS. Mammoth steppe: a high-productivity phenomenon. *Quat Sci Rev.* 2012;57:26–45.

Appendix III

An 'Aukward' Tale:

**A Genetic Approach to Discover the
Whereabouts of the Last Great Auks**

Article

An 'Aukward' Tale: A Genetic Approach to Discover the Whereabouts of the Last Great Auks

Jessica E. Thomas ^{1,2,*}, Gary R. Carvalho ^{1,†}, James Haile ^{2,†}, Michael D. Martin ³,
Jose A. Samaniego Castruita ², Jonas Niemann ², Mikkel-Holger S. Sinding ^{2,4},
Marcela Sandoval-Velasco ², Nicolas J. Rawlence ⁵, Errol Fuller ⁶, Jon Fjeldså ⁷,
Michael Hofreiter ⁸, John R. Stewart ⁹, M. Thomas P. Gilbert ^{2,3,†} and Michael Knapp ^{10,†}

¹ Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK; g.r.carvalho@bangor.ac.uk

² Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5–7, 1350 Copenhagen K, Denmark; drjameshaile@gmail.com (J.H.); jose.samaniego@snm.ku.dk (J.A.S.C.); j.niemann@snm.ku.dk (J.N.); mikkel.sinding@snm.ku.dk (M.-H.S.S.); marcela.velasco@snm.ku.dk (M.S.-V.); mtpgilbert@gmail.com (M.T.P.G.)

³ Department of Natural History, Norwegian University of Science and Technology, University Museum, NO-7491 Trondheim, Norway; mike.martin@ntnu.no

⁴ Natural History Museum, University of Oslo, P.O. Box 1172 Blindern, N-0318 Oslo, Norway

⁵ Otago Palaeogenetics Laboratory, Department of Zoology, University of Otago, Dunedin 9054, New Zealand; nic.rawlence@otago.ac.nz

⁶ 65 Springfield Road, Southborough, Tunbridge Wells TN4 0RD, Kent, UK; errolfuller123@btinternet.com

⁷ Center for Macroecology, Evolution and Climate, the Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark; jfjeldsaa@snm.ku.dk

⁸ Department of Mathematics and Natural Sciences, Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany; michi@palaeo.eu

⁹ Faculty of Science and Technology, Bournemouth University, Dorset BH12 5BB, UK; jstewart@bournemouth.ac.uk

¹⁰ Department of Anatomy, University of Otago, 270 Great King Street, Dunedin 9016, New Zealand; michael.knapp@otago.ac.nz

* Correspondence: bsp20a@bangor.ac.uk

† These authors contributed equally to this work.

Academic Editor: J. Peter W. Young

Received: 30 May 2017; Accepted: 9 June 2017; Published: 15 June 2017

Abstract: One hundred and seventy-three years ago, the last two Great Auks, *Pinguinus impennis*, ever reliably seen were killed. Their internal organs can be found in the collections of the Natural History Museum of Denmark, but the location of their skins has remained a mystery. In 1999, Great Auk expert Errol Fuller proposed a list of five potential candidate skins in museums around the world. Here we take a palaeogenomic approach to test which—if any—of Fuller's candidate skins likely belong to either of the two birds. Using mitochondrial genomes from the five candidate birds (housed in museums in Bremen, Brussels, Kiel, Los Angeles, and Oldenburg) and the organs of the last two known individuals, we partially solve the mystery that has been on Great Auk scholars' minds for generations and make new suggestions as to the whereabouts of the still-missing skin from these two birds.

Keywords: ancient DNA; extinct birds; mitochondrial genome; museum specimens; palaeogenomics

1. Introduction

Over the past three decades, the field of ancient DNA (aDNA) has grown considerably, from sequencing a small section of mitochondrial DNA from the Quagga, an extinct form of the plains zebra [1], to whole genome sequencing from samples up to 735,000 years old [2]. Ancient DNA has been used to answer and address a diverse range of ecological and evolutionary questions, providing insight into countless species' pasts, including our own. However, aDNA can also be a useful tool for museums, specifically for species identification and, under suitable circumstances for reconstructing the history of specimens where museum records are insufficient. This study traces the whereabouts of the skins from the last two documented Great Auks using a palaeogenomic approach.

The Great Auk (Figure 1), *Pinguinus impennis*, Bonnaterre (1790) (traditionally *Alca impennis*, Linnaeus, 1758), has been described as “perhaps the most curious of all vanished birds” [3]. It was a bird whose life and ultimate extinction has generated ongoing interest, with several scholars dedicating their lives to Great Auk research [3–7]. Even now, 173 years after the death of the last two recorded captured individuals, there are still many unanswered questions concerning aspect of its life-history, evolution, and extinction. One such mystery that surrounds the Great Auk is the whereabouts of the skins from the last documented pair. In order to be able to correlate the phenotype of the last birds with genomic information obtained from the well-preserved organs, and in view of the active role that researchers and research institutions played in pushing the Great Auk towards extinction, it is of relevance to be able to trace these skins.



Figure 1. A mounted Great Auk skin, The Brussels Auk (RBINS 5355) (MK135), from the collections at Royal Belgian Institute of Natural Sciences (Credit Thierry Hubin (RBINS)).

Once found in great numbers across the North Atlantic (Figure 2), this flightless bird was heavily hunted for its meat, oil, and feathers. By the start of the 19th century, populations in the North-West Atlantic had been decimated. The last few remaining birds were breeding on the skerries off the south-west coast of Iceland, but with their scarcity increasing, Great Auks were then also sought after as a desirable item for both private and institutional collections [3,5,8–10].

From 1830 to 1841, several trips were taken to Eldey Island (Figure 2) where Great Auks were caught, killed, and sold for exhibitions. Following a three-year period of no recorded captures of Great

Auks, Carl Siemsen commissioned an expedition to Eldey to search for any remaining birds. Between 2 and 5 June 1844, the expedition reached Eldey Island where two Great Auks were observed amongst smaller birds inhabiting the island. Both Auks were killed and their broken egg discarded. The birds, though, were never to reach Siemsen. The expedition leader sold them to Christian Hansen, who then sold them to the apothecary Möller, in Reykjavik, Iceland. Möller skinned the birds and sent them, as well as their preserved body parts, to Denmark [3,6,7].

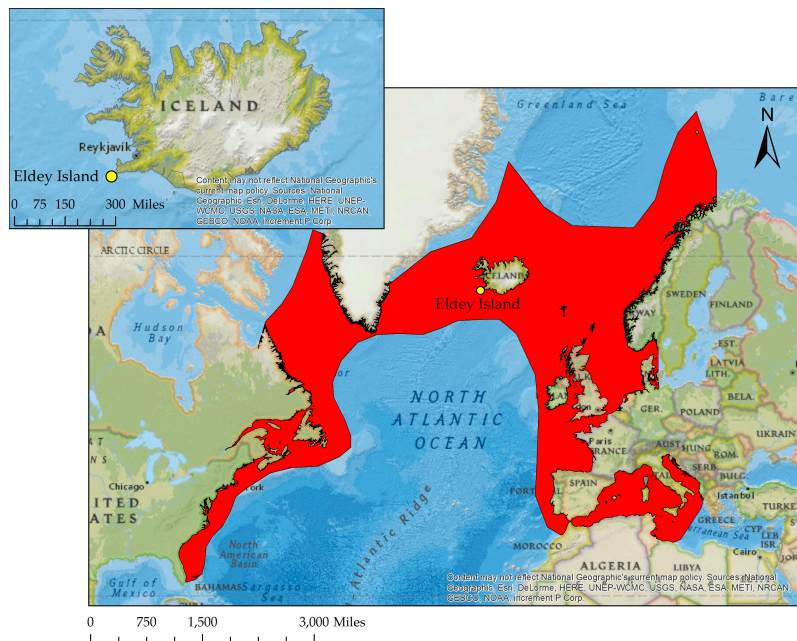


Figure 2. The Great Auk's breeding range across the North Atlantic, as indicated by the red area and the location of Eldey Island (yellow dot) off the south-west coast of Iceland, the site where the last documented Great Auks were killed. Maps were created using spatial data provided by BirdLife International/IUCN [11] with the National Geographic basemap in ArcGIS (ESRI, Redlands, CA, USA) [12].

The internal organs of these two birds now reside in the Natural History Museum of Denmark. However, the location of the skins of those individuals remains a mystery, despite considerable effort of notable Great Auk scholars to solve it.

Fuller [3] describes in detail the known history of the 80 or so specimens that are still in existence in collections today and concludes: *"Somehow, amid all the frantic Garefowl [another name for Great Auk] research of the nineteenth century, they [the skins] were lost track of. Several of the surviving stuffed specimens, notably those in Kiel, Bremen and Oldenburg were tentatively identified with them. The most likely candidates, however, are the birds now in Los Angeles and in Brussels"* [3] (p. 85).

Our study compares complete mitochondrial genome (mitogenome) sequences from the five candidate skins (those housed in Bremen, Brussels, Kiel, Los Angeles, and Oldenburg) to the internal organs of the last documented captured Great Auks (stored in Copenhagen) to test which—if any—of Fuller's candidate skins likely belong to one of the last two individuals.

2. Materials and Methods

2.1. Sample Information

Specimens from the candidate list proposed by Fuller [3] and the organs from the two 1844 Eldey Island individuals, were sampled using sterile equipment and the appropriate method for sample type, which caused minimal physical damage to the specimen (Table 1).

Table 1. Sample information. Lab ID number used during laboratory and analysis process. Mount name and description given by Fuller and its number in various published lists of Great Auk mounts [3]. Origin and date information as noted by Fuller [3]. Institution information relating to the present location of specimen and the curator/sample collector name.

| Lab ID | Bird Name, Number & Description | Origin & Date | Institution | Curator/Collector | Institution Number | Sample Type/Sampling Method |
|---------------|---|--|--|---------------------------|--------------------|--|
| MK131 | Last Great Auk 1 Oesophagus (male) | Eidey Island, Iceland. Date: June 1844 | Natural History Museum of Denmark Copenhagen, Denmark | J. Fieldså/ J. Thomas | NHMID 153069 | Oesophagus. Tissue cut from end of oesophagus. |
| MK132 | Last Great Auk 2 Oesophagus (female) | Eidey Island, Iceland. Date: June 1844 | Natural History Museum of Denmark Copenhagen, Denmark | J. Fieldså/ J. Thomas | NHMID 153070 | Oesophagus. Tissue cut from end of oesophagus. |
| MK133 | The Oldenburg Auk Fuller: Bird no. 47, Griever: no. 57, Hahn: no. 77 Adult in summer plumage | Iceland. Probably Eidey. Date: Unknown | Landesmuseum Natur und Mensch Oldenburg, Germany | C. Barilaro | AVE 8086 | Body tissue. Tissue cut from body of bird under wing. |
| MK134 | The Bremen Auk Fuller: Bird no. 36, Griever: no. 10, Hahn: no. 71 Adult in summer plumage | Unknown. Probably Eidey. Date: Unknown | Übersee-Museum Bremen Germany | M. Stiller | RKNr. 2392 | Toepad tissue. Tissue cut from feet. |
| MK135 | The Brussels Auk Fuller: Bird no. 3, Griever: no. 15, Hahn: no. 6 Adult in summer plumage | Probably Eidey Date: Unknown perhaps June, 1844 | Institut Royal des Sciences Naturelles de Belgique. Brussels, Belgium | G. Lenglet | RBINS 5355 | Toepad tissue. Tissue cut from feet |
| MK136 | Dawson Rowley's Los Angeles Auk Fuller: Bird no. 73, Griever no. 13, Hahn: no. 5 Adult in summer plumage, said to be female | Iceland. Probably Eidey. Date: Unknown perhaps June, 1844 | Natural History Museum of Los Angeles County, USA | K. Garrett | LACM 76476 | Feather. Feathers plucked from body of bird. |
| MK138 | The Schleswig-Holstein Auk Fuller: Bird no. 42, Griever: no. 31, Hahn: no. 74 Adult in summer plumage | Unknown Date: Unknown | Zoologisches Museum der Christian-Albrechts Universität zu Kiel, Germany | D. Brandis/ L. Kosotta | cat. No. A0585 | Toepad tissue. Tissue cut from feet. |
| LastGA2_Heart | Last Great Auk 2 Heart (female) | Eidey Island, Iceland. Date: June 1844 | Natural History Museum of Denmark Copenhagen, Denmark | J. Fieldså/ J. Halle | NHMID 153070 | Heart. Tissue cut from aorta. |

2.2. DNA Extraction

All lab work prior to polymerase chain reaction (PCR) amplification was carried out in designated aDNA laboratories that adhere to strict aDNA protocols [13]. For each DNA extraction and library build, negative controls were used to check for contamination by exogenous DNA. All post-PCR work on amplified DNA was carried out in separate laboratory facilities.

Genomic DNA was extracted from the oesophagus (Figure 3a), skin (Figure 3b), toepad tissue (Figure 3c), and feathers using a modified version of Dabney et al. [14] in which the initial digestion was carried out following the protocol by Gilbert et al. [15]. This digestion buffer is better suited to extraction from these tissues types than the Dabney et al. [14] digestion buffer, which was optimised for DNA extraction from bone. Subsequent DNA purification and elution was conducted following the approach described by Dabney et al. [14]. Genomic DNA was extracted from the heart tissue (Figure 3d) using the protocol by Campos et al. [16].



Figure 3. (a) Jars containing the oesophagus from the last two individuals killed on Eldey Island. The oesophagus from the larger jar represents that of the individual labelled male (NHMD153069) (MK131). The smaller jar contains the oesophagus from the female bird (NHMD153070) (MK132) (credit: J. Thomas). (b) Sampling of The Oldenburg Auk (AVE 8086) (MK133) to remove a section of body tissue for DNA extraction (credit: C. Barilaro, Landesmuseum Natur und Mensch Oldenburg). (c) Sampling the toe pad of The Bremen Auk (RKNr. 2392) (MK134) to remove tissue sample (credit M. Stiller, Übersee-Museum Bremen). (d) The hearts from the last two documented individuals. The heart from the female individual has been sampled for this study (top) (NHMD153070) (LastGA2_Heart) (credit Natural History Museum of Denmark).

2.3. Data Generation

Single stranded libraries were constructed for all samples, except LastGA2_Heart, following Gansauge & Meyer [17], with modifications as described by Bennett et al. [18], as this allowed for targeting of the smallest fragments of DNA, typical of highly degraded specimens. For LastGA2_Heart, the protocol described by Meyer & Kircher [19] was used. Enrichment for complete mitogenomes was performed using MYcroarray MYbaits, following the manufacturer's manual v2.3.1 [20] on all samples except MK138 and LastGA2_Heart. Samples were sequenced on Illumina platforms (HiSeq and MiSeq) by New Zealand Genomics Limited, Otago Branch, or the Danish National High-Throughput DNA Sequencing Centre.

2.4. Read Processing

Processing of raw sequence data was facilitated by the PALEOMIX v1.2.5 pipeline [21], which performs adapter trimming, read mapping to a reference genome, and quality-based filtering. Low-quality bases and adapter sequences were trimmed from the 3' ends of DNA reads with the software AdapterRemoval v2.1.7 [22,23] using a mismatch rate of 0.333 (command-line option—mm 3). Paired end reads overlapping by at least 11 base pairs (bp) were collapsed into a single read with re-calibrated base quality scores. Trimmed reads shorter than 25 bp were discarded.

Mapping to the Great Auk reference mitogenome (GenBank: KU158188.1) [24] was performed with Burrows–Wheeler Aligner (BWA) v0.5.10 [25] with seeding deactivated and otherwise default settings. PCR duplicates were removed with the MarkDuplicates function within Picard v1.82 [26] and the rmdup function within the software SAMtools [27]. Collapsed reads were filtered using a script included with PALEOMIX. Reads with mapping quality (MAPQ) scores <20 were removed from further analysis. Local realignment of reads misaligned to the reference mitogenome was performed with the RealignerTargetCreator and IndelRealigner tools included in the software Genome Analysis Toolkit (GATK) v3.6.0 [28]. The pipeline also utilised MapDamage2 [29] to recalibrate base qualities of aligned sequence reads in each sequencing library in order to remove the residual aDNA damage patterns. The UnifiedGenotyper algorithm within GATK v3.6.0 was used to determine haploid genotypes within individual samples.

A relaxed and strict filtering system was used to create consensus sequences and alignments from the processed data. In the first stage of filtering, both systems used VCFtools [30] to filter genotypes from the final alignment when their genotype quality scores were less than 30. For the relaxed alignment, the per-individual read depth was set to only include bases with a minimum of 3-fold coverage. Bases called for the consensus sequence had to be present at a frequency higher than 33%. To be included in the final alignment, no more than 33% of bases could be missing from the consensus sequence of an individual.

For the strict settings, the per-individual read depth was set to only include bases with at least 10-fold coverage. Geneious v-10.1.3 [31] was used to filter bases so that the majority base was present in more than 90% of reads. For an individual to be included in the final alignment, no more than 20% of sites could be missing from the individual's consensus sequence.

A custom script was used to convert the filtered Variant Call Format (VCF) file into a multiple sequence alignment in FASTA format.

Following read processing, the data was aligned using Seaview v4.0 [32] with the algorithm *Muscle -maxiters2 -diags*. The alignment was manually checked for errors using BioEdit v7.2.5 [33], and Tablet v-1.16.09.06 [34] was used to view the rescaled Binary Alignment Map (BAM) file for each sample.

MEGA v-7.0.21 [35] was used to generate a pairwise distance table for all sequenced individuals. Phylogenetic relationships between the individuals were reconstructed and visualized using a maximum-likelihood approach as implemented in MEGA v-7.0.21 [35]. jModelTest v-2.1.10 [36,37] was used to determine the most suitable nucleotide substitution model, which was a Hasegawa–Kishino–Yano (HKY) [38] model. Initial trees for the heuristic search were obtained by

applying Neighbour-Joining methods to a matrix of pairwise distances estimated using the maximum composite likelihood approach. Branch lengths are measured in number of substitutions per site. All positions containing gaps and missing data were removed. Phylogenies were reconstructed from 500 bootstrap pseudoreplicates to evaluate branch support.

3. Results

Mitogenome sequence data was obtained from all candidate specimens as well as from the two oesophagi of the last Great Auks. Unique coverage of the mitogenomes for these samples ranged from $6.2\times$ to $288.6\times$ (Table 2). As DNA extracted from the oesophagus of the female last Great Auk (MK132) yielded only a low coverage, poor quality mitogenome assembly, DNA from the heart of the same individual was also sequenced. This yielded a high coverage ($430\times$) mitogenome, which was used in all further analyses.

Table 2. Read processing results for all samples.

| Sample | GenBank Accession Number | Number of Reads | Number of Unique Reads Mapping to Reference Mitogenome | Estimated Coverage from Unique Hits | Relaxed Settings Sequence Length (bp ¹) | Strict Settings Sequence Length (bp) |
|---------------|--------------------------|-------------------------------------|--|-------------------------------------|---|--------------------------------------|
| MK131 | MF188883 | 300754 (read pairs) | 30,297 | 74.40 | 16,001 | 15,067 |
| MK132 | NA | 550631 (read pairs) | 2366 | 6.23 | 13,267 | 3312 |
| MK133 | MF188884 | 429392 (read pairs) | 8750 | 23.04 | 16,251 | 14,240 |
| MK134 | MF188885 | 343766 (read pairs) | 86,325 | 288.62 | 16,607 | 16,526 |
| MK135 | MF188886 | 579992 (read pairs) | 27,767 | 88.90 | 16,554 | 16,356 |
| MK136 | MF188887 | 563635 (read pairs) | 24,401 | 67.83 | 16,330 | 15,833 |
| MK138 | MF188888 | 10796460 (SE ² reads) | 2799 | 9.76 | 16509 | 7866 |
| LastGA2_Heart | MF188889 | 957970612 (SE reads) | 121,886 | 430.09 | 16,698 | 16,649 |

¹ Base pairs (bp); ² Single End (SE).

With the sequence data from the heart of the female last Great Auk (LastGA2_Heart), the alignment of all sequences assembled under the relaxed rules had a length of 15,790 bp after sites not covered by all consensus sequences were removed. For the strict alignment, MK138 did not meet criteria set by the strict filtering settings as more than 20% sites were missing. With this individual removed, we obtained a strict alignment length of 13,475 bp.

The pairwise distance matrix (Table 3) shows that the consensus sequence obtained from sample MK131, the oesophagus of the male, is identical to the consensus sequence obtained from MK135, The Brussels Auk. No other consensus sequences match. LastGA2_Heart, the female last Great Auk, groups with MK136 and MK134 in the maximum likelihood phylogeny (Figure 4), but there are 18 and 20 well-supported differences between the consensus sequences, respectively. Analysis presented here was generated using data from the relaxed filtering settings, but results were consistent with data from the strict filtering system. Thus, only the male last Great Auk has a corresponding DNA match among the candidate skin samples identified by Fuller [3].

Table 3. Pairwise distance matrix. Estimates of evolutionary divergence between sequences generated using the relaxed settings. The number of base differences per sequence from between sequences are shown. All positions containing gaps and missing data were removed, leaving a total of 15,790 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 [35].

| | MK131 | MK133 | MK134 | MK135 | MK136 | MK138 | LastGA2_Heart |
|-----------------|-------|-------|-------|-------|-------|-------|---------------|
| MK131_LastGA1 | | | | | | | |
| MK133_Oldenburg | 17 | | | | | | |
| MK134_Bremen | 18 | 23 | | | | | |
| MK135_Brussels | 0 | 17 | 18 | | | | |
| MK136_LA | 16 | 23 | 20 | 16 | | | |
| MK138_Kiel | 14 | 11 | 20 | 14 | 20 | | |
| LastGA2_Heart | 16 | 23 | 20 | 16 | 18 | 20 | |

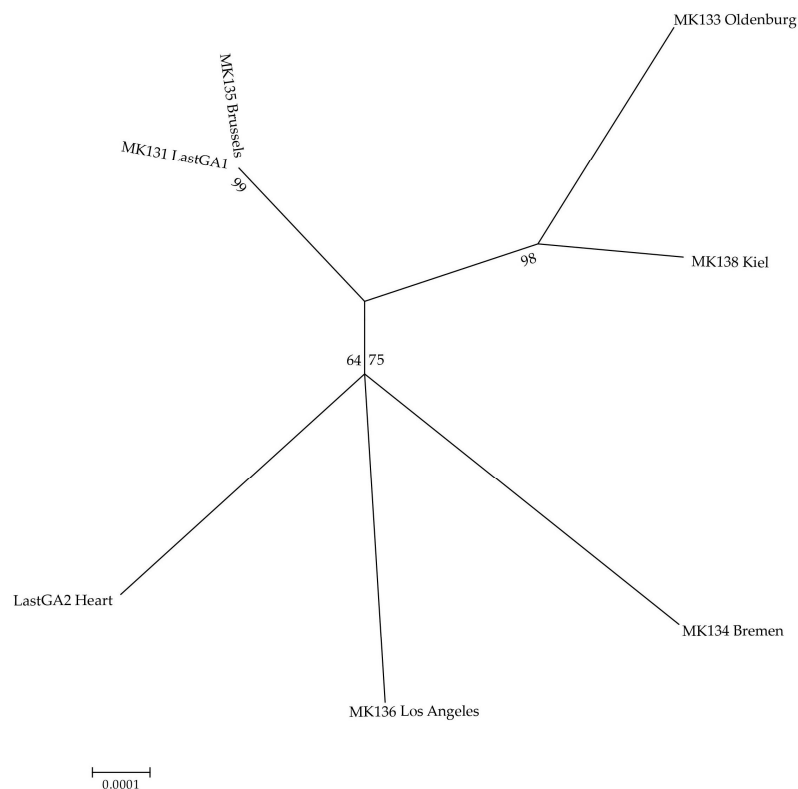


Figure 4. Maximum likelihood reconstruction of phylogenetic relationships between individuals, under the relaxed filtering settings. Branch labels are bootstrap support values for the respective sample. Evolutionary analyses were conducted in MEGA7 [35].

4. Discussion

The genetic analyses presented here help to partially resolve the mystery of the missing skins of the last two Great Auks. They provide evidence of matching mitochondrial genomes for the internal organs of the last male Great Auk held at the Natural History Museum of Denmark in Copenhagen and the Great Auk skin held at the Royal Belgian Institute of Natural Sciences, Brussels (Figure 1). Mitochondrial DNA cannot always be unambiguously used in identification of individuals. However, in a broader study of forty one Great Auk mitogenomes from across their range, Thomas et al. (in prep) [39], found that mitochondrial diversity in Great Auks remained high right up to their demise, with no other individuals found to have the same mitochondrial haplotype. Together with the information from the historical record, the match between the internal organs and The Brussels Auk therefore appears to be more than just a coincidence.

There are around 80 known mounted Great Auk skins in museums worldwide. However, the majority can be ruled out of any speculation that they belonged to the last pair due to their history (for example, if they were collected before 1844). Those tested in the current study were placed on the candidate specimen list due to several factors that led Fuller, as well as other experts like the University of Copenhagen Professor Japetus Steenstrup (dubbed 'Father of Garefowl History' by Grieve, 1885), and Grieve [4], to suspect that they originated from the 1844 Eldey pair. Details such as when and where they were acquired, from whom (i.e., the dealer), and suggestions by renowned Great Auk scholars made the birds in Bremen, Brussels, Kiel, Los Angeles, and Oldenburg the top candidates [3].

In the museum industry, accurate records and archiving are obviously of high priority, with labels and registers providing vital information about the specimens [40–42]; it therefore seems unexpected that the two bird skins could have been "lost". However, at the time, their significance as the final remnants of the species was not recognised. The story of the ending of these individuals lives is well documented due to the efforts of English naturalist John Wolley and Cambridge University Professor Alfred Newton, who travelled to Iceland in the late 1850s and spoke directly with those who were part of the 1844 Eldey Island voyage (details from Wolley's notebook 'Garefowl books' published in Newton, 1861 [7]). What happened once the skins and their organs reached Denmark, on the other hand, is poorly recorded and remains speculative [3].

In the archives of Cambridge University are the fragments of information that Newton learned of the birds. On notes dated 1861, it was recorded that Professor Reinhardt of the Royal Museum (Copenhagen) believed the skins and their organs had been purchased for the museum by Professor Eschricht of the University of Copenhagen. He is said to have taken the skins to the Congress of German Naturalists in Bremen in the autumn of 1844 [3].

The connection with the skins and the Congress in Bremen could be what led Steenstrup to inform Grieve of his suspicions that the specimen at the museum in Bremen (MK134) was indeed one of the last birds [4]. Yet, this bird was bought by the museum at the time of the Congress from the Hamburg dealer Salmin, not Eschricht. Therefore, while the possibility may be there for Salmin to have first had the bird from Eschricht and then sold it on, it is also likely that it was a bird he had in his stocks prior to 1844 [3]. This study shows The Bremen Auk is not a match with either of the organs from the last pair, suggesting that it did indeed come from an earlier raid of Eldey.

The specimen in Kiel, the Schleswig–Holstein Auk (MK138), was purchased in 1844. With such a suggestive purchase date it is a contender in the mystery [3]. Professor Steenstrup was quoted by Grieve as saying, "*If really purchased in 1844, it might perhaps be the second of these two Garefowls got in 1844, but traditionally I never heard that mentioned*" [4] (Grieve Appendix p. 13 [4]). Our study shows this specimen was not a match, so Steenstrup was correct in his belief.

With regard to The Oldenburg Auk (MK133), this specimen was once regarded by nineteenth century scholars as belonging to one of the last birds. However, the records for this bird shows it was obtained prior to 1844 and is therefore ruled out [3]. It was tested in this study due to the suggestions of these early researchers but was not a match.

The history of The Brussels Auk (MK135) and Dawson Rowley's Los Angeles Auk (MK136) can be traced back to 1845 when they were said to be in the hands of a well-known, and well connected, Great Auk dealer, Israel of Copenhagen. Israel is known to have had excellent links with Iceland and spent his winters in Copenhagen and his summers in Amsterdam [3]. Fuller suggests that perhaps Israel, if he did not receive them direct from Iceland, purchased the birds in Bremen from Eschricht. The birds have a detailed history, passing through the hands of several dealers. From Israel, they were bought by Lintz, a Hamburg merchant, and in 1845 were sold on to the Amsterdam branch of the dealer, Frank. In Newton's notes at Cambridge it was recorded that Frank believed the two skins he bought were from the last pair. The Brussels Auk was purchased in 1847 by Viscount Bernard Du Bus Ghisignies, director of the Brussels Museum [3]. The history of The Brussels Auk therefore strongly supports our positive match with MK131.

If the bird in Brussels, which came from Israel of Copenhagen, is from one of the last birds, then this would suggest that the second bird he had would also be from Eldey in 1844 and therefore be a positive match with the second set of organs. Israel's second bird has an even longer story than that of MK135, but it now resides in the Natural History Museum of Los Angeles County [3]. This specimen, Dawson Rowley's Los Angeles Auk (MK136), was tested, and the results showed it did not match LastGA2_Heart. With this negative result, we can only speculate which of the remaining untested birds could be identified as the second individual.

A possible scenario to explain the mismatch between Dawson Rowley's Los Angeles Auk (MK136) and the internal organs from the Natural History Museum of Denmark involves a mix up of skins. Dawson Rowley's Los Angeles Auk, was once one of two Great Auks owned by George Dawson Rowley. During the 1930s, they were passed to Captain Vivian Hewitt who owned two additional specimens. The four specimens are currently held in Cardiff, Birmingham, Los Angeles, and Cincinnati. At Hewitt's death, his collection had been put under the control of Spink and Son Ltd., a London dealer, who offered them for sale. While organising Hewitt's affairs, the four birds were mixed up. The identity of the birds now in Birmingham and Cardiff could be easily resolved, but those now in Los Angeles and Cincinnati are harder to determine. It is thought that their identities could be determined from annotated photographs taken in 1871 by George Dawson Rowley when they were in his possession [3]. However, we speculate that their identities were not correctly resolved and that perhaps the bird in Cincinnati was the original bird from Israel of Copenhagen. If this were the case, then it would explain why the Los Angeles bird fails to match with either of the last Great Auk organs held in Copenhagen.

In summary, we suggest that The Brussels Auk is the skin from the last male Great Auk killed on Eldey Island in June 1844. The skin of the female killed at the same time remains unaccounted for, but a common history with The Brussels Auk makes the skin currently held at Cincinnati Museum of Natural History and Science, a likely candidate. A re-evaluation of the historical records may reveal further candidate skins amongst those currently held in museums around the world.

5. Conclusions

Ancient DNA has been used to evaluate museum collections in the past, albeit usually for taxonomic identification of unidentified or misidentified accessions. Our study shows an alternative use of the technology. It demonstrates the utility of molecular tools and advanced sequencing to contribute to questions, which are not primarily biological or molecular but rather historical in nature. The unraveling of the mystery surrounding the whereabouts of the skins of the last two Great Auks represents a fascinating element in the story of extinction and human involvement in that process.

Acknowledgments: We are very grateful to the institutions that provided samples to this project and the curators/sample collectors within them (Jon Fjeldså at Natural History Museum of Denmark, Christina Barilao at Landesmuseum Natur und Mensch Oldenburg, Germany, Michael Stiller at Übersee-Museum Bremen, Germany, Georges Lenglet at Institut Royal des Sciences Naturelles de Belgique, Brussels, Belgium (RBINS), Kimball Garret at Los Angeles County Museum of Natural History, Los Angeles, USA and Dirk Brandis and Lina Rosotta at Zoologisches Museum der Christian-Albrechts-Universität zu Kiel, Germany). We thank Jan Bolding Kristensen for his help and advice with the sampling of the Great Auk organs and photograph of the Great Auk hearts. We also thank Thierry Hubin (RBINS), Michael Stiller (Kiel), and Christina Barilao for providing photographs for use in this publication. Funding was provided through NERC Ph.D. Studentship (NE/L501694/1), the Genetics Society-Hereditry Fieldwork Grant, and European Society for Evolutionary Biology–Godfrey Hewitt Mobility Award. MK is supported by a Rutherford Discovery Fellowship from the Royal Society of New Zealand. We thank members of the Molecular Ecology and Fisheries Genetics Laboratory at Bangor University, EvoGenomics & GeoGenetics at University of Copenhagen, and the Biological Anthropology group at the University of Otago, for guidance in the laboratory, analysis, and useful discussions. Sequencing was provided by either The Danish National High-Throughput DNA Sequencing Centre or New Zealand Genomics Limited.

Author Contributions: J.E.T. conceived the study, J.E.T., J.H., G.R.C., M.T.P.G., and M.K. designed the experiments; J.E.T., J.H., M-H.S.S., and M.S.-V. conducted the experiments; J.E.T., M.D.M., J.A.S.C., and J.N. analysed the data; N.J.R., M.H., and J.R.S. provided the initial framework for the study; all authors contributed to writing the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Higuchi, R.; Bowman, B.; Freiberger, M.; Ryder, O.A.; Wilson, A.C. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **1984**, *312*, 282–284. [CrossRef]
- Orlando, L.; Ginolhac, A.; Zhang, G.; Froese, D.; Albrechtsen, A.; Stiller, M.; Schubert, M.; Cappellini, E.; Petersen, B.; Moltke, I.; et al. Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. *Nature* **2013**, *499*, 74–78. [CrossRef]
- Fuller, E. *The Great Auk*; Errol Fuller: Kent, UK, 1999; 448p; ISBN 0-9533553-0-6.
- Grieve, S. *The Great Auk, or Garefowl. Its History, Archaeology and Remains*, Digitally Printed Version 2015 ed.; Cambridge University Press: Cambridge, UK, 1885; ISBN 978-1-108-08147-4.
- Bengtson, S.-A. Breeding ecology and extinction of the great auk (*Pinguinus impennis*): Anecdotal evidence and conjectures. *Auk* **1984**, *101*, 1–12.
- Gaskell, J. *Who Killed the Great Auk?* Oxford University Press: New York, NY, USA, 2000; 240p; ISBN 978-0-19856478-2.
- Newton, A. XIII.—Abstract of Mr. J. Wolley’s researches in Iceland respecting the gare-fowl or great auk (*Alca impennis*, linn.). *Ibis* **1861**, *3*, 374–399. [CrossRef]
- Meldgaard, M. The great auk, *Pinguinus impennis* (L.) in greenland. *Hist. Biol.* **1988**, *1*, 145–178. [CrossRef]
- Serjeantson, D. The great auk and the gannet: A prehistoric perspective on the extinction of the great auk. *Int. J. Osteoarchaeol.* **2001**, *11*, 43–55. [CrossRef]
- Montevocchi, W.A.; David, A. Kirk. *Great Auk (Pinguinus impennis) Birds of North America Online*; Rodewald, P.G., Ed.; Cornell Lab of Ornithology: Ithaca, NY, USA, 1996.
- International Union for Conservation of Nature, Birdlife International and Handbook of the Birds of the World (2016). *Pinguinus impennis*, The IUCN Red List of Threatened Species. Version 2016–3; 2016. Available online: <http://maps.iucnredlist.org/map.html?id=22694856> (accessed on 16 May 2017).
- ArcGIS. *Arcgis Desktop, Arcmap*; 10.5.0.6491; ESRI: Redlands, CA, USA, 2016.
- Knapp, M.; Clarke, A.C.; Horsburgh, K.A.; Matisoo-Smith, E.A. Setting the stage—Building and working in an ancient DNA laboratory. *Ann. Anat.* **2012**, *194*, 3–6. [CrossRef] [PubMed]
- Dabney, J.; Knapp, M.; Glocke, I.; Gansauge, M.-T.; Weihmann, A.; Nickel, B.; Valdiosera, C.; García, N.; Pääbo, S.; Arsuaga, J.-L.; et al. Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15758–15763. [CrossRef]
- Gilbert, M.T.P.; Tomsho, L.P.; Rendulic, S.; Packard, M.; Drautz, D.I.; Sher, A.; Tikhonov, A.; Dalén, L.; Kuznetsova, T.; Kosintsev, P.; et al. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **2007**, *317*, 1927–1930. [CrossRef] [PubMed]
- Campos, P.F.; Gilbert, T.M.P. DNA extraction from keratin and chitin. In *Ancient DNA: Methods and Protocols*; Shapiro, B., Hofreiter, M., Eds.; Humana Press: Totowa, NJ, USA, 2012; pp. 43–49; ISBN 978-1-61779-516-9.
- Gansauge, M.-T.; Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **2013**, *8*, 737–748. [CrossRef]
- Bennett, E.A.; Massilani, D.; Lizzo, G.; Daligault, J.; Geigl, E.M.; Grange, T. Library construction for ancient genomics: Single strand or double strand? *BioTechniques* **2014**, *56*, 289–290, 292–286, 298, passim. [CrossRef]
- Meyer, M.; Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, *2010*. [CrossRef]
- MYcroarray. *Mybaits Manual—Sequence Enrichment for Targeted Sequencing v2.3.1*; MYcroarray, 2014; Available online: <http://www.mycroarray.com/mybaits/manuals.html> (accessed on 13 June 2017).
- Schubert, M.; Ermini, L.; Sarkissian, C.D.; Jónsson, H.; Ginolhac, A.; Schaefer, R.; Martin, M.D.; Fernández, R.; Kircher, M.; McCue, M.; et al. Characterization of ancient and modern genomes by snp detection and phylogenomic and metagenomic analysis using paleomix. *Nat. Protoc.* **2014**, *9*, 1056–1082. [CrossRef]
- Lindgreen, S. Adapterremoval: Easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **2012**, *5*, 337. [CrossRef]
- Schubert, M.; Lindgreen, S.; Orlando, L. Adapterremoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **2016**, *9*, 88. [CrossRef]
- Anmarkrud, J.A.; Lifjeld, J.T. Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Mol. Ecol. Resour.* **2017**, *17*, 334–341. [CrossRef]

25. Li, H.; Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)]
26. BroadInstitute. Picard v1.82. Available online: <http://broadinstitute.github.io/picard/> (accessed on 13 June 2017).
27. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)]
28. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernysky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The genome analysis toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [[CrossRef](#)]
29. Jónsson, H.; Ginolhac, A.; Schubert, M.; Johnson, P.L.F.; Orlando, L. Mapdamage2.0: Fast approximate bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **2013**, *29*, 1682–1684. [[CrossRef](#)]
30. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and vcftools. *Bioinformatics* **2011**, *27*, 2156–2158. [[CrossRef](#)]
31. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C.; et al. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647–1649. [[CrossRef](#)]
32. Gouy, M.; Guindon, S.; Gascuel, O. Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **2010**, *27*, 221–224. [[CrossRef](#)]
33. Hall, T.A. Bioedit: A user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symp. Ser.* **1999**, *41*, 95–98.
34. Milne, I.; Stephen, G.; Bayer, M.; Cock, P.J.A.; Pritchard, L.; Cardle, L.; Shaw, P.D.; Marshall, D. Using tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **2013**, *14*, 193–202. [[CrossRef](#)]
35. Kumar, S.; Stecher, G.; Tamura, K. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [[CrossRef](#)]
36. Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. Jmodeltest 2: More models, new heuristics and high-performance computing. *Nat. Methods* **2012**, *9*, 772. [[CrossRef](#)]
37. Guindon, S.; Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **2003**, *52*, 696–704. [[CrossRef](#)]
38. Hasegawa, M.; Kishino, H.; Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **1985**, *22*, 160–174. [[CrossRef](#)]
39. Thomas, J.E.; Haile, J.; Martin, M.D.; Samaniego Castuita, J.A.; Niemann, J.; Sinding, .M.-H.S.; Sandoval-Velasco, M.; Soares, A.E.R.; Rawlence, N.J.; Fuller, E.; et al. The evolution and extinction of the Great Auk. (manuscript in preparation).
40. Boessenkool, S.; Star, B.; Scofield, R.P.; Seddon, P.J.; Waters, J.M. Lost in translation or deliberate falsification? Genetic analyses reveal erroneous museum data for historic penguin specimens. *Proc. R. Soc. B Biol. Sci.* **2010**, *277*, 1057–1064. [[CrossRef](#)]
41. Rawlence, N.J.; Kennedy, M.; Waters, J.M.; Scofield, R.P. Morphological and ancient DNA analyses reveal inaccurate labels on two of buller’s bird specimens. *J. R. Soc. N. Z.* **2014**, *44*, 163–169. [[CrossRef](#)]
42. Shepherd, L.D.; Tennyson, A.J.D.; Lambert, D.M. Using ancient DNA to enhance museum collections: A case study of rare kiwi (*Apteryx* spp.) specimens. *J. R. Soc. N. Z.* **2013**, *43*, 119–127. [[CrossRef](#)]



Appendix IV

Demographic reconstruction from ancient DNA supports rapid extinction of the great auk

Demographic reconstruction from ancient DNA supports rapid extinction of the great auk

Jessica E Thomas^{1,2†*}, Gary R Carvalho^{1†}, James Haile², Nicolas J Rawlence³, Michael D Martin⁴, Simon YW Ho⁵, Arnór Þ Sigfússon⁶, Vigfús A Jósefsson⁶, Morten Frederiksen⁷, Jannie F Linnebjerg⁷, Jose A Samaniego Castruita², Jonas Niemann², Mikkel-Holger S Sinding^{2,8}, Marcela Sandoval-Velasco², André ER Soares⁹, Robert Lacy¹⁰, Christina Barilaro¹¹, Juila Best^{12,13}, Dirk Brandis¹⁴, Chiara Cavallo¹⁵, Mikelo Elorza¹⁶, Kimball L Garrett¹⁷, Maaïke Groot¹⁸, Friederike Johansson¹⁹, Jan T Lifjeld²⁰, Göran Nilson¹⁹, Dale Serjeanston²¹, Paul Sweet²², Errol Fuller²³, Anne Karin Hufthammer²⁴, Morten Meldgaard²⁵, Jon Fjeldsá²⁶, Beth Shapiro⁹, Michael Hofreiter²⁷, John R Stewart^{28†}, M Thomas P Gilbert^{2,4†}, Michael Knapp^{29†*}

¹Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Bangor University, Bangor, United Kingdom; ²Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark; ³Otago Palaeogenetics Laboratory, Department of Zoology, University of Otago, Dunedin, New Zealand; ⁴Department of Natural History, University Museum, Norwegian University of Science and Technology, Trondheim, Norway; ⁵School of Life and Environmental Sciences, University of Sydney, Sydney, Australia; ⁶Verkís Consulting Engineers, Reykjavik, Iceland; ⁷Department of Bioscience, Aarhus University, Roskilde, Denmark; ⁸Greenland Institute of Natural Resources, Nuuk, Greenland; ⁹Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, United States; ¹⁰Department of Conservation Science, Chicago Zoological Society, Brookfield, United States; ¹¹Landesmuseum Natur und Mensch Oldenburg, Oldenburg, Germany; ¹²Department of Archaeology, Anthropology and Forensic Science, Faculty of Science and Technology, Bournemouth University, Poole, United Kingdom; ¹³School of History, Archaeology and Religion, Cardiff University, Cardiff, United Kingdom; ¹⁴Zoological Museum, University of Kiel, Kiel, Germany; ¹⁵Amsterdam Centre for Ancient Studies and Archaeology, University of Amsterdam, Amsterdam, Netherlands; ¹⁶Arqueología Prehistórica, Sociedad de Ciencias Aranzadi, San Sebastián, Spain; ¹⁷Natural History Museum of Los Angeles County, Los Angeles, United States; ¹⁸Institut für Prähistorische Archäologie, Freie Universität Berlin, Berlin, Germany; ¹⁹Gothenburg Museum of Natural History, Gothenburg, Sweden; ²⁰Natural History Museum, University of Oslo, Oslo, Norway; ²¹Humanities Archaeology, University of Southampton, Southampton, United Kingdom; ²²Department of Ornithology, American Museum of Natural History, New York, United States; ²³Independent researcher, Kent, United Kingdom; ²⁴Department of Natural History, University Museum of Bergen, Bergen, Norway; ²⁵University of Greenland, Nuuk, Greenland; ²⁶Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark; ²⁷Evolutionary Adaptive Genomics, Institute for Biochemistry and Biology, Department of Mathematics and Natural Sciences,

***For correspondence:**

drjethomas@hotmail.com (JET);
michael.knapp@otago.ac.nz (MK)

†These authors contributed
equally to this work

Present address: †Department
of Bioscience, College of
Science, Swansea University,
Swansea, United Kingdom

Competing interest: See
page 12

Funding: See page 12

Received: 08 April 2019

Accepted: 22 October 2019

Published: 26 November 2019

Reviewing editor: Christian
Rutz, University of St Andrews,
United Kingdom

© Copyright Thomas et al. This
article is distributed under the
terms of the [Creative Commons
Attribution License](#), which
permits unrestricted use and
redistribution provided that the
original author and source are
credited.

University of Potsdam, Potsdam, Germany; ²⁸Faculty of Science and Technology, Bournemouth University, Dorset, United Kingdom; ²⁹Department of Anatomy, University of Otago, Dunedin, New Zealand

Abstract The great auk was once abundant and distributed across the North Atlantic. It is now extinct, having been heavily exploited for its eggs, meat, and feathers. We investigated the impact of human hunting on its demise by integrating genetic data, GPS-based ocean current data, and analyses of population viability. We sequenced complete mitochondrial genomes of 41 individuals from across the species' geographic range and reconstructed population structure and population dynamics throughout the Holocene. Taken together, our data do not provide any evidence that great auks were at risk of extinction prior to the onset of intensive human hunting in the early 16th century. In addition, our population viability analyses reveal that even if the great auk had not been under threat by environmental change, human hunting alone could have been sufficient to cause its extinction. Our results emphasise the vulnerability of even abundant and widespread species to intense and localised exploitation.

Introduction

The great auk (*Pinguinus impennis*) was a large, flightless diving bird thought to have once numbered in the millions (Birkhead, 1993). A member of the family Alcidae in the order Charadriiformes, its closest extant relative is the razorbill (*Alca torda*) (Moum et al., 2002). The great auk was distributed around the North Atlantic and breeding colonies could be found along the east coast of North America, especially on the islands off Newfoundland (Figure 1). The species also bred on islands off Iceland and Scotland, and was found throughout Scandinavia (Norway, Denmark, and Sweden), with evidence of bone finds existing as far south as Florida and in to the Mediterranean (Fuller, 1999; Grieve, 1885).

The archaeological and historical records show a long history of humans hunting great auks. In prehistoric times, they were hunted for their meat and eggs by the Beothuk in North America (Fuller, 1999; Gaskell, 2000), the Inuit of Greenland (Meldgaard, 1988), Scandinavians (Hufthammer, 1982), Icelanders (Bengtson, 1984), in Britain (Best, 2013; Best and Mulville, 2016), Magdalenian hunter-gatherers in the Bay of Biscay (Laroulandie et al., 2016), and possibly even Neanderthals (Halliday, 1978). Around 1500 AD intensive hunting began by European seamen visiting the fishing grounds of Newfoundland (Bengtson, 1984; Fuller, 1999; Gaskell, 2000; Steenstrup, 1855). Towards the end of the 1700s, the development of commercial hunting for the feather trade intensified exploitation levels (Fuller, 1999; Gaskell, 2000; Kirkham and Montevecchi, 1982). As their rarity increased, great auk specimens and eggs became desirable for private and institutional collections. The last reliably recorded breeding pair were killed in June 1844 on Eldey Island, Iceland, to be added to a museum collection (Bengtson, 1984; Fuller, 1999; Gaskell, 2000; Grieve, 1885; Newton, 1861; Steenstrup, 1855; Thomas et al., 2017).

There are scattered records of great auks dating to later than 1844, including in 1848 near Vardø, Norway (Fuller, 1999; Newton, 1861), and 1852 in Newfoundland (Fuller, 1999; Grieve, 1885; Newton, 1861). BirdLife International/IUCN recognises the last sighting as 1852 (BirdLife International, 2016a). However, uncertainty remains about the reliability of all of these later sightings (Fuller, 1999; Grieve, 1885). There is little doubt that the extensive hunting pressure on the species contributed significantly to its demise. Nevertheless, despite the well documented history of exploitation since the 16th century, it is unclear whether hunting alone could have been responsible for the demise of the great auk, or whether the species was already in decline due to non-anthropogenic environmental changes (Bengtson, 1984; Birkhead, 1993; Fuller, 1999). For example, there is evidence of a decrease in great auk numbers on the eastern side of the North Atlantic, as reflected in a decline in bone finds in England, Scotland, and Scandinavia, which remains unexplained and could have been caused by hunting as well as environmental change (Bengtson, 1984; Best and Mulville, 2014; Grieve, 1885; Hufthammer, 1982; Serjeantson, 2001). To quote Bengtson (1984), 'In the absence of more detailed information about rate of decline of the

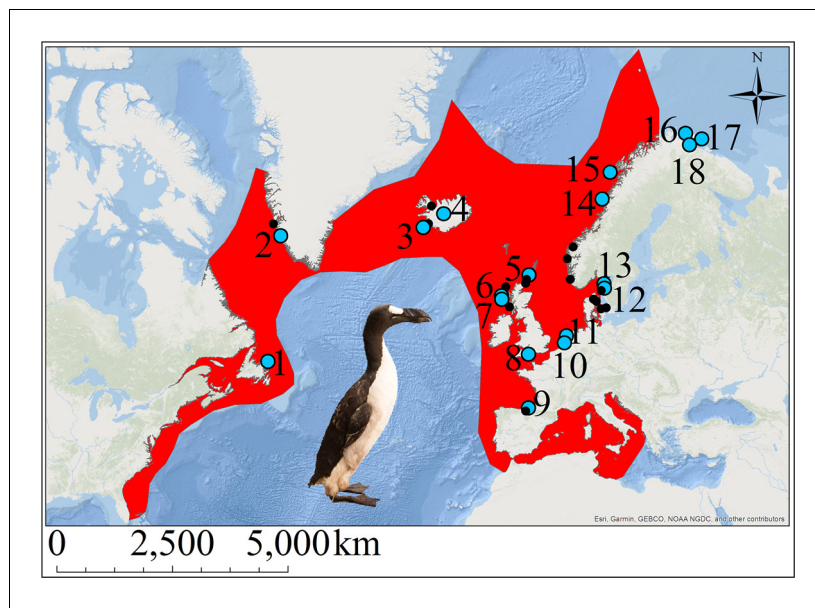


Figure 1. The great auk and its former distribution in the North Atlantic. Red shading indicates the geographic distribution of the great auk, as defined by BirdLife International/IUCN (*BirdLife International, 2016a*). Sites marked with blue dots represent samples used in our analyses. Black dots denote other sites from which material was obtained, but for which samples were not sequenced or for which sequences did not pass filtering settings. Numbers associated with blue dots correspond to the following sites: 1: Funk Island (n = 14), 2: Qeqertarsuaq (n = 1), 3: Eldey Island (n = 2), 4: Iceland (n = 5), 5: Tofts Ness (n = 2), 6: Bornais (n = 1), 7: Cladh Hallan (n = 1), 8: Portland (n = 1), 9: Santa Catalina (n = 2), 10: Schipluiden (n = 1), 11: Velsen (n = 1), 12: Sotenkanalen (n = 2), 13: Skalbänk Otterön (n = 2), 14: Kirkehlleren (n = 1), 15: Storbåthelleren (n = 1), 16: Iversfjord (n = 1), 17: Vardø (n = 2), and 18: Nyelv (n = 1).

bird populations, hunting pressure and environmental changes, we cannot separate the effects of hunting and that of climate change' (p10).

Reconstructing specific environmental influences on an extinct species can be difficult when there is limited knowledge of the species' biology. However, if the species had been at risk of extinction prior to the onset of intensive hunting in the 16th century, we would expect to see genetic signatures of population decline, including limited genetic diversity and pronounced population structure. In contrast, the lack of an observable loss in genetic diversity during the last few centuries prior to the extinction would be consistent with a rapid demographic decline at the end. At the same time, human hunting alone can only be considered a reasonable explanation for the extinction of the great auk, if population viability analyses show that extinction could have been caused by harvest rates that would have been realistic for the time and circumstances of the harvest.

Here, we examine the drivers of the extinction of the great auk by analysing whole mitochondrial genome (mitogenome) sequences from across its geographic range, population viability, and harvest rates. We combined these with data from GPS-equipped drifting capsules deployed in the North Atlantic, which allow us to suggest potential migration routes among breeding sites.

Results

Mitogenome sequence data

Using hybridisation capture combined with high-throughput sequencing, we generated short-read sequence data from 66 bone samples of great auk (See *Supplementary file 1a* for sample information). Following read processing and filtering, 35 samples passed the quality requirements (see Materials and methods) and were suitable for further analysis. In addition to the sequences generated from bones, we included six previously published mitogenome sequences from tissue or feather samples (*Thomas et al., 2017*) (*Supplementary file 1a*).

The combined data set comprised 41 complete mitogenomes, representing individuals from across the former range of the great auk and spanning the period 170–15,000 years before present (ybp). For samples in the final data set, the mean average read length of aligned bases to the reference great auk mitogenome (GenBank accession KU158188.1 [Anmarkrud and Lifjeld, 2017]) was 55.12 base pairs (bp), with a range of 41.21–86.95 bp. Unique mitogenome coverage of these samples ranged from $6.39 \times$ to $430.09 \times$, with average coverage of $72.5 \times$ (Supplementary file 1c). The final alignment length was 16,641 bp, including 9994 bp (after removal of gaps) that were shared across all 41 mitogenomes.

Genetic diversity and population structure

Haplotype diversity among the great auk mitogenomes was high, with only two individuals yielding identical haplotypes across the 9994 bp covered by all 41 mitogenomes. The two identical sequences differed in age, so that when divided into different age groups, each age group contained a unique set of haplotypes. No reduction of haplotype diversity could be identified in more recent samples (Figure 2).

We observed no structure in the distribution of haplotypes using any of our four approaches to reconstruct phylogeographic and temporal relationships among the samples: Bayesian analyses using BEAST (Appendix 1—figure 1 and Appendix 1—figure 2); maximum-likelihood phylogenetic

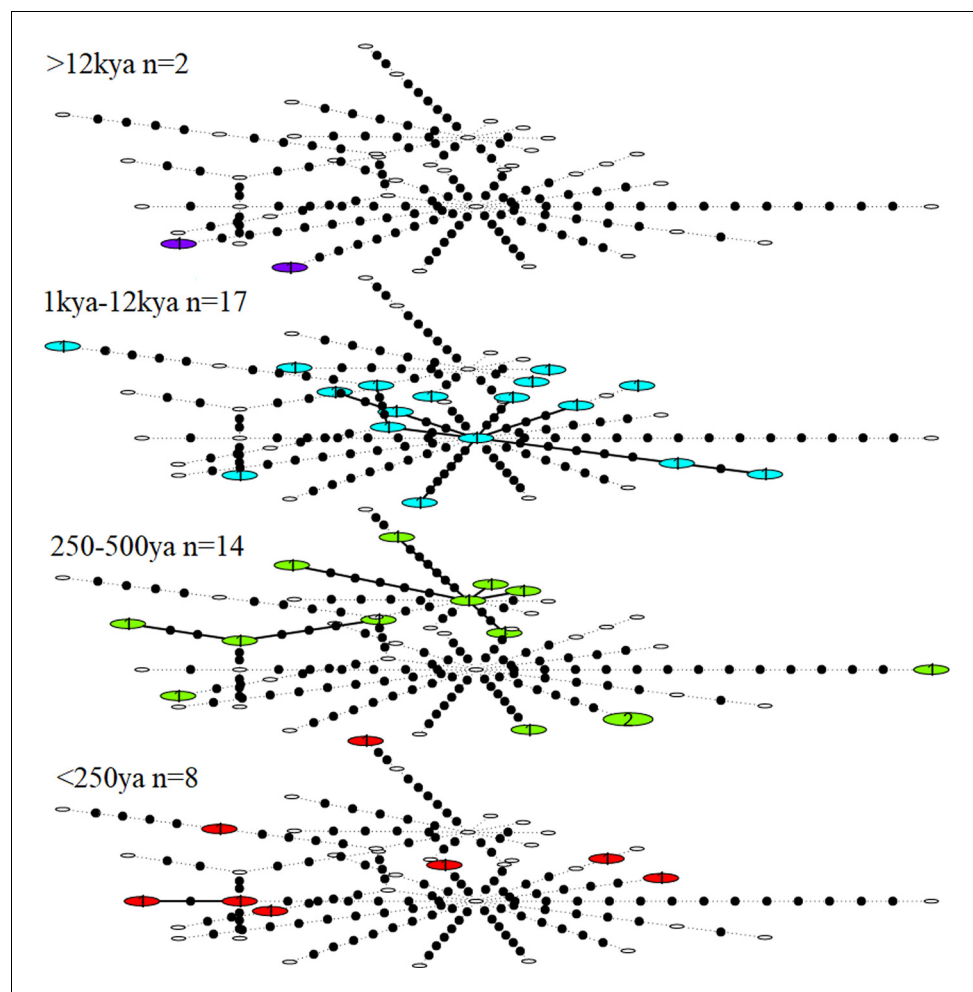


Figure 2. Statistical parsimony network showing haplotype diversity of great auk mitogenomes through time. In each age category observed haplotypes are shown in colour, absent haplotypes are shown as empty circles, and mutations between haplotypes are marked as black dots. All samples have been included in this figure.

analysis using RAxML; statistical parsimony network analysis using TempNet (**Figure 2**); and median-joining network analysis using PopART (**Figure 3**).

Ocean current data

To evaluate potential reasons for the observed lack of population structure, we sourced data from GPS-equipped drifting capsules that had been deployed in the North Atlantic as part of the ‘Message in a Bottle’ project by Verkís Consulting Engineers. As the great auk was flightless, ocean currents might have influenced its migration patterns. The route taken by the capsules connects some of the main breeding colonies in St Kilda (Scotland), Geirfuglasker/Eldey Island (Iceland), and Funk Island (Canada) (**Figure 4**).

The extrapolation of present-day ocean current data into the past and the interpretation of the data in the context of great auk movements is merely speculative. However, if ocean currents today are comparable to those of past millennia, then the data do at least provide a possible explanation for how great auks travelled across their former range and between breeding colonies (**Figure 4**). A full description of the routes taken by the capsules is provided in Appendix 2.

Demographic history and effective population size

We reconstructed the demographic history of the great auk using the 25 dated mitogenomes (see Materials and Methods for definition of ‘dated’ samples) and found support for a constant population size through time, with no evidence of a population decline. Despite having a high haplotype diversity, our samples had a shallow divergence and their most recent common ancestor was dated to 42,188 ybp (95% credibility interval 24,743–84,894 ybp; see Appendix 3). The effective female population size (N_{ef}) was estimated at 9558 (95% credibility interval 4548–19,665), assuming a generation interval of 12 years (*BirdLife International, 2016a*). To examine the effect of including the undated samples, we repeated the analysis on the complete data set while accounting for the uncertainty in the ages of the undated samples. This second analysis also yielded support for a constant population size, with an effective female population size of 7331 (95% credibility interval 2477–19,492). Census size (N_c) estimates based on the effective population size and the range of known N_e/N_c ratios (*Frankham, 1995*) yielded an expected wide range of 12,292–756,346 individuals.

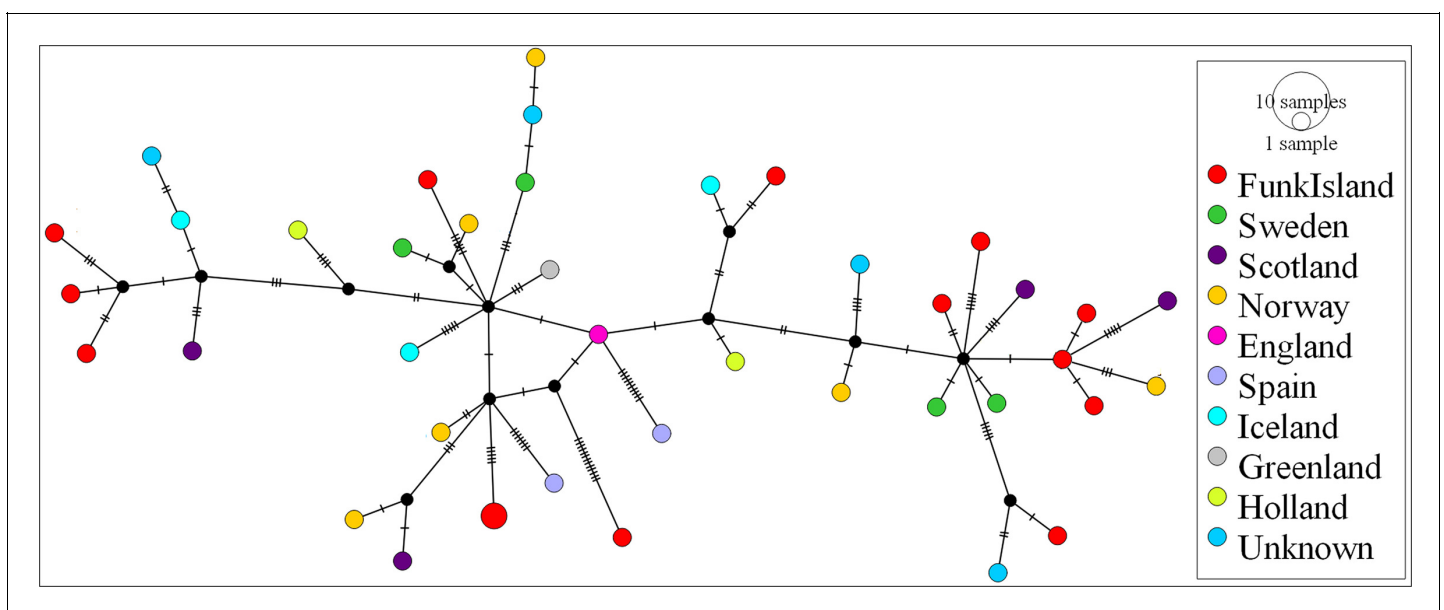


Figure 3. Median-joining network of great auk mitogenomes. The network was inferred in PopART18 and shows a lack of phylogeographic structure among the dated and undated samples of great auks. Haplotypes are coloured according to sampling location.

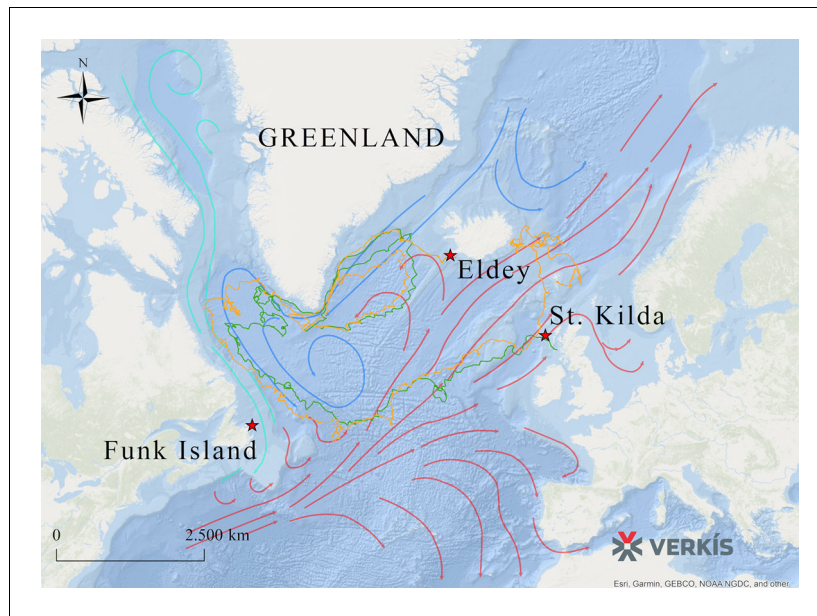


Figure 4. Routes taken by GPS capsules in the North Atlantic. The map shows GPS data from two capsules (green and yellow lines). These tracks show possible routes that the great auk might have used to move between colonies, aided by ocean currents, waves, and wind. Legend: Red Star: Known breeding sites of the great auk (Funk Island, New Newfoundland; Eldey Island, Iceland; St Kilda, Scotland). Green line: GPS capsule 1. Yellow line: GPS capsule 2. Pink arrows: Warm sea currents (Gulf Stream and North Atlantic Drift). Dark blue arrows: Cold sea currents (East Greenland Current and Labrador Current).

Population viability analyses and sustainable harvest rates

To assess the feasibility of a ‘hunting-only’ scenario of extinction, we used population viability analysis to estimate the proportion of the population that would need to have been harvested in order to cause extinction within 350 years. Population sizes for our simulations were conservatively based on the upper margin of the census size estimates outlined above, consistent with the large census sizes described in historic documents (*Birkhead, 1993*) (see also Appendix 8). The estimate of 756,346 mature birds is slightly below the census size estimates for the great auk’s closest relative, the razorbill (*Alca torda*; ~1 million mature birds) and significantly below those of common and thick billed murre, also from the Alcidae family (*Uria aalge* and *Uria lomvia*; 3 million mature birds each) (*BirdLife International, 2016a*; *BirdLife International, 2016b*; *BirdLife International, 2016c*; *BirdLife International, 2017*). Given historic reports of millions of great auks (*Birkhead, 1993*) and in order to reduce the risk of underestimating the census size of great auks, we ran simulations for population sizes of 1 million and 3 million mature birds (2 million and 6 million birds total size including juveniles). All simulation settings were ‘optimistic’ and biased strongly towards survival. This included conservatively high estimates of reproductive success and conservatively low estimates of natural mortality. For a subset of simulations, we also introduced a further, population density dependent, linear reduction of natural mortality to half our already low rates of natural mortality. Furthermore, in order to provide maximum sustainable harvest rate estimates for more ‘realistic’ settings, we ran simulations using estimates for reproductive success and natural mortality obtained from the razorbill.

We found that under our conservative settings, annual harvest rates up to 9% of the pre-hunting population were sustainable. For example, for a pre-hunting population size of 2 million individuals, this corresponds to an annual harvest rate of 180,000 birds. In contrast, an annual harvest rate of 10% of the pre-hunting population combined with an annual egg harvest rate of 5% led to extinction in a large proportion of our simulations. A harvest rate of 10.5% (egg harvest rate 5%) of the pre-hunting population led to extinction within 350 years in all of our simulations. Assuming a density-dependent reduction of mortality had only a small effect on sustainable harvest rates (*Table 1*). Furthermore, even if no eggs at all were harvested, the population was still at risk of extinction at

Table 1. Population viability analysis.

Extinction is defined as 'only one sex remains'. The number of mature individuals was estimated in Vortex 10.2.8.0, assuming a stable age distribution and given our fixed mortality rates. 'Maximum- number of eggs' refers to the number of eggs that would be produced if all mature individuals were breeding. 'Harvest rate' describes the percentage of the population that is harvested annually, with egg harvest rate calculated from the maximum number of eggs in parentheses. 'DD' refers to density-dependent reduction of mortality. 'Number of birds' is the total number of birds killed annually, which was split between the age cohorts (see Appendix 8). 'Number of eggs' is total number of eggs harvested annually.

Conservative settings

| Population size (total) | Mature birds (>4 years) | Maximum number of eggs | Harvest rate (% of starting population size) | DD | Number of birds | Number of eggs | Probability of extinction within 350 years |
|-------------------------|-------------------------|------------------------|--|-----|-----------------|----------------|--|
| 2,000,000 | 1,027,532 | 513,766 | 9 (5) | No | 180,000 | 25,688 | 0.00 |
| 2,000,000 | 1,027,532 | 513,766 | 10 (5) | No | 200,000 | 25,688 | 0.79 |
| 2,000,000 | 1,027,532 | 513,766 | 10 (5) | Yes | 200,000 | 25,688 | 0.22 |
| 2,000,000 | 1,027,532 | 513,766 | 10.5 (5) | Yes | 210,000 | 25,688 | 1.00 |
| 2,000,000 | 1,027,532 | 513,766 | 10.5 (0) | No | 210,000 | 0 | 0.71 |
| 2,000,000 | 1,027,532 | 513,766 | 10.5 (0) | Yes | 210,000 | 0 | 0.19 |
| 6,000,000 | 3,082,594 | 1,541,297 | 9 (5) | No | 540,000 | 77,065 | 0.00 |
| 6,000,000 | 3,082,594 | 1,541,297 | 10 (5) | No | 600,000 | 77,065 | 0.86 |
| 6,000,000 | 3,082,594 | 1,541,297 | 10 (5) | Yes | 600,000 | 77,065 | 0.33 |
| 6,000,000 | 3,082,594 | 1,541,297 | 10.5 (5) | Yes | 630,000 | 77,065 | 1.00 |
| 6,000,000 | 3,082,594 | 1,541,297 | 10.5 (0) | No | 600,000 | 0 | 0.81 |
| 6,000,000 | 3,082,594 | 1,541,297 | 10.5 (0) | Yes | 630,000 | 0 | 0.15 |

'Realistic' settings

| Population size (total) | Mature birds (>5 years) | Maximum number of eggs | Harvest rate (% of starting population size) | DD | Number of birds | Number of eggs | Probability of extinction within 350 years |
|-------------------------|-------------------------|------------------------|--|-----|-----------------|----------------|---|
| 2,000,000 | 1,027,532 | 513,766 | 2 (0) | Yes | 40,000 | 0 | 0.19–0.33 (range across multiple repeat simulations) |

10.5% bird harvest rate, with extinction probabilities between 15% (population size 6 million, density-dependent mortality) and 81% (population size 6 million, no density-dependent mortality, [Table 1]). These results were robust to the definition used for extinction. For comparison, when using the much higher mortality rate of the razorbill, with a starting population of 2 million birds and slightly more realistic settings for reproductive age and success, harvest rates are only sustainable up to about 40,000 birds per year even if no eggs are harvested and mortality is gradually reduced to 50% of the starting value as the population density declines (see *Supplementary file 2b*).

Discussion

Our analyses of the demographic history of great auks support a constant population size within the temporal resolution of our data (back to the most recent common ancestor of all samples 24,000–85,000 ybp). Therefore, we find no evidence of a decline in the population prior to the onset of intensive hunting. We also observed high haplotype diversity across the sampling period, right up to the demise of the species. If the great auk had been at risk of extinction prior to the onset of intensive human hunting, for example as a result of long-term suboptimal habitat or environmental change, we would expect to see genetic evidence of such stress, as for example observed in studies of cave bears (*Stiller et al., 2010*) and bison (*Shapiro et al., 2004*). If, on the other hand, the population declined rapidly, for example as a result of extensive hunting, genetic data would have only very limited power to detect such a decline in a long-lived species. Mitochondrial DNA studies of New Zealand moa found no evidence of a population decline prior to extinction (*Allentoft et al.,*

2014; Rawlence et al., 2012) and a study of the endemic Hawaiian Petrel came to a similar conclusion (Welch et al., 2012). In fact, even a recent whole-genome study of two extinct New Zealand songbirds (huia and South Island kōkako), which disappeared after human settlement within 700 years, found no genetic evidence of population decline prior to the disappearance of the species (Dusseux et al., 2019). Therefore, our results are consistent with a rapid decline of great auks. It is important to keep in mind, though, that our results simply indicate that the demise of the great auk was beyond the detection limit of genetic data. They do not necessarily confirm whether the rapid demise that must have taken place prior to extinction started before or after the onset of extensive human hunting, nor do the results provide an indication of whether there was more than one population decline. A localised, unexplained decline in great auk numbers on the eastern side of the North Atlantic over the past 2,000 years, for example, which has been inferred from a decline in bone finds in England, Scotland, and Scandinavia (Bengtson, 1984; Best and Mulville, 2014; Grieve, 1885; Hufthammer, 1982; Serjeantson, 2001), does not appear to have been severe enough to leave a genetic signature.

The estimated female effective population size is considerably smaller than the census size, which has been estimated to be in the millions (Birkhead, 1993). This is noteworthy because it suggests that the species went through a severe bottleneck in the recent past. The shallow divergence of less than 90,000 years between the sequenced individuals suggests a population decline in the late Pleistocene, potentially associated with climate fluctuations. However, the wide 95% credibility intervals of our divergence-time estimates prevent us from narrowing down the cause of the bottleneck to any specific event. In any case, the high percentage of singleton haplotypes in our data, which is characteristic of a population expansion following a bottleneck (Slatkin and Hudson, 1991), together with the large census size at the onset of intensive hunting, suggest that the great auk had successfully recovered from the bottleneck.

Our genetic analyses failed to detect any female population structure in space or time, indicating a lack of marked barriers to dispersal among populations across the species' range. This is inconsistent with predictions of limited or no interbreeding between populations from either side of the North Atlantic (Burness and Montevecchi, 1992), and suspected regional philopatry in this species (Bengtson, 1984; Montevecchi and Kirk, 1996). Such a lack of structure is, however, common in seabirds, and has been observed in several relatives of the great auk, such as the thick-billed murre (*Uria lomvia*; no structure within ocean basins) (Tigano et al., 2015), common murre (*Uria aalge*; structure in the Atlantic but not in the Pacific) (Morris-Pocock et al., 2008), ancient murrelets (*Synthliboramphus antiquus*; no genetic differentiation in the North Pacific) (Pearce et al., 2009), and little auk (*Alle alle*; no structure in the Arctic) (Wojczulanis-Jakubas et al., 2014). While all of the great auk's closest relatives are capable of flight, which would aid population connectivity, a lack of population structure has similarly been reported from some penguin species. For example, little or no population structure has been reported for the emperor penguin (*Aptenodytes forsteri*) (Cristofari et al., 2016), chinstrap penguin (*Pygoscelis antarcticus*) (Mura-Jornet et al., 2018), and Adélie penguin (*P. adeliae*) (Gorman et al., 2017; Roeder et al., 2001).

We can only speculate what factors may have driven this lack of population structure, but the data collected from the GPS-enabled drifting capsules are consistent with hypotheses put forward by a number of authors. It has been suggested that migrations occurred in both northward and southward directions between breeding and wintering sites, aided by ocean currents such as the East Greenland Current (Brown, 1985; Meldgaard, 1988; Montevecchi and Kirk, 1996). However, as these preliminary data were only available from two GPS-enabled drifting capsules and as ocean currents may have changed significantly over the past few centuries, the conclusions that we can draw from such data are somewhat limited. Furthermore, it is possible that these currents can change throughout the year. Thus, these data must be considered with caution and pending far more detailed studies of ocean currents in the North Atlantic throughout the year. Nevertheless, high vagility of the great auk is further supported by its ability to track its habitat in response to climate change, as evidenced by archaeological records (Bengtson, 1984; Campmas et al., 2010; Meldgaard, 1988; Serjeantson, 2001).

We find no evidence in our genetic data that would suggest that great auk populations were at risk of extinction at the time when human hunting intensified. However, the strength of our conclusions is limited in a number of respects. The mitochondrial genome is only a single genetic marker and our samples were insufficiently preserved to yield nuclear SNP data (Appendix 9), which would

have offered a greater degree of resolution with the potential to detect population structure. Similarly, as a result of limitations in sample preservation and availability, the sample size of 41 is relatively small for population genetic analysis and could have limited our ability to resolve changes in population structure and size.

The key question, therefore, is whether it is at all feasible to assume that the intensive hunting of the 16th–19th centuries alone led to the extinction of the great auk. Our population viability analysis shows that, independent of the population size, harvest rates that would cause extinction under all of the conditions explored in our simulations are well below reasonable estimates of harvest rates as inferred from historical sources. For example, a total population size of 2 million birds corresponds to 1 million mature individuals. This is higher than the upper margin of our census size estimates and is consistent with the census size currently estimated for the great auk's closest relative, the razorbill. At this census size, an annual harvest of 210,000 birds and fewer than 26,000 eggs would have caused the extinction of the great auk within 350 years.

Actual hunting pressure on great auks is likely to have far exceeded 210,000 birds annually. From 1497 AD, when Europeans discovered the rich fishing grounds of Newfoundland, fleets of 300 to 400 ships from various European countries were drawn annually to this region, which is likely to have had the highest population density of great auks (*Bengtson, 1984; Steenstrup, 1855*). Fishing stations were set up near colonies of the great auk and other seabirds, and these colonies were heavily exploited (*Pope, 2009*). Great auks were also likely to have been caught by fishing lines and in fishing nets (*Montevecchi and Kirk, 1996; Piatt and Nettleship, 1985; Piatt and Nettleship, 1987; Pope, 2009*). Contemporary reports document a case in which approximately 1000 great auks were caught and killed within half an hour by two fishing vessels off the coast of Funk Island (*Bengtson, 1984; Grieve, 1885*). Thus, if each of the 400 vessels in the region spent only half an hour a year harvesting great auks at this rate, that would already correspond to 200,000 birds a year.

At a total population size of 6 million birds, corresponding to the estimated 3 million mature individuals of common murre and thick-billed murre in the North Atlantic, an annual harvest of 630,000 birds and 77,000 eggs would cause certain extinction. Even this number does not appear unrealistically high when considering that great auks were also targeted for the feather trade, with hunters living on Funk Island throughout the summer with the purpose of killing the birds (*Gaskell, 2000; Kirkham and Montevecchi, 1982*). Adding to the effects of excessive hunting, the great auk laid only one egg a year, which was not replaced if removed (*Bengtson, 1984*). Thus, replenishing the large number of birds lost annually would have been highly improbable (*Gaskell, 2000*).

Critically, our estimates of harvest rates leading to extinction are likely to be conservatively high, because they are based on some unrealistically optimistic assumptions. For example, our settings assume that 100% of mature birds breed, that they had 100% breeding success, and that their offspring was independent from the time the egg was laid (hence no negative effect of parents being killed). Furthermore, we assumed the lowest natural mortality observed among all alcids for each age class and in some simulations reduced these mortality rates by half when population density declined, thereby considering the positive effects of increased availability of resources and reduced competition. Detrimental effects of small population sizes, such as inbreeding depression, were not included in our simulations. Because very little is known about the biology of the great auk, we chose to use such conservative settings to reduce the risk of underestimating the sustainable harvest rate. However, this brings an increased risk of overestimating the number of birds that could have been sustainably harvested. Using the mortality rate of the razorbill and allowing for more variation in reproductive success (see *Supplementary file 2a*) reduces the sustainable harvest rate for a population of 2 million birds to as few as 40,000 birds per year. However, the razorbill can produce a second egg per season if the first one is lost, so applying razorbill mortality rates to the great auk likely leads to an underestimation of the sustainable harvest rate.

Our conservative simulations require high harvest rates to cause the extinction of the great auk, but these values are largely consistent with harvest rates for present-day species. For example, until recently, between 200,000 and 300,000 murre (*Uria* spp.) were killed legally every year off the eastern Canadian coast (*Wilhelm et al., 2008*). Harvest rates were even higher before the mid-1990s, when between 300,000 and 700,000 thick-billed murre alone were being harvested annually (*Wilhelm et al., 2008*). In Iceland, 150,000 to 233,000 Atlantic puffins were once killed annually, representing about 2–3% of the population. In contrast, 25–30% of the populations of species of black-backed gulls are killed annually (*Merkel and Barry, 2008*). Although current figures for annual

harvest rates of auk species are considerably lower than those given above and continue to decline (e.g., ~25,000 puffins were killed in Iceland in 2016 compared with ~233,000 in 1995 [*Statistics Iceland, 2016*]; also see *Frederiksen et al., 2016*), the harvesting rates required to cause the extinction of the great auk would not be considered excessive even by modern standards.

The roles of humans and environmental changes in causing extinctions have long been debated, not only for the great auk but also for other lost species (*Cooper et al., 2015; Lorenzen et al., 2011; Shapiro et al., 2004*). In contrast with most studies of Pleistocene extinctions, which have argued for at least some level of climate-driven environmental contributions to species extinction, we have found little evidence that the great auk was at risk of extinction prior to the onset of intensive human hunting. Critically, this does not mean that our study provides unequivocal evidence that humans alone were the cause of great auk extinction. To test this hypothesis, simulations of great auk population dynamics in response to environmental change throughout the Holocene would be required. However, with little information about great auk biology, such simulations would be highly speculative. What our study has demonstrated though, is that human hunting pressure alone was very likely to have been high enough to cause extinction even if the great auk population was not already under threat of extinction through environmental change.

Our findings highlight how industrial-scale commercial exploitation of natural resources have the potential to drive even an abundant, wide-ranging, highly vagile, and genetically diverse species to extinction within a short period of time. This echoes the conclusions drawn for the passenger pigeon (*Murray et al., 2017*), which occurred in enormous numbers prior to its extinction in the early 20th century. Our findings emphasise the need for thorough monitoring of commercially harvested species, particularly in poorly researched environments such as our oceans. This will lay the platform for sustainable ecosystems and ensure the evidence-based conservation management of biodiversity.

Materials and methods

Sampling and DNA extraction

We obtained great auk material for ancient DNA (aDNA) analyses from various institutions (*Supplementary file 1a*). Samples were chosen to represent individuals from the major centres of the former geographic distribution of the species (*Figure 1*), spanning as wide a time period as possible (*Supplementary file 1a*). The samples range from about 170 years old to about 13,000–15,000 years old. Sample dates are stratigraphically assigned (archaeological material), based on documented information (e.g., dates on which mounted specimens were killed), or estimated from known site information to give dated constraints (e.g., Funk Island material was collected from the top layers of the islands, so the bones are most likely from individuals killed during the intense hunting period that began ~500 years ago). Bones were sampled via drilling using a Dremel 107 2.4 mm engraving cutter to obtain powdered bone (*Figure 5*) or using a Dremel cutting wheel, which allowed removal of sections of bones that were later powdered using a sonic dismembrator.

All laboratory work prior to polymerase chain reaction (PCR) amplification was carried out in the designated aDNA laboratories of the Natural History Museum of Denmark and the University of Otago. Strict aDNA protocols were followed to avoid contamination. For each DNA extraction and library build, no-template controls were used to test for contamination by exogenous DNA. All post-PCR work was carried out in separate laboratory facilities (*Knapp et al., 2012*).

Genomic DNA was extracted from 20 to 60 mg of bone powder (*Supplementary file 1b*) using the method described by *Dabney et al. (2013)*. In short, the bone powder was digested using an EDTA-based extraction buffer and DNA purified using a Qiagen MinElute column. After washing with ethanol-based wash buffers (Qiagen), the DNA was eluted in TE buffer for storage.

DNA sequence data

Single-stranded sequencing libraries were prepared from aDNA extracts following the protocol by *Gansauge and Meyer (2013)*, with modifications as described by *Bennett et al. (2014)*. For some samples, double-stranded libraries were also built using the protocol described by *Meyer and Kircher (2010)* (*Supplementary file 1b*). Hybridisation capture was used to enrich libraries for great auk mitochondrial DNA following the MYcroarray MYbaits Sequence Enrichment protocol v2.3.1



Figure 5. Great auk humeri following sampling. Great auk humeri, collected from Funk Island, following sampling to collect bone powder for use in DNA extraction. Bones part of the collection at the American Museum of Natural History (Credit: J. Thomas).

(*MYcroarray MYbaits, 2014*). Bait design details can be found in Appendix 4 and **Appendix 4—figure 1**.

Samples were sequenced on Illumina platforms (HiSeq 2500 and MiSeq; further details in **Supplementary file 1b**) at the Danish National High-Throughput DNA Sequencing Centre or by New Zealand Genomics Limited. Demultiplexing of raw sequence data was performed by the respective sequencing centres. Read processing of demultiplexed sequence data was performed as described by *Thomas et al. (2017)* using the PALEOMIX v1.2.5 pipeline (*Schubert et al., 2014*), details of which can be found in Appendix 5.

Demographic history analyses

To reconstruct the demographic history of the great auk through time, we performed a Bayesian phylogenetic analysis of the mitogenome sequences from the 25 dated samples ('dated' being defined here as those with associated date information, such as stratigraphically assigned dates; undated refers to those for which there is no associated dating information, such as the Funk Island samples) (**Supplementary file 1e**). The sequence alignment was analysed using BEAST 1.8.4 (*Drummond et al., 2012*). Full details of the BEAST analysis, including details of the data-partitioning scheme, can be found in Appendix 6.

To test hypotheses of constant population size through time vs. population size increase or decline, we compared the marginal likelihoods of constant-size and exponential-growth coalescent tree priors for our data set. The exponential-growth coalescent tree prior with a positive growth rate yielded a higher marginal likelihood than the constant-size tree prior, suggesting that it was the best model of population dynamics in the great auk. However, the posterior distribution of the population growth rate was highly right-skewed with a mode very close to zero, so we conservatively used the constant-size coalescent tree prior for our analysis.

A second analysis was performed in BEAST, in which the 16 undated mitogenomes were included in the data set. A uniform prior of either (0,1000) or (0,5000) was specified for the ages of these mitogenomes, depending on independent information about the context of the samples (*Shapiro et al., 2011*). All other settings and priors matched those used in the analysis of the 25 dated samples. The extended data set was still best described by a constant-size coalescent prior.

Network analyses

Population structure was investigated by inferring a haplotype network using median joining (*Bandelt et al., 1999*) in PopART (*Leigh and Bryant, 2015*). Genetic diversity through space and time was visualised using statistical parsimony and a temporal haplotype network, as implemented

in TempNet (*Prost and Anderson, 2011*) (see Appendix 7 for details on TempNet age categories and *Supplementary file 1e*).

Population viability analysis

We performed a population viability analysis using the software Vortex 10.2.8.0 (*Lacy and Pollak, 2014*) in order to estimate the number of great auks that were hunted annually, as well as the rate at which a given intensity of hunting would result in population collapse and extinction. Full details of the simulations performed and parameter justifications can be found in Appendix 8 and *Supplementary file 2a, 2b and 2c*.

Tracking migration routes using GPS capsules

To achieve a better understanding of the feasibility of great auk movement between colonies of the North Atlantic, we accessed data that were initially generated as part of the 'Message in a Bottle' project by Verkís Consulting Engineers in Iceland. Two GPS-equipped drifting capsules were released on 10th January 2016 from a helicopter around 40 km southeast of the Reykjanes peninsula (southwestern Iceland). Each of the capsules contained a North Star TrackPack GPS tracking device (<https://www.northstarst.com/asset-trackers/trackpack/>), which uploaded precise location data six times a day for up to two years, through the GlobalStar satellite network.

Acknowledgements

We are very grateful to the archaeological site directors, sample collectors, curators, and institutions that provided samples for this project. We thank Áki Thoroddsen (Verkís) for producing *Figure 4*. Funding was provided through a NERC PhD Studentship (NE/L501694/1) to JET, ERC Consolidator Award (681396-Extinction Genomics) to MTPG, the Genetics Society-Heredity Fieldwork Grant, and European Society for Evolutionary Biology–Godfrey Hewitt Mobility Award to JET. MK is supported by a Rutherford Discovery Fellowship from the Royal Society of New Zealand. We thank members of the Molecular Ecology and Fisheries Genetics Laboratory at Bangor University, EvoGenomics and GeoGenetics at University of Copenhagen, and the Biological Anthropology group at the University of Otago for guidance in the laboratory and on data analysis and for useful discussions. Sequencing was provided by The Danish National High-Throughput DNA Sequencing Centre and New Zealand Genomics Limited.

Additional information

Competing interests

Arnór Þ Sigfússon: Arnór Þ Sigfússon is affiliated with Verkís Consulting Engineers. The author has no financial interests to declare. Vigfús A Jósefsson: Vigfús A Jósefsson is affiliated with Verkís Consulting Engineers. The author has no financial interests to declare. The other authors declare that no competing interests exist.

Funding

| Funder | Grant reference number | Author |
|---|---------------------------------|--------------------|
| NERC Environmental Bioinformatics Centre | NE/L501694/1 | Jessica E Thomas |
| European Research Council | 681396-Extinction Genomics | M Thomas P Gilbert |
| Genetics Society | Heredity Fieldwork Grant | Jessica E Thomas |
| European Society for Evolutionary Biology | Godfrey Hewitt Mobility Award | Jessica E Thomas |
| Royal Society of New Zealand | Rutherford Discovery Fellowship | Michael Knapp |

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Jessica E Thomas, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Visualization, Writing—original draft, Project administration, Writing—review and editing; Gary R Carvalho, Conceptualization, Supervision, Writing—original draft, Project administration, Writing—review and editing; James Haile, Mikkel-Holger S Sinding, Marcela Sandoval-Velasco, Investigation, Writing—original draft; Nicolas J Rawlence, Conceptualization, Writing—original draft, Writing—review and editing; Michael D Martin, Formal analysis, Investigation, Writing—original draft; Simon YW Ho, Data curation, Software, Formal analysis, Investigation, Visualization, Writing—original draft, Writing—review and editing; Arnór P Sigfússon, Vigfús A Jósefsson, Investigation, Visualization, Writing—original draft; Morten Frederiksen, Jannie F Linnebjerg, Christina Barilaro, Juila Best, Dirk Brandis, Chiara Cavallo, Mikelo Elorza, Kimball L Garrett, Maaïke Groot, Friederike Johansson, Jan T Lifjeld, Göran Nilson, Dale Serjeanston, Paul Sweet, Errol Fuller, Anne Karin Hufthammer, Morten Meldgaard, Jon Fjeldså, Resources, Writing—original draft; Jose A Samaniego Castruita, Jonas Niemann, Data curation, Software, Writing—original draft; André ER Soares, Formal analysis, Writing—original draft; Robert Lacy, Software, Formal analysis, Investigation, Writing—original draft, Writing—review and editing; Beth Shapiro, Formal analysis, Writing—original draft, Writing—review and editing; Michael Hofreiter, John R Stewart, Conceptualization, Supervision, Writing—original draft, Writing—review and editing; M Thomas P Gilbert, Conceptualization, Supervision, Funding acquisition, Investigation, Writing—original draft, Project administration, Writing—review and editing; Michael Knapp, Conceptualization, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Jessica E Thomas  <https://orcid.org/0000-0002-9043-646X>

James Haile  <http://orcid.org/0000-0002-8521-8337>

Simon YW Ho  <https://orcid.org/0000-0002-0361-2307>

Michael Knapp  <https://orcid.org/0000-0002-0937-5664>

Decision letter and Author response

Decision letter <https://doi.org/10.7554/eLife.47509.SA1>

Author response <https://doi.org/10.7554/eLife.47509.SA2>

Additional files

Supplementary files

- Source data 1. Nuclear SNP bait design.
- Supplementary file 1. Sample Information. **Supplementary file 1a** Sample information for all samples collected. Information listed shows institution name and number where sample was sourced, the site location and country where sample was discovered (if known), and any associated date/age information, if known. Those highlighted indicate samples that ultimately passed the filtering settings and were used in the final analysis. Asterisks indicate samples from *Thomas et al. (2017)*. **Supplementary file 1b** Lab process table for all samples collected. Table includes information on sample type, weight used in extraction, which library build method was used, if hybridization capture was used, and which type of sequencing was performed. Those highlighted indicate samples that ultimately passed the filtering settings and were used in the final analysis. Asterisks indicate samples from *Thomas et al. (2017)*. **Supplementary file 1c** PALEOMIX summary data for mitogenome samples. Summary statistics table from all great auk samples sent for sequencing. Library type: PE = Paired end, SE = Single end, *=both. Samples highlighted were used in final analysis. **Supplementary file 1d** GenBank accession numbers. GenBank accession numbers for samples used in analysis. **Supplementary file 1e** Age information for samples used in analysis. Age information for samples used in the BEAST and TempNet analyses.

- Supplementary file 2. Population Viability Analysis Settings. **Supplementary file 2a** Settings used in Population Viability Analysis. Details of the settings used for Population Viability Analysis performed in Vortex 10.2.8.0. Information displayed corresponds to the various setting sections in the software and the variables that were changed. Further details on justification for these settings can be found in Appendix 8. **Supplementary file 2b** Details of mortality rates used in Population Viability Analysis. Details of the mortality rate settings used in Population Viability Analysis performed in Vortex 10.2.8.0, showing formula information for including density-dependent change and additional justification. **Supplementary file 2c** Harvest rate calculations. Example of how harvest rates of birds and eggs were calculated for the Population Viability Analysis.
- Transparent reporting form

Data availability

Sequence data are available on NCBI GenBank under the Popset IDs 1735592912 and 1208276182.

The following dataset was generated:

| Author(s) | Year | Dataset title | Dataset URL | Database and Identifier |
|--|------|--|---|-------------------------|
| Thomas JE, Carvalho GR, Haile J, Rawlence NJ, Martin MD, Ho SYW, Sigfusson AP, Josefsson VA, Frederiksen M, Linnebjerg JF, Samaniego Castruita JA, Niemann J, Sinding M-HS, Sandoval-Velasco M, Soares AER, Lacy R, Barilaro C, Best J, Brandis D, Cavallo C, Elorza M, Garrett KL, Groot M, Johansson F, Liffjeld JT, Nilson G, Serjeantson D, Sweet P, Fuller E, Hufthammer AK, Meldgaard M, Fjeldsa J, Shapiro B, Hofreiter M, Stewart JR, Gilbert MTP, Knapp M | 2019 | Pinguinus impennis mitochondrion, partial genome | https://www.ncbi.nlm.nih.gov/popset/?term=1735592912 | NCBI Popset, 1735592912 |

The following previously published dataset was used:

| Author(s) | Year | Dataset title | Dataset URL | Database and Identifier |
|---|------|---|---|-------------------------|
| Thomas JE, Carvalho GR, Haile J, Martin MD, Castruita JAS, Niemann J, Sinding MS, Sandoval-Velasco M, Rawlence NJ, Fuller E, Fjeldsa J, Hofreiter M, Stewart JR, Gilbert MTP, Knapp M | 2017 | Pinguinus impennis mitochondrion, complete genome | https://www.ncbi.nlm.nih.gov/popset/?term=1208276182 | NCBI PopSet, 1208276182 |

References

- Allentoft ME, Heller R, Oskam CL, Lorenzen ED, Hale ML, Gilbert MT, Jacomb C, Holdaway RN, Bunce M. 2014. Extinct New Zealand megafauna were not in decline before human colonization. *PNAS* **111**:4922–4927. DOI: <https://doi.org/10.1073/pnas.1314972111>, PMID: 24639531
- Anijalg P, Ho SYW, Davison J, Keis M, Tammeleht E, Bobowik K, Tumanov IL, Saveljev AP, Lyapunova EA, Vorobiev AA, Markov NI, Kryukov AP, Kojola I, Swenson JE, Hagen SB, Eiken HG, Paule L, Saarma U. 2018. Large-scale migrations of Brown bears in Eurasia and to North America during the late Pleistocene. *Journal of Biogeography* **45**:394–405. DOI: <https://doi.org/10.1111/jbi.13126>
- Anmarkrud JA, Lijfeld JT. 2017. Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular Ecology Resources* **17**:334–341. DOI: <https://doi.org/10.1111/1755-0998.12600>, PMID: 27654125
- Bandelt HJ, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**:37–48. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a026036>, PMID: 10331250
- Bengtson S-A. 1984. Breeding ecology and extinction of the great auk (*Pinguinus impennis*): Anecdotal evidence and conjectures. *The Auk* **101**:1–12. DOI: <https://doi.org/10.1093/auk/101.1.1>
- Bennett EA, Massilani D, Lizzo G, Daligault J, Geigl EM, Grange T. 2014. Library construction for ancient genomics: single strand or double strand? *BioTechniques* **56**:289–300. DOI: <https://doi.org/10.2144/000114176>, PMID: 24924389
- Best J. 2013. PhD Thesis: Living in liminality: an osteoarchaeological investigation into the use of avian resources in North Atlantic Island environments. Cardiff University. <http://orca.cf.ac.uk/58668/>
- Best J, Mulville J. 2014. A bird in the hand: data collation and novel analysis of avian remains from South Uist, Outer Hebrides. *International Journal of Osteoarchaeology* **24**:384–396. DOI: <https://doi.org/10.1002/oa.2381>
- Best J, Mulville J. 2016. Birds from the water: reconstructing avian resource use and contribution to diet in prehistoric Scottish island environments. *Journal of Archaeological Science: Reports* **6**:654–664. DOI: <https://doi.org/10.1016/j.jasrep.2015.11.024>
- BirdLife International. 2016a. *Pinguinus impennis*. The IUCN Red List of Threatened Species 2016: IUCN.
- BirdLife International. 2016b. The IUCN Red List of Threatened Species. *Uria lomvia*. <http://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T22694847A86853272>
- BirdLife International. 2016c. *Uria aalge*. <http://www.iucnredlist.org/details/22694841/0>
- BirdLife International. 2017. *Alca torda* (amended version of 2016 assessment). The IUCN Red List of Threatened Species 2017: IUCN. DOI: <https://doi.org/10.2305/IUCN.UK.2016.RLTS.T22694852A110637027.en>
- Birkhead TR. 1993. *Great Auk Islands: a field biologist in the Arctic*. Poysler.
- Broad Institute. 2019. Picard Tools - by Broad Institute. <http://broadinstitute.github.io/picard/>
- Brown RG. 1985. The Atlantic Alcidae at sea. In: Nettleship D. N, Birkhead T. R (Eds). *The Atlantic Alcidae*. London: Academic Press. p. 264–318.
- Burness GP, Montevecchi WA. 1992. Oceanographic-related variation in the bone sizes of extinct great auks. *Polar Biology* **11**:545–551. DOI: <https://doi.org/10.1007/BF00237947>
- Campmas E, Laroulandie V, Michel P, Amani F, Nespoulet R, El Hajraoui MA. 2010. A Great Auk (*Pinguinus impennis*) in North Africa: Discovery of a bone remain in Neolithic layer of El Harhoura 2 Cave (Temara, Morocco). In: *Birds in Archaeology. Proceedings of the 6th Meeting of the ICAZ Bird Working Group in Groningen*. Groningen Institute of Archaeology. p. 233–240.
- Chang D, Knapp M, Enk J, Lippold S, Kircher M, Lister A, MacPhee RDE, Widga C, Czechowski P, Sommer R, Hodges E, Stümpel N, Barnes I, Dalén L, Derevianko A, Germonpré M, Hillebrand-Voiculescu A, Constantin S, Kuznetsova T, Mol D, et al. 2017. The evolutionary and phylogeographic history of woolly mammoths: a comprehensive mitogenomic analysis. *Scientific Reports* **7**:44585. DOI: <https://doi.org/10.1038/srep44585>, PMID: 28327635
- Cooper A, Turney C, Hughen KA, Brook BW, McDonald HG, Bradshaw CJ. 2015. PALEOECOLOGY. abrupt warming events drove late Pleistocene Holarctic megafaunal turnover. *Science* **349**:602–606. DOI: <https://doi.org/10.1126/science.aac4315>, PMID: 26250679
- Cristofari R, Bertorelle G, Ancel A, Benazzo A, Le Maho Y, Ponganis PJ, Stenseth NC, Trathan PN, Whittington JD, Zanetti E, Zitterbart DP, Le Bohec C, Trucchi E. 2016. Full circumpolar migration ensures evolutionary unity in the Emperor penguin. *Nature Communications* **7**:11842. DOI: <https://doi.org/10.1038/ncomms11842>, PMID: 27296726
- Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga JL, Meyer M. 2013. Complete mitochondrial genome sequence of a middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *PNAS* **110**:15758–15763. DOI: <https://doi.org/10.1073/pnas.1314445110>, PMID: 24019490
- De Santo TL, Nelson SK. 1995. Chapter 3: Comparative Reproductive Ecology of the Auks (Family Alcidae) with Emphasis on the Marbled Murrelet. In: Ralph J. F, John C. H, George L, Martin R, Piatt G (Eds). *Ecology and Conservation of the Marbled Murrelet*. Albany, CA: Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture. 33–47.
- Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* **29**:1969–1973. DOI: <https://doi.org/10.1093/molbev/mss075>, PMID: 22367748

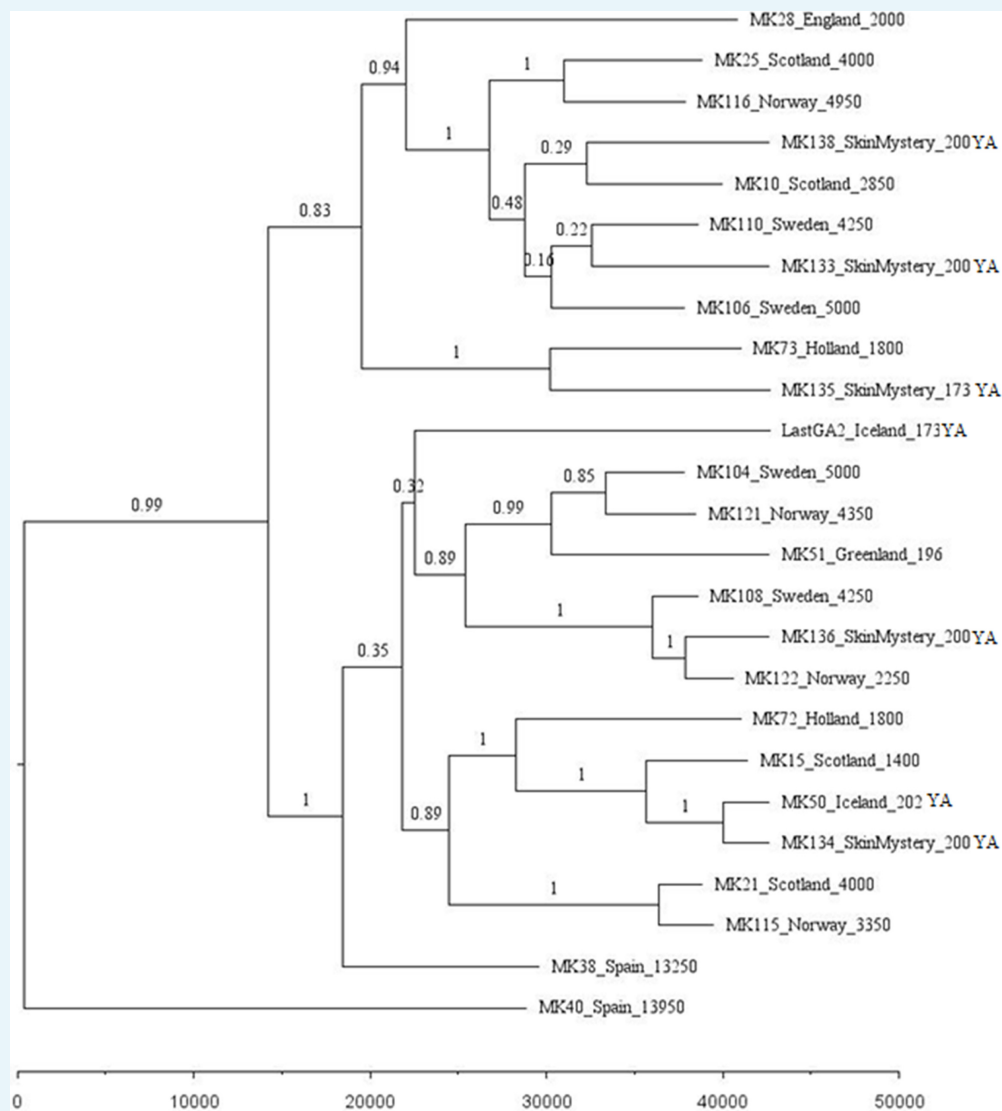
- Duchêne S**, Duchêne D, Holmes EC, Ho SY. 2015. The performance of the Date-Randomization test in phylogenetic analyses of Time-Structured virus data. *Molecular Biology and Evolution* **32**:1895–1906. DOI: <https://doi.org/10.1093/molbev/msv056>
- Dussex N**, von Seth J, Knapp M, Kardailsky O, Robertson BC, Dalén L. 2019. Complete genomes of two extinct new zealand passerines show responses to climate fluctuations but no evidence for genomic erosion prior to extinction. *Biology Letters* **15**:20190491. DOI: <https://doi.org/10.1098/rsbl.2019.0491>, PMID: 31480938
- Felsenstein J**. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783–791. DOI: <https://doi.org/10.1111/j.1558-5646.1985.tb00420.x>
- Frankham R**. 1995. Effective population size/adult population size ratios in wildlife: a review. *Genetical Research* **66**:95–107. DOI: <https://doi.org/10.1017/S0016672300034455>
- Franklin IR**. 1980. Evolutionary change in small populations. In: Soule M. E, Wilcox B. A (Eds.), *Conservation Biology: An Evolutionary Ecological Perspective*. Sinauer Associates: Sunderland, Mass. pp. 135–140.
- Frederiksen M**, Descamps S, Erikstad KE, Gaston AJ, Gilchrist HG, Grémillet D, Johansen KL, Kolbeinson Y, Linnebjerg JF, Mallory ML, McFarlane Tranquilla LA, Merkel FR, Montevecchi WA, Mosbech A, Reiertsen TK, Robertson GJ, Steen H, Strøm H, Thórarinnsson TL. 2016. Migration and wintering of a declining seabird, the thick-billed murre *Uria lomvia*, on an ocean basin scale: Conservation implications. *Biological Conservation* **200**:26–35. DOI: <https://doi.org/10.1016/j.biocon.2016.05.011>
- Fuller E**. 1999. *The Great Auk*. Kent, England: Errol Fuller.
- Gansauge M-T**, Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols* **8**:737–748. DOI: <https://doi.org/10.1038/nprot.2013.038>
- Gaskell J**. 2000. *Who Killed the Great Auk?* Oxford University Press.
- Gorman KB**, Talbot SL, Sonsthagen SA, Sage GK, Gravely MC, Fraser WR, Williams TD. 2017. Population genetic structure and gene flow of Adélie penguins (*Pygoscelis adeliae*) breeding throughout the western Antarctic Peninsula. *Antarctic Science* **29**:499–510. DOI: <https://doi.org/10.1017/S0954102017000293>
- Gouy M**, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**:221–224. DOI: <https://doi.org/10.1093/molbev/msp259>, PMID: 19854763
- Grieve S**. 1885. *The Great Auk, or Garefowl, Its History, Archaeology and Remains (Digitally)*. Cambridge, United Kingdom: Cambridge University Press.
- Hall TA**. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. In: *Nucleic Acids Symposium Series* **4141**. scinapse 9595 9898.
- Halliday T**. 1978. *Vanishing Birds: Their Natural History and Conservation*. Holt, Rinehart and Winston.
- Hufthammer AK**. 1982. *Geirfuglens Utbredelse Og Morfologiske Variasjon I Skandinavia*. Universitetet i Bergen.
- Jónsson H**, Ginolhac A, Schubert M, Johnson PL, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**:1682–1684. DOI: <https://doi.org/10.1093/bioinformatics/btt193>, PMID: 23613487
- Keane TM**, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology* **6**:29. DOI: <https://doi.org/10.1186/1471-2148-6-29>, PMID: 16563161
- Kirkham IR**, Montevecchi WA. 1982. The breeding birds of Funk Island, Newfoundland: an historical perspective. *American Birds* **36**:111–118.
- Knapp M**, Clarke AC, Horsburgh KA, Matisoo-Smith EA. 2012. Setting the stage - building and working in an ancient DNA laboratory. *Annals of Anatomy - Anatomischer Anzeiger* **194**:3–6. DOI: <https://doi.org/10.1016/j.aanat.2011.03.008>, PMID: 21514120
- Kumar S**, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* **35**:1547–1549. DOI: <https://doi.org/10.1093/molbev/msy096>
- Lacy RC**, Pollak JP. 2014. Vortex: A Stochastic Simulation of the Extinction Process. Version 10.0. *Chicago Zoological Society*.
- Laroulandie V**, Elorza M, Berganza E. 2016. Les oiseaux marins du Magdalénien supérieur de Santa Catalina (Lekeitio, Biscaye, Espagne) : approches taphonomique et archéozoologique. In: Dupont C, Marchand G (Eds). *Archéologie Des Chasseurs-Cueilleurs Maritimes. De La Fonction Des Habitats À L'organisation De L'espace Littoral, Actes De La Séance De La Société Préhistorique Française De Renne, Seapeople 2014*. Paris: Société préhistorique française (Séances SPF 6). p. 37–57.
- Leigh JW**, Bryant D. 2015. Popart : full-feature software for haplotype network construction. *Methods in Ecology and Evolution* **6**:1110–1116. DOI: <https://doi.org/10.1111/2041-210X.12410>
- Li H**, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079. DOI: <https://doi.org/10.1093/bioinformatics/btp352>, PMID: 19505943
- Li H**. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. <https://arxiv.org/abs/1303.3997>.
- Li H**, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**:1754–1760. DOI: <https://doi.org/10.1093/bioinformatics/btp324>, PMID: 19451168
- Lindgreen S**. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Research Notes* **5**:337. DOI: <https://doi.org/10.1186/1756-0500-5-337>, PMID: 22748135
- Lorenzen ED**, Nogués-Bravo D, Orlando L, Weinstock J, Binladen J, Marske KA, Ugan A, Borregaard MK, Gilbert MT, Nielsen R, Ho SY, Goebel T, Graf KE, Byers D, Stenderup JT, Rasmussen M, Campos PF, Leonard JA,

- Koepfli KP, Froese D, et al. 2011. Species-specific responses of late quaternary megafauna to climate and humans. *Nature* **479**:359–364. DOI: <https://doi.org/10.1038/nature10574>, PMID: 22048313
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303. DOI: <https://doi.org/10.1101/gr.107524.110>, PMID: 20644199
- Meldgaard M. 1988. The great auk, *Pinguinus impennis* (L.) in Greenland. *Historical Biology* **1**:145–178. DOI: <https://doi.org/10.1080/08912968809386472>
- Merkel F, Barry T. 2008. *Seabird Harvest in the Arctic: Conservation of Arctic Flora and Fauna*. <http://hdl.handle.net/11374/190>.
- Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* **2010**:pdb.prot5448. DOI: <https://doi.org/10.1101/pdb.prot5448>, PMID: 20516186
- Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using tablet for visual exploration of second-generation sequencing data. *Briefings in Bioinformatics* **14**:193–202. DOI: <https://doi.org/10.1093/bib/bbs012>, PMID: 22445902
- Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution* **25**:1459–1471. DOI: <https://doi.org/10.1093/molbev/msn090>, PMID: 18408232
- Montevicchi WA, Kirk DA. 1996. Great Auk (*Pinguinus impennis*). *The Birds of North America Online*.
- Morris-Pocock JA, Taylor SA, Birt TP, Damus M, Piatt JF, Warheit KI, Friesen VL. 2008. Population genetic structure in Atlantic and Pacific ocean common murrelets (*Uria aalge*): natural replicate tests of post-Pleistocene evolution. *Molecular Ecology* **17**:4859–4873. DOI: <https://doi.org/10.1111/j.1365-294X.2008.03977.x>, PMID: 19140977
- Moum T, Arnason U, Árnason E. 2002. Mitochondrial DNA sequence evolution and phylogeny of the Atlantic Alcidae, including the extinct great auk (*Pinguinus impennis*). *Molecular Biology and Evolution* **19**:1434–1439. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a004206>
- Mura-Jornet I, Pimentel C, Dantas GPM, Petry MV, González-Acuña D, Barbosa A, Lowther AD, Kovacs KM, Poulin E, Vianna JA. 2018. Correction to: chinstrap penguin population genetic structure: one or more populations along the southern ocean? *BMC Evolutionary Biology* **18**:117. DOI: <https://doi.org/10.1186/s12862-018-1231-0>, PMID: 30045693
- Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL, Gilbert MTP, Heintzman PD, Letts B, McIntosh G, O’Connell BL, Peck M, Pipes ML, Rice ES, Santos KM, Sohrweide AG, et al. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. *Science* **358**:951–954. DOI: <https://doi.org/10.1126/science.aao0960>, PMID: 29146814
- MYcroarray MYbaits. 2014. MYcroarray MYbaits sequence Enrichment for Targeted Sequencing User Manual. <https://arborbiosci.com/wp-content/uploads/2018/04/MYbaits-manual-v2.pdf> [Accessed May 17, 2017].
- MYcroarray MYbaits. 2016. Manual for MYbaits target enrichment kit v3. <https://manualzz.com/doc/7072779/mybaits-v2-manual> [Accessed September 15, 2017].
- Nei M, Kumar S. 2000. *Molecular Evolution and Phylogenetics*. In Oxford University Press.
- Newton A. 1861. XLII.-Abstract of Mr. J. Wolley’s Researches in Iceland respecting the Gare-fowl or Great Auk (*Alea impennis*, Linn.). *Ibis* **3**:374–399. DOI: <https://doi.org/10.1111/j.1474-919X.1861.tb08857.x>
- Pearce RL, Wood JJ, Artukhin Y, Birt TP, Damus M, Friesen VL. 2009. Mitochondrial DNA suggests high gene flow in ancient murrelets. *The Condor* **104**:84–91. DOI: <https://doi.org/10.1093/condor/104.1.84>
- Piatt JF, Nettleship DN. 1985. Diving depths of four alcids. *The Auk* **102**:293–297. DOI: <https://doi.org/10.2307/4086771>
- Piatt JF, Nettleship DN. 1987. Incidental catch of marine birds and mammals in fishing nets off Newfoundland, Canada. *Marine Pollution Bulletin* **18**:344–349. DOI: [https://doi.org/10.1016/S0025-326X\(87\)80023-1](https://doi.org/10.1016/S0025-326X(87)80023-1)
- Pope PE. 2009. Early migratory fishermen and Newfoundland’s Seabird Colonies. *Journal of the North Atlantic* **102**:57–70. DOI: <https://doi.org/10.3721/037.002.s107>
- Prost S, Anderson CNK. 2011. TempNet: a method to display statistical parsimony networks for heterochronous DNA sequence data. *Methods in Ecology and Evolution* **2**:663–667. DOI: <https://doi.org/10.1111/j.2041-210X.2011.00129.x>
- Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* **16**:395–399. DOI: <https://doi.org/10.1093/bioinformatics/16.4.395>, PMID: 10869038
- Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution* **2**:vew007. DOI: <https://doi.org/10.1093/ve/vew007>, PMID: 27774300
- Ramsden C, Melo FL, Figueiredo LM, Holmes EC, Zanotto PM, VGDN Consortium. 2008. High rates of molecular evolution in hantaviruses. *Molecular Biology and Evolution* **25**:1488–1492. DOI: <https://doi.org/10.1093/molbev/msn093>, PMID: 18417484
- Rawlence NJ, Metcalf JL, Wood JR, Worthy TH, Austin JJ, Cooper A. 2012. The effect of climate and environmental change on the megafaunal moa of New Zealand in the absence of humans. *Quaternary Science Reviews* **50**:141–153. DOI: <https://doi.org/10.1016/j.quascirev.2012.07.004>
- Roeder AD, Marshall RK, Mitchelson AJ, Visagathilagar T, Ritchie PA, Love DR, Pakai TJ, McPartlan HC, Murray ND, Robinson NA, Kerry KR, Lambert DM. 2001. Gene flow on the ice: genetic differentiation among adélie

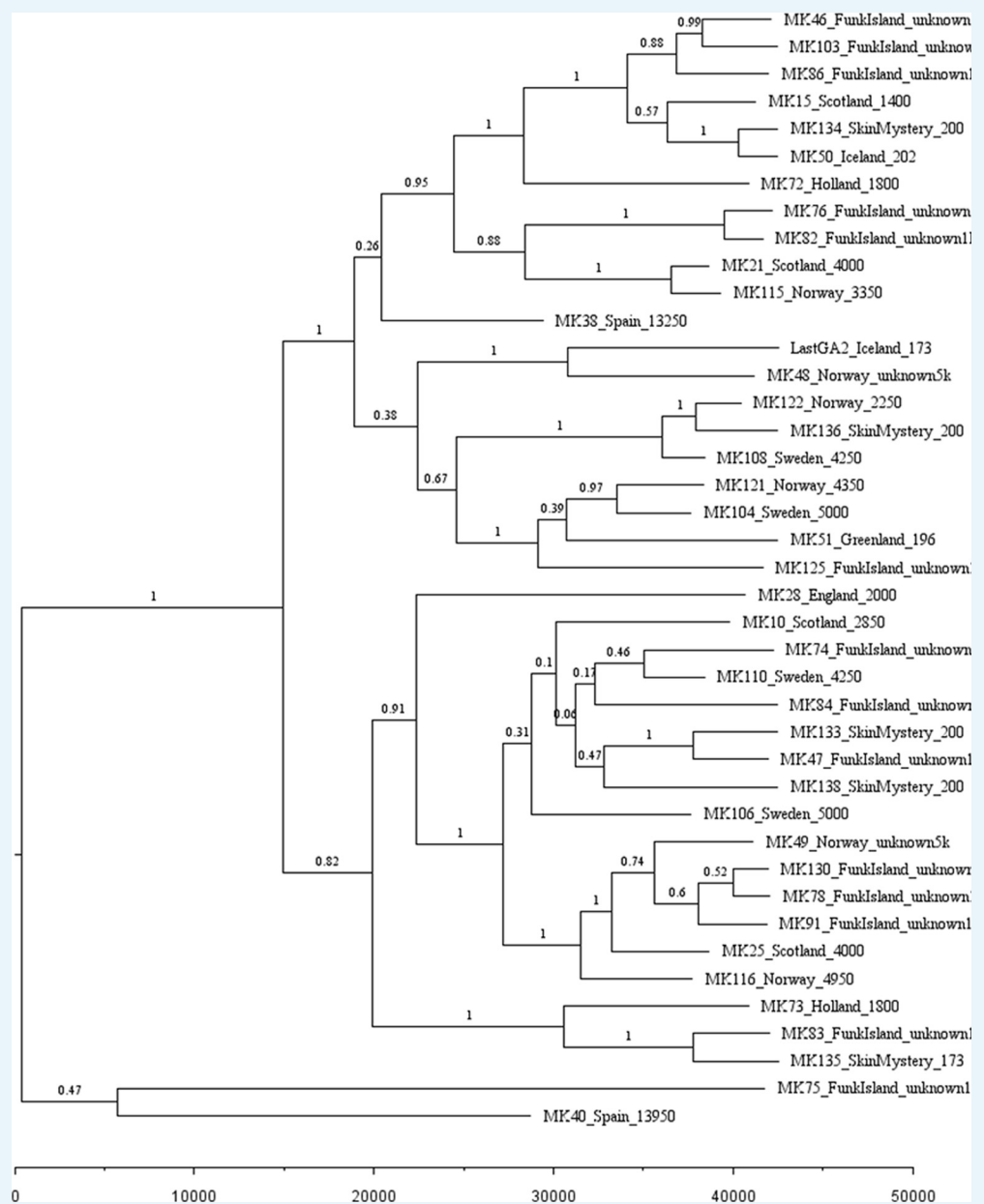
- penguin colonies around Antarctica. *Molecular Ecology* **10**:1645–1656. DOI: <https://doi.org/10.1046/j.0962-1083.2001.01312.x>, PMID: 11472533
- Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:406–425. DOI: <https://doi.org/10.1093/oxfordjournals.molbev.a040454>, PMID: 3447015
- Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**:1056–1082. DOI: <https://doi.org/10.1038/nprot.2014.063>, PMID: 24722405
- Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* **9**:88. DOI: <https://doi.org/10.1186/s13104-016-1900-2>, PMID: 26868221
- Serjeantson D. 2001. The great auk and the gannet: a prehistoric perspective on the extinction of the great auk. *International Journal of Osteoarchaeology* **11**:43–55. DOI: <https://doi.org/10.1002/oa.545>
- Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, Gilbert MT, Barnes I, Binladen J, Willerslev E, Hansen AJ, Baryshnikov GF, Burns JA, Davydov S, Driver JC, Froese DG, Harington CR, Keddie G, Kosintsev P, et al. 2004. Rise and fall of the beringian steppe Bison. *Science* **306**:1561–1565. DOI: <https://doi.org/10.1126/science.1101074>, PMID: 15567864
- Shapiro B, Ho SY, Drummond AJ, Suchard MA, Pybus OG, Rambaut A. 2011. A bayesian phylogenetic method to estimate unknown sequence ages. *Molecular Biology and Evolution* **28**:879–887. DOI: <https://doi.org/10.1093/molbev/msq262>, PMID: 20889726
- Slatkin M, Hudson RR. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**:555–562. PMID: 1743491
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. DOI: <https://doi.org/10.1093/bioinformatics/btu033>
- Statistics Iceland. 2016. Hunting 1995–2016. *Statistics Iceland*. http://px.hagstofa.is/pxen/pxweb/en/Atvinnuvegir/Atvinnuvegir_landbunadur_landveidi/SJA10303.px [Accessed February 13, 2018].
- Steenstrup JJ. 1855. Et Bidrag Til Geirfuglens, Alca Impennis, Naturhistorie, Og Særligt Til Kundskaben Om Dens Tidligere Udbredningskreds. In: *Videnskabelige Meddelelser Fra Den Naturhistoriske Forening I Kjöbenhavn*. Naturhistoriske forening i Kjöbenhavn. p. 33–118.
- Stiller M, Baryshnikov G, Bocherens H, Grandal d’Anglade A, Hilpert B, Münzel SC, Pinhasi R, Rabeder G, Rosendahl W, Trinkaus E, Hofreiter M, Knapp M. 2010. Withering away—25,000 years of genetic decline preceded cave bear extinction. *Molecular Biology and Evolution* **27**:975–978. DOI: <https://doi.org/10.1093/molbev/msq083>, PMID: 20335279
- Thomas JE, Carvalho GR, Haile J, Martin MD, Castruita JAS, Niemann J, Sinding M-HS, Sandoval-Velasco M, Rawlence NJ, Fuller E, Fjeldså J, Hofreiter M, Stewart JR, Gilbert MTP, Knapp M. 2017. An ‘Aukward’ Tale: A Genetic Approach to Discover the Whereabouts of the Last Great Auks. *Genes* **8**:164. DOI: <https://doi.org/10.3390/genes8060164>
- Tigano A, Damus M, Birt TP, Morris-Pocock JA, Artukhin YB, Friesen VL. 2015. The arctic: glacial refugium or area of secondary contact? inference from the population genetic structure of the Thick-Billed murre (*Uria lomvia*), with implications for management. *Journal of Heredity* **106**:238–246. DOI: <https://doi.org/10.1093/jhered/esv016>, PMID: 25825313
- To TH, Jung M, Lycett S, Gascuel O. 2016. Fast dating using Least-Squares criteria and algorithms. *Systematic Biology* **65**:82–97. DOI: <https://doi.org/10.1093/sysbio/syv068>, PMID: 26424727
- Welch AJ, Wiley AE, James HF, Ostrom PH, Stafford TW, Fleischer RC. 2012. Ancient DNA reveals genetic stability despite demographic decline: 3,000 years of population history in the endemic hawaiian petrel. *Molecular Biology and Evolution* **29**:3729–3740. DOI: <https://doi.org/10.1093/molbev/mss185>, PMID: 22844071
- Wilhelm SI, Gilliland SG, Robertson GJ, Ryan PC, Elliot RD. 2008. Development and validation of a wing key to improve harvest management of alcids in the northwest atlantic. *Journal of Wildlife Management* **72**:1026–1034. DOI: <https://doi.org/10.2193/2007-232>
- Wojczulanis-Jakubas K, Kilikowska A, Harding AMA, Jakubas D, Karnovsky NJ, Steen H, Strøm H, Welcker J, Gavrilov M, Lifjeld JT, Johnsen A. 2014. Weak population genetic differentiation in the most numerous arctic seabird, the little auk. *Polar Biology* **37**:621–630. DOI: <https://doi.org/10.1007/s00300-014-1462-5>
- Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. 2011. Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology* **60**:150–160. DOI: <https://doi.org/10.1093/sysbio/syq085>, PMID: 21187451
- Zhang G, Li B, Li C, Gilbert MTP, Jarvis ED, Wang J. 2014. Comparative genomic data of the avian phylogenomics project. *GigaScience* **3**:26. DOI: <https://doi.org/10.1186/2047-217X-3-26>

Appendix 1

Phylogenetic trees



Appendix 1—figure 1. Phylogenetic tree showing the relationships among dated mitogenomes from the great auk. This maximum-clade-credibility tree was inferred by Bayesian analysis in BEAST. Nodes are labelled with posterior probabilities. The tree is drawn to a timescale, as indicated by the horizontal scale bar. Samples included in the analysis are those with associated date information (see *Supplementary file 1e*). For samples with a stratigraphically assigned date the median age has been used. Tip labels give the sample names, sampling locations, and sample ages (years before present, with the exception of mounted specimens labelled YA- years ago).



Appendix 1—figure 2. Phylogenetic tree showing the relationships among dated and undated mitogenomes from the great auk. This maximum-clade-credibility tree was inferred by Bayesian analysis in BEAST. Nodes are labelled with posterior probabilities. The tree is drawn to a timescale, as indicated by the horizontal scale bar. Samples included in the analysis are those with and without associated date information (**Supplementary file 1e**). For samples with a stratigraphically assigned date the median age has been used. Tip labels give the sample names, sampling locations, and sample ages (years before present, with the exception of mounted specimens labelled YA- years ago).

Appendix 2

GPS-equipped drifting capsules: Full result

Following release, easterly winds prevailed and the two GPS-equipped drifting capsules drifted westwards past the tip of the Reykjanes peninsula and past Eldey Island (**Figure 4**). Over the next two weeks, the capsules drifted towards Greenland and when located near the continental shelf started drifting southwards along the coast. The capsules then followed the track of the Icelandic Low, a low-pressure area found between Iceland and Southern Greenland in winter.

The Icelandic Low took the capsules in an anti-clockwise circle back towards Iceland, and onward again towards the west coast of Greenland. The Icelandic Low weakens in summer, so in late April the capsules turned westwards past the southern tip of Greenland and into the Labrador Sea. In summer, they drifted slowly towards the Labrador coast until the beginning of August when they started drifting south-eastwards along the coast of Labrador and Newfoundland and past Funk Island and around 500 km east. By the end of October, the capsules start to follow the trail of the winter low pressures across the Atlantic.

At the beginning of January, capsule one drifted eastwards, around 50 km south of St Kilda and came ashore on the island of Tiree (15.01.2017). Capsule two drifted northwards, passing around 70 km west of St Kilda and west of the Faeroes towards Iceland. In early March, the capsule was around 20 km from the east coast of Iceland when it turned eastwards and then towards south by the beginning of April. It drifted towards the Faeroes where it came ashore on the island of Sandoy (13.05.2017).

The forces driving the capsules are currents, wind, and waves. The capsules got trapped in the Iceland Low where the wind direction is in a counter clockwise circle in winter in the Denmark Strait. In spring, when the Iceland Low starts dissolving, they pass Cape Farewell and, in summer, they drift slowly in calmer summer winds and followed the cold current towards the Labrador coast and then along the coast of Labrador and Newfoundland. In autumn, they hit the path of lows crossing the Atlantic as well as following the warmer Gulfstream. In spring, when at the east coast of Iceland, the weather was calmer and thus capsule two drifted slowly towards and then away from the coast and ended up in the Faeroes.

Appendix 3

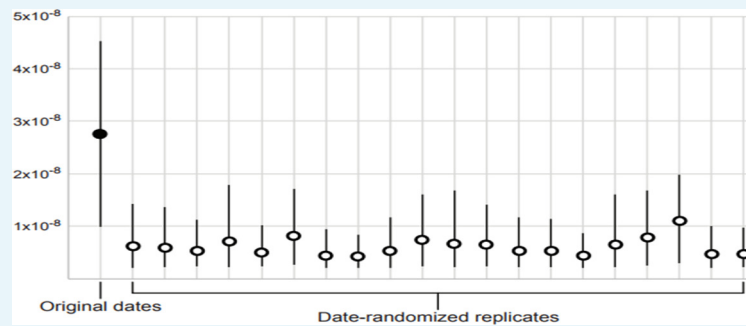
Molecular dating

Estimating the age of the most recent common ancestor (MRCA) of all our samples is not essential to understanding the causes of the extinction of the great auk, but can help with the interpretation of our reconstructions of population dynamics. Our data set is unable to yield reliable information about population dynamics beyond the MRCA of all samples. To infer the evolutionary rate and timescale, we performed a Bayesian phylogenetic analysis of the mitogenome sequences from the 25 dated samples. The analyses were conducted using the same settings and data-partitioning scheme as described in the Methods for our Bayesian phylogenetic analyses.

The sequence alignment was analysed using BEAST 1.8.4 ([Drummond et al., 2012](#)). The evolutionary timescale was estimated using a strict clock model, with the sampling times of the mitogenomes serving as calibrations for the clock ([Rambaut, 2000](#)). Furthermore, to test for the presence of temporal structure in the data set, we performed a date-randomisation test ([Ramsden et al., 2008](#)). We estimated mutation rates from 20 replicate data sets in which the sampling times were permuted and compared these with the rate estimate from the original data set. Two different criteria can be used to determine whether the data set has sufficient temporal structure for generating a reliable estimate of the mutation rate ([Duchêne et al., 2015](#)): if the mean or median estimate from the original data set is not contained within the 95% credibility intervals of the rate estimates from the date-randomised replicates (CR1), or if the 95% credibility intervals of the rate estimates from the date-randomised replicates do not overlap with the 95% credibility interval of the rate estimate from the original data set (CR2).

For comparison, we used two additional methods to estimate the mutation rate. First, we used TempEst ([Rambaut et al., 2016](#)) to estimate the mutation rate using regression of root-to-tip distances against sampling times. Second, we analysed the data using least-squares dating in LSD ([To et al., 2016](#)). For both of these methods, a phylogram was estimated from the dated mitogenome sequences using maximum likelihood in RAxML 8 ([Stamatakis, 2014](#)). Rooting of the tree was inferred by maximising the R-squared value in TempEst and by minimising the objective function in LSD.

Our Bayesian phylogenetic analysis of the dated mitogenomes produced a posterior median estimate of 42,188 years (95% credibility interval 24,743–84,894 years) for the age of the most recent common ancestor. The median posterior estimate of the mutation rate was 2.74×10^{-8} mutations/site/year (95% credibility interval 9.83×10^{-9} – 4.53×10^{-8}). The data set showed some evidence of temporal structure, passing the more lenient criterion CR1 but not the more stringent CR2 of the date-randomisation test ([Appendix 3—figure 1; Duchêne et al., 2015](#)). Thus, with all caution required given the limited temporal structure in our data, our inference of a constant population size for the great auks should be reliable reaching back to the late Pleistocene. However, our data set is not likely to be suitable for drawing strong conclusions about population dynamics of the great auk beyond the last glacial period.

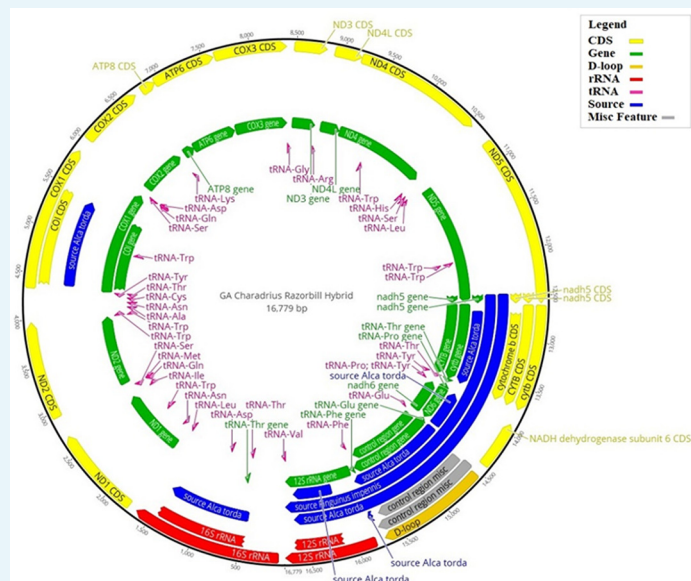


Appendix 3—figure 1. Date-randomisation test for temporal structure in dated mitogenome sequences. The filled circle indicates the median posterior estimate of the mutation rate from the original data set, whereas the empty circles show the median posterior estimates from 20 date-randomised replicate data sets. The 95% credibility intervals (vertical black lines) of the estimates from the date-randomised replicates do not overlap with the median estimate from the original data set, providing some evidence of temporal structure in the data set (criterion CR1). However, the 95% credibility intervals of the estimates from the date-randomised replicates overlap with the 95% credibility interval of the estimate from the original data set, indicating that the data set does not meet the more stringent criterion CR2.

Appendix 4

Bait design

100mer mitochondrial DNA baits (MYcroarray MYbaits) with 50 bp tiling were designed using a hybrid reference mitogenome. This was constructed using the mitogenome from killdeer (*Charadrius vociferus*; assembled from whole genome data, BioProject: PRJNA212867 [Zhang et al., 2014]), with orthologous gene regions replaced by those of great auk where available (GenBank: AJ242685), and those from the razorbill (*Alca torda*; GenBank accessions AJ301680, EF380281, EF380318, and X73916) when great auk data were unavailable (Appendix 4—figure 1).



Appendix 4—figure 1. Hybrid reference mitogenome used for bait design. Illustration of the hybrid reference mitogenome constructed using the killdeer (*Charadrius vociferus*) mitogenome, with orthologous gene regions replaced by those of the great auk (*Pinguinus impennis*), or razorbill (*Alca torda*), when great auk data were unavailable. Annotations correspond to the various regions of the mitogenome: those in blue show where great auk or razorbill genes have been used; yellow corresponds to coding regions; green shows all gene regions; the D-loop is shown in gold; rRNA regions are in red; tRNA regions are in pink; and any miscellaneous features are in grey. The numbers on the outer black circle correspond to the base position of the mitogenome.

Appendix 5

Read processing

Read processing was performed using the PALEOMIX v1.2.5 pipeline (**Schubert et al., 2014**). The procedure included software tools to remove adapters, filter bases based on quality (AdapterRemoval v2.1.7 [**Lindgreen, 2012; Schubert et al., 2016**]), and map reads to the reference mitogenome (Burrows-Wheeler Aligner v0.5.10 [**Li and Durbin, 2009**]). At the time of these analyses, a great auk mitogenome had been published (GenBank: KU158188.1 [**Anmarkrud and Lifjeld, 2017**]), and was thus available for the mapping assembly of our mitogenomes rather than mapping against the composite mitogenome used for bait design (see above).

PCR duplicates were removed using MarkDuplicates within Picard v1.8.2 (**Broad Institute, 2019**) and the rmdup function within SAMtools (**Li et al., 2009**). The Genome Analysis Toolkit (GATK) v3.6.0 was used to correct for misaligned reads to the reference mitogenome using the RealignerTargetCreator and IndelRealigner functions (**McKenna et al., 2010**). Finally, MapDamage2 (**Jónsson et al., 2013**) was employed to rescale base-quality scores according to their probability of being damaged, thereby removing residual aDNA damage patterns. The UnifiedGenotyper algorithm within GATK v3.6.0 was used to determine haploid genotypes for individual samples.

Consensus sequences were produced using the following filtering settings. The per-individual read depth was set to include only bases with a minimum of 3-fold coverage. Bases called for the consensus sequence had to be present at a frequency higher than 33%. To be included in the final alignment, no more than 33% of bases could be missing from the consensus sequence of an individual compared with the reference sequence. All bases failing to meet these criteria were called as 'N' (**Chang et al., 2017**).

Following read processing, the sequences were aligned using Seaview v4.0 (**Gouy et al., 2010**) with the algorithm *Muscle -maxiters2 -diags*. The alignment was manually checked for errors using BioEdit v7.2.5 (**Hall, 1999**). Tablet v1.16.09.06 (**Milne et al., 2013**) was used to view the rescaled Binary Alignment Map (BAM) file for each sample.

Sequence data from all samples included in the analysis have been deposited in GenBank. The GenBank accession numbers for samples included in the final analysis can be found in **Supplementary file 1d**.

Appendix 6

Population dynamics analysis: Settings, partitioning schemes and further details

Six partitioning schemes were compared for the data, varying in the degree of partitioning and the resulting number of data subsets (**Appendix 6—table 1**). For each data subset, the best-fitting model of nucleotide substitution was selected using the Bayesian information criterion in Modelgenerator (**Keane et al., 2006**). A partitioning scheme with six data subsets provided the best fit to the data.

Appendix 6—table 1. Marginal likelihoods of six partitioning schemes and two tree priors for the 25 dated mitogenomes.

| Partitioning scheme ^a | Marginal likelihood ^b | |
|---|----------------------------------|--------------------|
| | Constant size | Exponential growth |
| Unpartitioned | −24,151.6 | −24,143.6 |
| two subsets: (CR rRNA tRNA) (PC1 PC2 PC3) | −24,222.3 | −24,212.4 |
| three subsets: (CR) (rRNA tRNA) (PC1 PC2 PC3) | −24,162.4 | −24,150.1 |
| four subsets: (CR) (rRNA tRNA) (PC1 PC2) (PC3) | −23,659.7 | −23,647.5 |
| five subsets: (CR) (rRNA tRNA) (PC1) (PC2) (PC3) | −23,248.7 | −23,235.9 |
| six subsets: (CR) (rRNA) (tRNA) (PC1) (PC2) (PC3) | −23,229.1 | −23,217.5 |

^aComponents of the mitogenome are the ribosomal RNA genes (rRNA), transfer RNA genes (tRNA), three codon positions of the protein-coding genes (PC1, PC2, and PC3), and the control region (CR). ^bMarginal likelihoods were estimated by stepping-stone sampling with 25 path steps, each with a chain length of 2,000,000 steps.

Constant-size and exponential-growth coalescent tree priors were compared for the data. Analyses using a skyride coalescent prior (**Minin et al., 2008**) were attempted but invariably failed to converge, which strongly suggested overparameterisation. The marginal likelihood was computed for each combination of partitioning scheme and tree prior, using stepping-stone sampling with 25 path samples (**Xie et al., 2011**).

The evolutionary timescale was estimated using a strict clock model, with the sampling times of the mitogenomes serving as calibrations for the clock (**Rambaut et al., 2016**). A uniform prior of $(10^{-10}, 10^{-4})$ was used for the mutation rate, with a separate rate assigned to each subset of the data defined by the partitioning scheme. This approach is consistent with previous analyses of time-structured mitogenomic data sets (e.g. **Anijalg et al., 2018**).

Posterior distributions of parameters were estimated by Markov chain Monte Carlo (MCMC) sampling. Samples were drawn every 5000 steps from a chain with a total length of 50,000,000 steps. The MCMC analysis was run in duplicate to check for convergence and the first 10% of samples were discarded as burn-in. Effective sample sizes of the model parameters were estimated to ensure that they were all over 200, which indicates sufficient sampling.

Appendix 7

TempNet age categories

Age categories were chosen based on changes in climate and hunting pressure. Samples were divided into four groups (**Supplementary file 1e**): >12,000 years old (i.e., Late Pleistocene samples); 1,000–12,000 years old (i.e., Holocene samples when hunting pressure was low and opportunistic); ~500 years old (i.e., the period in which intense hunting began but when diversity should be representative of the previous 12,000 years); and <250 years old (i.e., samples from during the period of intense hunting, including samples from the last reliably seen pair, killed in 1844). For samples with available date information, the median age was used to determine age group. The 16 samples without date information were placed in the most appropriate group based on other information that allowed us to estimate their ages. For example, the samples from Funk Island are unlikely to be over 1000 years old and are most likely to be around 500 years old.

Appendix 8

Population viability analysis: Details and justification (See also Supplementary file 2a, 2b and 2c)

Simulation scenarios were set to run for a 350 year period, as intense hunting began in ~1500 AD (**Bengtson, 1984; Fuller, 1999; Gaskell, 2000; Steenstrup, 1855**) and no confirmed sightings of great auks occurred later than 1852 (**BirdLife International, 2016a; Fuller, 1999; Grieve, 1885**). Data produced in this study show a lack of population structuring in the great auk (see **Figure 3**), and we therefore consider great auks of the North Atlantic to form a single panmictic population. Scenarios were run as a population-based model. Models were also run under scenarios with various definitions of extinction to evaluate any impacts on our results. Extinction was defined as: only one sex remains; population size below the critical limit of 50; or population size below the critical limit of 500. These values are based on the '50:500 rule', which refers to a species' risk of extinction as defined by **Franklin (1980)**.

The outcomes of our simulations were unaffected by the choice of definition used for extinction, which is unsurprising because hunting pressure did not cease towards the extinction of the species. Given this hunting pressure, even 500 birds were well below the sustainable population size, so independent of whether the population size declined to 500 or 50 birds or there was just one sex remaining, the species was bound for extinction. These results might have looked different if our simulations had assumed a complete cessation of hunting when only 500 or only 50 birds were remaining. However, the historical record clearly shows that this was not the case. In fact, as the rarity of the great auk increased, it became more desirable for inclusion in private and institutional collections, as was the case for the last breeding pair killed on Eldey Island in June 1844 (**Bengtson, 1984; Fuller, 1999; Gaskell, 2000; Grieve, 1885; Newton, 1861; Steenstrup, 1855; Thomas et al., 2017**). All results reported are from simulations run under extinction defined as 'only one sex remains'.

Age of first breeding for the great auk is estimated to be 4–7 years old (**Bengtson, 1984**), and a conservative value of 4 years was therefore adopted for the model. The younger the age of first breeding, the less susceptible to extinction the species is. The species is assumed to have been monogamous, laying only one egg per breeding season,, and it is thought they did not replace the egg if it was lost (**Bengtson, 1984; Birkhead, 1993; Fuller, 1999**). An assumed sex ratio of 1:1 has been applied. Life expectancy is estimated to have been 20–25 years (**Bengtson, 1984**) and we assume that breeding remained possible until death. As several alcid species breed annually once they reach sexual maturity (**De Santo and Nelson, 1995**), we set reproductive rate to 100% adult females breeding and all females producing exactly one egg per year.

Mortality rates were estimated based on records from extant alcids. **De Santo and Nelson (1995)** report survival rates for alcid species at various life stages. Mortality at age 0–1 includes hatchlings and fledglings. For the great auk, we estimate mortality to be 9% (SD: 1), consistent with the lowest mortality reported for any alcid species for this age category by **De Santo and Nelson (1995)** (Japanese murrelet, *Synthliboramphus wumizusume*). With regard to the simulation model, juvenile mortality includes mortality in the age groups 1–2, 2–3, and 3–4; therefore, our juvenile mortality rate was divided between these groups. The lowest mortality for this age group reported by **De Santo and Nelson (1995)** is that of the crested auklet (*Aethia cristella*; 34%). This corresponds to approximately 13% (SD: 1) mortality per year over three years, if the population size of the respective previous year is used as reference in each year. Annual adult survival rate is estimated to be quite high for great auks, because of their large size (**Bengtson, 1984; Montevecchi and Kirk, 1996**). Annual adult survival in other alcids is also high, with the razorbill being the highest reported at 93% (**De Santo and Nelson, 1995**). We therefore used an annual mortality rate of 7% (SD:1) for adult great auks. Strictly applying the rule that we use the lowest mortality rate of any alcid species found in the literature leads to some settings that are questionable from a biological perspective. For example, our 0–1 year hatchling mortality is lower than our 1–4 years juvenile mortality. However, as we have no information about actual mortality rates in great auks, any

adjustment of these settings would be arbitrary. We therefore chose to strictly use the lowest mortality rates found in the literature for each age class. For comparison, we added a simulation based on known mortality rates of the razorbill, which have a more biological realistic distribution of mortality rates, albeit perhaps somewhat too high for the great auk (see Discussion and **Supplementary file 2a**).

A reduction of population size, even by harvesting, might have a positive effect on reproductive rate and mortality by freeing up resources and reducing competition. As our reproductive rate was already 100%, a way to simulate such effects was to introduce a linear, density-dependent reduction of mortality rates to half the initial value, following the formula: $(0.5 + (0.5 * PS1)) * [\text{initial mortality rate}]$, with PS1 being defined as initial population size (N) divided by carrying capacity (K). Simulations were run with and without this density-dependent reduction in mortality rates (DD) (**Supplementary file 2b**).

We initially estimated the census size (Nc) for our population viability analyses from our estimated effective female population size (Ne) by doubling the effective female population size and dividing the result by Ne/Nc ratios typical for birds as summarized by **Frankham (1995)**. However, the range of known, typical Ne/Nc ratios for birds extends over two orders of magnitude, from 0.052 to 0.74 (**Frankham, 1995**). Given these ratios, our estimates for the census size of great auks ranged from 12,292 to 756,346. As we did identify a Pleistocene population bottleneck, and given the large population size reported in historic sources, the actual census size was likely close to or even higher than the upper margin of these estimates, and this is consistent with census sizes currently estimated for the great auk's closest relative, the razorbill (*Alca torda*). Within a range similar to that of the great auk, the IUCN Red List estimates that the razorbill (*Alca torda*) currently has a population size of 979,000–1,020,000 mature individuals. Within the same range, the common murre (*Uria aalge*) and the thick-billed murre *Uria lomvia* are estimated to have population sizes of 2,350,000–3,060,000 and 1,920,000–2,840,000 mature individuals, respectively (**BirdLife International, 2016b; BirdLife International, 2016c; BirdLife International, 2017**). Therefore, we conservatively aimed for mature population sizes of 1,000,000 and 3,000,000 great auks.

Razorbills and murrelets can fly and therefore have access to a larger number of breeding sites than the great auk. They are also much smaller birds, which could facilitate larger population sizes in the same range. On the other hand, razorbill and murre populations may be more affected by hunting today than great auk populations were at the time intensive hunting started. Overall, we feel that our population-size estimates are a reasonably realistic reflection of great auk population sizes. We used Vortex 10.2.8.0 to estimate the census size from the number of mature individuals, assuming that birds reach maturity at 4 years of age, that they show a stable age distribution, and that the different age classes follow the fixed mortality rates described below. This resulted in census sizes for our simulations of 2,000,000 and 6,000,000 birds respectively.

To estimate hunting pressure, we compared models in which various proportions of the population were harvested (see **Supplementary file 2c** for example of how harvest rates were calculated). The age categories for harvest rate are 0–1, 1–2, 2–3, 3–4, and over four for both males and females. We allocated 75% of the harvest rate to the over four category as it was assumed that predominantly adult birds were harvested due to being easily accessible when breeding. The remaining 25% was then split evenly between the other four age categories (0–1, 1–2, 2–3, 3–4), as although it has been reported that young were used as bait (**Grieve, 1885**), it is unlikely they were harvested at the same intensity as the adults and represented a smaller proportion of the overall population.

As we know eggs were collected as well, we allowed for this in the model. The harvest rate for eggs was set at 5%, corresponding to 25,688 and 77,065 respectively for the two initial population sizes tested in our simulations. As great auks nested in dense groups (**Bengtson, 1984**) eggs would have been easy to collect. Based on estimates of breeding pairs at Funk Island (>100,000) (**Birkhead, 1993**), these two values allowed us to test the impact of a quarter or three quarters of all the eggs laid annually on Funk Island alone being harvested, with no harvest occurring anywhere else in the great auk range. We also ran simulations with no egg harvesting to evaluate whether this significantly changed our conclusions. With these

egg harvest settings, an annual bird harvest rate of 10% of the number of birds in the pre-hunting population was identified as critical limit, with significant numbers of simulations leading to extinction. At 10.5% bird harvest rate, all simulations that included egg harvesting and a significant proportion of simulations excluding egg harvesting resulted in extinction.

Our comparative simulations with more 'realistic' rather than conservative settings, including razorbill mortality rates were conducted under the settings outlined in **Supplementary file 2a**.

Appendix 9

Nuclear SNP data

As the results of our mitochondrial genome revealed a lack of population genetic structure and high genetic diversity in the great auk, we attempted to target nuclear DNA (nuDNA) to further investigate these results, and to obtain a more detailed picture of great auk evolution and extinction. Initially, twelve samples were chosen for capture of 495 nuclear markers.

Samples were chosen based on the percentage of reads retained in preliminary mitogenome capture dataset, as a rough indication for sample preservation and quality, as well as their geographical location to represent individuals from as much of the former distribution as possible. DNA extraction and library preparation methods were as described for the mitogenome work (see Materials and methods main text). Great auk shotgun genome data (Gilbert et al., bait design available see **Source data 1**) mapped against the razorbill genome (Feng et al. In Review) were used as data basis for bait design. Target gene regions for hybridisation capture enrichment were selected using the following filters:

Paralog genes were excluded from the capture by using UniProtIDs and EnsemblIDs in the razorbill (*Alca torda*) annotation (Feng et al. In Review).

Genes that were missing coverage for more than 20% of their length when mapping great auk reads against the razorbill genome were excluded.

Great auk consensus genes were generated by replacing the razorbill genes with the homozygous SNPs found in great auk.

Genes with the highest percentage divergence between the razorbill and great auk, that didn't contain any N's in their sequence, and which were less than 5kbp in length, were used to build the 20K probes resulting in 495 genes.

MYcroarray probes of 120 bps long with 3x tiling (40 bps shifts) were made from CDS regions and intron regions that were adjacent to the exons of the 495 genes. Enrichment for nuclear genes was performed using MYcroarray MYbaits, following the MYcroarray Mybaits manual v3 (MYcroarray/**MYcroarray MYbaits, 2016**), using 24 hr hybridisation time, at 65°C and final elution into 30 µl nuclease free water. Samples were sequenced on an Illumina MiSeqPE75 platform by New Zealand Genomics Limited, Otago.

Sequencing reads were processed using the PALEOMIX v1.2.5 pipeline (**Schubert et al., 2014**) following a procedure similar to that described by the authors. Briefly, we used AdapterRemoval v2.1.17 (**Schubert et al., 2016**) to trim the reads for adapters and low quality bases (BaseQ <5 or Ns), and to exclude those reads shorter than 30 bp or with more than 50 bp of missing data. Filtered reads from each sample were mapped against the razorbill reference genome (Gilbert, unpublished) using BWA-MEM v0.7.12 (**Li, 2013**), and those with low mapping quality (MapQ <15) removed. After the initial alignment step, Picard (v1.128, <https://broadinstitute.github.io/picard>) was used to exclude reads that were PCR or optical duplicates. Subsequently, GATK v3.5.0 (**McKenna et al., 2010**) was used to perform a realignment step around indels. As we are dealing with historical samples, we also quantified the extent of DNA damage in our samples using mapDamage v2.0.6 (**Jónsson et al., 2013**). We characterised rates of deamination in double strands (DeltaD) and single strands (DeltaS), as well as the probability of reads not terminating in overhangs (Lambda, transformed into 1/Lambda - 1, a proxy for the overhang length of overhanging regions). From these analyses, we also rescaled base quality scores according to the probability of each base being affected by post-mortem damage.

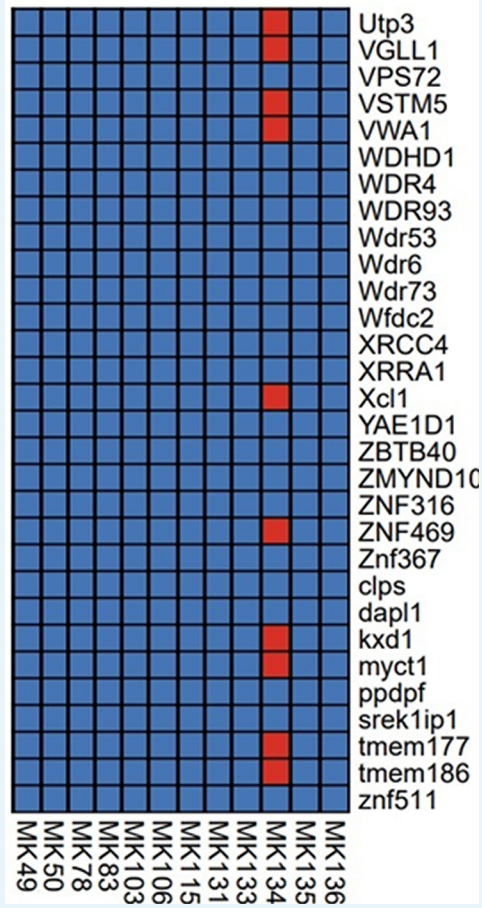
Read processing of the twelve samples initially sequenced revealed low coverage of both the 495 targeted markers (0.0018x MK78 - 1.2592x MK134), and the razorbill genome overall (0.00006x MK83 - 0.0190x MK50) (**Appendix 9—table 1** and **Appendix 9—table 2**). Only one sample, MK134, had any genes with at least 3-fold coverage (**Appendix 9—figure 1**). Therefore, further analysis that would provide any meaningful results could not be performed.

Appendix 9—table 1. Estimated coverage information from the twelve sequenced samples. The estimated coverage of the 495 targeted genes and estimated coverage of the reads that mapped to the razorbill genome is reported.

| Sample | Country | Estimated coverage of razorbill genome | Estimated coverage of targeted genes |
|--------|--------------|--|--------------------------------------|
| MK49 | Norway | 0.0101 | 0.0152 |
| MK50 | Iceland | 0.0190 | 0.0155 |
| MK78 | Funk Island | 0.0022 | 0.0018 |
| MK83 | Funk Island | 0.00006 | 0.0071 |
| MK103 | Funk Island | 0.0011 | 0.0150 |
| MK106 | Sweden | 0.0172 | 0.0105 |
| MK115 | Norway | 0.0012 | 0.0021 |
| MK131 | Iceland | 0.0090 | 0.0423 |
| MK133 | Skin Mystery | 0.0190 | 0.0154 |
| MK134 | Skin Mystery | 0.0179 | 1.2592 |
| MK135 | Skin Mystery | 0.0073 | 0.0106 |
| MK136 | Skin Mystery | 0.0021 | 0.0128 |

Appendix 9—table 2. Coverage range of captured markers. Numbers in square brackets represent the number of markers which have 0 coverage. Genes with the highest coverage are shown in brackets.

| Sample | Country | Coverage range of captured markers |
|--------|--------------|------------------------------------|
| MK49 | Norway | 0 [125] – 0.4898 (Fam174b) |
| MK50 | Iceland | 0 [157] – 0.2204 (Isca2) |
| MK78 | Funk Island | 0 [379] – 0.1087 (Mrp130) |
| MK83 | Funk Island | 0 [223] – 0.2960 (Nipbl) |
| MK103 | Funk Island | 0 [164] – 0.7049 (Glr5) |
| MK106 | Sweden | 0 [190] – 0.2403 (Pcp4) |
| MK115 | Norway | 0 [366] – 0.2263 (Tmem60) |
| MK131 | Iceland | 0[78] – 1.5238 (Ssna1) |
| MK133 | Skin mystery | 0 [129] – 0.3061 (Fam174b) |
| MK134 | Skin mystery | 0.0628 (TPK1) – 17.7232 (Ssna1) |
| MK135 | Skin mystery | 0 [172] – 0.2580 (myct1) |
| MK136 | Skin mystery | 0 [142] – 0.4067 (myct1) |

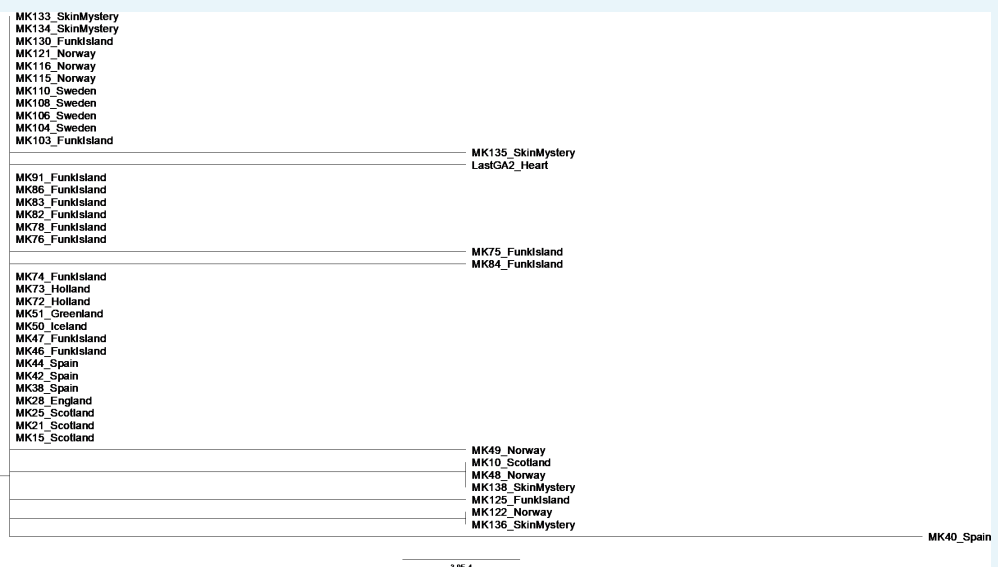


Appendix 9—figure 1. Section of the presence/absence matrix showing coverage of 30/495 captured genes (listed on the right-hand side) for each sample sent for sequencing. Presence is defined as coverage ≥ 3 , indicated by a red square, absence is indicated by a blue square.

Appendix 10

Additional analyses Spanish samples

In the phylogenetic tree of Great Auks that yielded sufficient sequence data as per our filtering criteria (Appendix 5) the sample MK40_Spain appeared to be differentiated from the rest of the samples which came from the northern regions of their distribution. This raises the question whether the sample could represent a refugial population in Spain. In order to test this, we re-examined the phylogenetic relationships between samples with the addition of the other Spanish samples we sequenced but which did not fulfil the filtering criteria for inclusion in our final dataset. These samples included MK37, MK42, MK44 and MK45. These samples were characterised by poor coverage (average coverage ranged from 0.18 to 2.07) and over 33% of bases missing from consensus sequence (consensus sequence length ranged from 36 bp to 5468 bp). Sequences generated as described using the Paleomix pipeline (Appendix 5) for samples MK37, MK42, MK44 and MK45 were manually aligned to the reference genome using Bioedit v7.2.5 (Hall, 1999) and Tablet v1.16.09.06 (Milne et al., 2013) to view the rescaled Binary Alignment Map (BAM). As MK37 and MK45 were of very poor quality, we were unable to use them in this additional analysis. However, we were able to produce an alignment of 859 bp that included the additional Spanish samples MK42 and MK44. A Neighbour-joining analysis of the alignment based on p-distances (Saitou and Nei, 1987) using MEGAX (Kumar et al., 2018) yielded a very poorly resolved phylogeny (Appendix 10—figure 1). Critically, the new Spanish samples do not group with the outlier MK40, thereby not supporting a hypothesis of a Spanish refugial population.



Appendix 10—figure 1. Phylogenetic tree showing the relationship between all samples that passed filtering criteria, plus additional Spanish samples previously excluded from analysis. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987) in MEGAX (Kumar et al., 2018). The optimal tree with the sum of branch length = 0.01164144 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein, 1985). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method (Nei and Kumar, 2000) and are in the units of the number of base differences per site. This analysis involved 43 nucleotide sequences. All positions containing gaps and missing data were eliminated (complete deletion option). There were a

total of 859 positions in the final dataset. Tip labels give the sample names and sampling locations.