Statistical Methods for Identifying Demographic Structure in DNA Sequence Alignments

Adam Benjamin Rohrlach

Thesis submitted for the degree of

Doctor of Philosophy

in

Applied Mathematics

at

The University of Adelaide

(Faculty of Engineering, Computer and Mathematical Sciences)

School of Mathematical Sciences

January 15, 2019

Contents

Abstract

All life on Earth, from viruses and bacteria, trees and flowers, to birds and human beings, can be traced back to a single common ancestor. However, the evolutionary history that led to this diversity of life is a complicated story that we do not yet fully understand.

Since the discovery of the structure of deoxyribonucleic acid (DNA) in 1953, and the development of DNA sequencing technology, researchers have been using similarities and differences in the genomes of organisms to better understand the relationships between species. However, due to the complexity of the evolutionary history of life, simplifying assumptions must be made to make mathematical models tractable. It must then be of paramount importance for researchers to be able to identify when the simplifying assumptions of a specific model are unreasonable.

In this thesis we present two projects, and although they are different in implementation, both attempt to investigate simplifying assumptions in the closely related fields of population genetics and phylogenetics. However, we also present applications of our projects where the results of our work are not used in assessing assumptions for further analyses, but are of standalone interest to researchers.

Our first project is concerned with the development of a method for constructing coordinate representations for single-copy DNA, such as mitochondrial DNA (mtDNA) or Y-chromosomal DNA, analogous to the use of PCA for nuclear DNA. We construct a coordinate system such that, given p informative sites in an alignment of n individuals, returns p-dimensional coordinates for each n individuals. We order the

dimensions by the proportion of variability each dimension captures in the overall genetic diversity.

From these coordinates in "genetic space" researchers may perform a number of down stream analyses. It is possible to optimally visualise high-dimensional sequence data in two or three dimensions. One may use our method to identify closely related individuals, identify sites in the alignment that are closely linked, or to use the same coordinate space to find sites that are closely linked with groups of individuals. Finally, one may choose to test for significant relationships between the structure of the coordinates in genetic space, and metadata recorded on sequenced individuals, indicating demographic variables that are highly related to the evolutionary history of an alignment.

This final application of our method, where one may test for demographic structure in sequence data, is of key importance to the theme of discovering when simplifying assumptions of analyses are not reasonable. Through the comparison of coordinates in gene space, and *any* demographic variables of interest, researchers may explore whether or not the individuals in the alignment indicate population substructure. For example, one may investigate if there appears to be a phylogeographic structure to the individuals forming distinct subpopulations, and if migration appears to occur between subpopulations.

Through empirical data, we show that our method can readily recover tree-like structure, identify strong genetic groupings based on qualitative traits and show that we are able to recover phylogeographic signal given provenanced sampling information. We show that our method can even be used to suggest routes of migration based on mtDNA. Finally we apply our method to modern Aboriginal Australian mtDNA to show strong evidence for discrete geographic populations of Aboriginal Australian peoples that display permanence on the Australian landscape dating back to the original colonisation of Australia 50 thousand years before present (kya).

Our second project is concerned with identifying departures from a tree-like evolu-

tionary history at the species level. It is not uncommon for closely related species (Species A and C say) to still be capable of interbreeding, and producing viable "hybrid" offspring (Species B say). Under these conditions, a phylogenetic *tree* cannot describe the evolutionary history of the hybrid species, and instead an admixture graph may be a better description.

We begin by considering the evolutionary history of three species: a hybrid organism that has undergone some independent evolution (Species B), and two "parent" organisms, Species A and C. Relatively long, contiguous regions of the genome of Species B will have undergone no recombination since the admixture event. These regions will have been contributed by either Species A (and hence will be more closely related to Species A), or Species C. We aim to estimate the proportion of the genome contributed by Species A, and denote this γ by considering the proportion of informative site patterns that indicate evidence for the two possible ancestries.

The mixing proportion is the parameter of interest in our analyses. However, due to the classical problem of the non-identifiability of mixing parameters in multinomial distributions, we describe two Bayesian methods for estimating γ. Our first method places prior distributions on the parameters of the model, and uses Approximate Bayesian Computation (ABC) to estimate the marginal posterior distribution of γ. Our second, closely related method, instead estimates the marginal posterior distribution of γ via numerical integration.

We show via a simulation study that our methods can accurately estimate the true value of γ, and perform well under biologically reasonable scenarios. However, we also find that our methods suffer from a relatively small positive bias for small values of γ, *i.e.*, when one species of the parent species contributes very little to the genome of the hybrid species. We compare the performance of our method to the popular method of the ratio of f_4 statistics. We do this by estimating the proportion of Neanderthal ancestry in pre-ice age European human samples and comparing our results to the finding of Fu *et al.* [18]. We show that our method recovers extremely

similar estimates of Neanderthal ancestry with no apparent systematic bias when compared to the results of Fu *et al.*.

Finally we apply our method to the genomes of Late Pleistocene European bison (*Bison bonasus*) and Steppe Bison (*Bison priscus*) to understand the evolutionary history of bovid megafauna in Europe over the last seventy thousand years. It was thought that before 10 kya the only bovid present in Europe was the Steppe bison. However, from bone samples found dating from the present day, and back to approximately 70 kya, mtDNA indicated a second bison species was also roaming Europe before 10 kya, more closely related to modern cattle than the Steppe bison.

After nuclear DNA was sequenced, we were able to show that this new species of bovid was actually a hybrid offspring of Aurochs (the ancestor of modern cattle) and Steppe bison, an event that occurred approximately 120 kya. We used our method, in concert with the ratio of f_4 statistics, to show that the hybrid species contained approximately 10% Aurochs and 90% Steppe bison ancestry.

Signed Statement

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

SIGNED: DATE: 05/11/2018

viii

Dedication

I dedicate this thesis to my family, as well as my friends and supervisors; a set for which the union and intersection look suspiciously similar.

Acknowledgements

First, to my supervisors Nigel Bean, Jono Tuke and Barbara Holland. It seems to me that this whole thing should have seemed like less fun, and more work. Thank you for your guidance and helping me achieve more than I believed possible. Nigel, there is no doubt in my mind that you are the type of academic I wish to be. At this stage, I would settle for "half as good", or even "epsilon as good". Jono, thank you for being my mentor for so long. You instilled in me the importance of hard work, but never at the expense of loving what you do. Your enthusiasm for statistics is the reason I got so excited about data science in the first place, and it is as infectious as it is persistent. Barbara, thank you for inspiring me to take up a field of research that fills my every working day with wonder. It was you all those years ago that made me decide to take on a field for which I had little training, but a deep interest. Thank you for making me feel so welcome every time I visited you in Hobart, and apologies for always suggesting *one* more stout.

To my father Brian and step-mother Kerry, thank you for your unwavering support these long years. Thank you for *gently* encouraging me to complete my PhD, and move on to the next phase of my life. I do not think I could have aimed so high, and then seen it through, without your sage advice and the benefit of your experience.

To my mother Wendy, thank you for your unconditional love. Thank you for being there during the good and bad times. One could not ask for a better mother. You have ingrained in your children a belief that we can achieve anything, which can be the difference between success and failure.

To my sisters Paige and Prue, thank you for your persistent belief in me. Although I may not always show it, some of my favourite times are family meals with you all. Even the kids.

To my friends Dan and Brett. Thank you for finding a way to let me stop thinking about my work at every moment. You know when I need to take a break from my work. You have kept me sane during the most stressful period of my life, and for that I can never repay you. Nevertheless, the first round is on me.

To Alan Cooper, Graham Gower, Ray Tobler, Bastien Llamas, Yassine Souilmi, Maria Lekis and everyone at the Australian Centre for Ancient DNA, thank you for making me feel welcome in your research group. If you had not included me in your meetings, I would not have been able to work on so many interesting projects, with so many interesting people. I have no idea why you allowed me in, but good luck dislodging me.

To everyone in the School of Mathematical Sciences at the University of Adelaide, thank you for such an enjoyable time, from the start of my undergraduate degree, to the end of my PhD study. This is an amazing place to come to work everyday due to the people that fill its offices. However, I would like to mention a few special people. David Price, thank you for being a friend and a colleague. Your advice, both statistical and personal, was always welcomed, and always useful. Mingmei Teo, you may be the hardest working individual I have ever met. Our morning coffees are sorely missed, as is the conversation that came with them. Nic Rebuli, thank you for making work fun. Your company, and almost sitcom-like entries into my office, never ceased to make me laugh. Stephen Crotty, thank you for the laughs. I am suspicious that had we gone ahead with those joint PhD meetings, we would have achieved nothing, and probably been asked to leave the premises due to noise complaints. Our conference trips were filled with classic moments, and I shall never again underestimate the supernatural defensive capabilities of the oft overlooked suitcase rack.

Finally, special thanks must go to Gary Glonek. Though you were never officially one of my supervisors, your input into almost every statistical problem I tackled made my work possible. I was particularly impressed with your patience and calm demeanour when you would slowly explain a simple concept to me. Especially on the twelfth attempt. Your guidance, both professionally and academically, will have a lasting effect on my career until the day I decide to hang up my boots.

Chapter 1

Introduction

Evolutionary biology is the study of the processes acting on populations of organisms that have led to the diversity of life on Earth. Within evolutionary biology are two closely related, and mathematically and statistically rich, sub-fields: phylogenetics and population genetics. Population genetics is the study of the changes in the frequencies of types of individuals in populations due to natural selection, mutation, genetic drift and gene flow. Phylogenetics is the study of the evolutionary relationships between individuals, or groups of organisms, such as populations or species.

The processes describing the evolutionary history of even simple, well-defined populations can be extremely complex, and simplifying assumptions must be made such that mathematical models describing these processes are tractable. For example, the process by which DNA accumulates substitutions over time is almost universally modelled using reversible, continuous-time Markov models. This is a clear oversimplification of the true underlying process, but an extremely useful tool for analysing genetic data [17].

In all statistical and mathematical models, it is of key importance to address whether or not the assumptions of the model are reasonable. Methods for the analysis of genetic data have, in some cases, formal tests for modelling assumptions [33, 37,

25]. In some cases, informal methods exist to identify departures from modelling assumptions [39, 51]. However, in many cases no such statistical tests, formal or informal, exist for analyses. In this work we look at the problem of identifying when the demographic history of a population (or populations) of individuals is sufficiently complex that simple models for the analysis of genetic data are no longer reasonable. We give two examples, one each in the sub-fields of population genetics and phylogenetics.

1.1 Motivation

A particularly important modelling tool used in population genetics and phylogenetics to separate and describe the effects of genetic drift and mutation on the allelic frequencies for a population of organism is the coalescent model [23]. The simplest form of the coalescent model assumes no recombination, no natural selection and no population structure or gene flow. However, once these assumptions have been deemed to be unreasonable, the coalescent model may be readily modified by a simple rescaling of time to incorporate these departures from the simple case [49, 37].

The adaptability of the coalescent model to a large number of biological scenarios means that the coalescent model can be used as the underlying model to reconstruct species phylogenies, efficiently simulate sequence data, and estimate demographic parameters such as population size, migration rates and recombination rates [5, 7, 24].

Hence, it is important for researchers to be able to identify, in some cases from sequence data alone, when the simplifying assumptions of the coalescent model appear unreasonable, such that they may employ a better model of the sequence evolution of their sample. For this reason unsupervised learning methods, such as principal components analysis (PCA), for the exploration of sequence data have been an extremely popular tool in modern genetic analyses involving nuclear DNA [29]. However, to our knowledge, no such analogous method of unsupervised exploration exists for single-

copy DNA, such as mtDNA and Y-chromosomal DNA. This is addressed, and the resulting method explained, in Chapters 2 and 3 of this thesis.

In maximum likelihood reconstruction of phylogenetic trees the simplifying assumption that the evolutionary history of the sequenced organisms can be adequately described by a *tree* is employed. While this assumption is clearly reasonable for non-recombining DNA, such as mitochondrial (mtDNA), it may not be reasonable for recombining DNA, such as nuclear DNA, over relatively short periods of evolutionary time due to the effects of linkage disequilibrium [41]. Further, even over longer periods of evolutionary time, populations of organisms that were once separated and unable to interbreed may be reintroduced, and begin to create admixed offspring [18, 46].

Thus, it is important to be able to identify that the evolutionary history of a collection of samples will not be adequately described by a single tree. The so-called D-statistic is one method that identifies departures from a tree-like evolutionary history [34]. If a tree-like evolutionary history is deemed unreasonable, researchers may look to fit a mixture model of underlying trees, a so-called admixture graph. Alternatively, researchers may wish to forgo reconstructing the parameter-rich admixture graph, and be more interested in estimating the ancestry proportions for a specific species [32, 18, 20]. This topic is addressed in Chapters 4 and 5 of this thesis.

Note that we motivate the work presented in this thesis under a unifying theme of identifying departures from simplifying assumptions for population genetics and phylogenetics. However, in many cases researchers may already know that admixture has occurred, and may simply be interested in quantifying the proportions of ancestry for an organism. Similarly, researchers may well know that some population demographic substructure exists, and simply wish to find a low-dimensional coordinate representation of the sequence data for confirmation, or visualisation, of the demographic substructure.

1.2 Thesis Structure

This thesis describes two mathematical and statistical projects within the field of evolutionary biology, and together these projects comprise Chapters 2 through 5. Broadly speaking, Chapters 2 and 3 are concerned with the problem of the unsupervised detection of population structure in single-copy DNA alignments, although this method can be extended to include any type of sequence data. Chapters 4 and 5 are concerned with the problem of detecting and quantifying admixture, a clear departure from a tree-like evolutionary history. We narrow our focus to consider situations where we have two identifiably different populations of individuals, that are reintroduced and produce viable offspring.

In Chapter 2 we fully describe and develop the underlying mathematics used for the spectral decomposition of the qualitative sequence data into a continuous coordinate representation. We then develop the method further to include projecting metadata and new sequences into the same coordinate space, followed by descriptions of formal statistical tests to identify and quantify relationships with demographic variables of interest. Through the use of a toy example, and previously published sequence alignments, we then show how we may use the method to visualise alignment data in as little as two dimensions, and identify variables of interest under several biological scenarios. This work is currently under review for publication.

In Chapter 3 we present work, for which I was a joint first author, that was published in *Nature* on the 8th of March, 2017. In this publication my contribution was an application of the theory developed in Chapter 2, applied directly to a unique data set containing aboriginal mtDNA for which we also had reliable pre-European provenance. We showed that the Aboriginal peoples entered northern Australia approximately 49 thousand years ago (kya), and rapidly migrated along the east and west coasts of Australia, and settled to form strong regional patterns that persist to this day.

For this publication the biological lab work was performed entirely by our collaborators at the Australian Centre for Ancient DNA, for which we take no credit. Our contribution was the complete development of the statistical framework for the analysis, and software implementation of the method. Our method was used to show a strong relationship between the continuous coordinate representation obtained from our spectral decomposition method (genetic space), and the demographic variables, longitude and latitude. We also showed a strong relationship between the distances calculated in genetic space, and distance calculated from geographical information. The analyses we were able to perform via the method we had developed allowed us to find significant evidence for the continuous presence of populations in discrete geographical areas. This presence was shown to date back to the initial peopling of Australia, agreeing with Aboriginal Australian cultural attachment to their country.

In Chapter 4 we develop a rigorous statistical method for modelling the distribution of site patterns which are informative for the underlying topology of a three taxon tree. We extend this method to include estimating the contribution of the two major bifurcating topologies to an admixture graph. To avoid the classical issue of non-identifiability for parameter estimation in mixture models for multinomial distributions, we provide two methods for investigating the marginal distribution of the mixing parameter when considering constraints on the branch length parameters. We first describe a method via Approximate Bayesian Computation (ABC), and then a closely related method using a Dirichlet prior assumption for the probabilities of the multinomial distribution, and find the marginal distribution of the mixing parameter via numerical integration.

Our method compares favourably in terms of complexity to the popular ratio of f_4 statistics due to the reduced number of parameters in our model, although our method is unable to incorporate incomplete lineage sorting [34]. We use a simulation study to assess the performance of our methods, and apply our method to previously published sequence data to estimate the proportion of Neanderthal (*Homo nean-*

derthalensis) ancestry in pre-ice age European human (*Homo sapiens*) individuals. We directly compare the results of our method to those obtained via the ratio of f_4 statistics. This work is presented as a traditional chapter and will be considered for publication later.

In Chapter 5 we present work that was published in *Nature Communications* on the 18th of October, 2016. In this publication we consider the European bison wisent (*Bison bonasus*), prior to the holocene (11.7 kya). We used complete ancient mitochondrial genomes and genome-wide nuclear DNA surveys to reveal that the wisent is the product of hybridisation between the extinct steppe bison (*Bison priscus*) and ancestors of modern cattle (aurochs, *Bos primigenius*) before 120 kya, and contains up to 10% aurochs genomic ancestry.

For this publication the biological lab work was performed entirely by our collaborators at the Australian Centre for Ancient DNA, for which we take no credit. Our contribution was the complete development of the statistical framework for the analysis, and software implementation of the method. Our method was used in parallel with f_4 statistics and the so-called 'ABBA-BABA' test, and since the conclusions were very similar, was used to strengthen the findings of both analyses of ancestry proportions [28].

In Chapter 6 we conclude our findings, and discuss possible extensions to the work presented in this thesis.

Chapter 2

Unsupervized Quantification of Demographic Structure for Single-copy Alignments

2.1 Introduction

In this chapter we present the paper titled "Unsupervized Detection of Demographic Structure for Single-copy Alignments". This paper is currently under review for publication.

The purpose of this work was to define a rigorous mathematical framework for the unsupervised exploration of single-copy DNA. This work was born of the desire to find an informative, low-dimensional, continuous-coordinate, representation for mitochondrial DNA (mtDNA) analogous to the use of principal components analysis for nuclear DNA. In principle, our method may be used to explore pseudo-haploid DNA, microsatellite data, and even nuclear DNA without the need to filter triallelic or quadrallelic sites, although this remains future work.

We begin by defining the transformation of aligned sequence data to a contingency

table of the same form used in analyses of linkage disequilibrium for recombining DNA. We then describe the singular value decomposition of the contingency table, and the scaling factors to calculate the row scores (coordinates for individuals in the alignment) and the column scores (coordinates for the sites in the genome used in the analysis). We also motivate our choice of scaling factor for the column scores and prove that our choice yields coordinates such that the row and column scores are directly comparable on the same space.

We then describe the method by which we may project new sequences and metadata into the space defined by previously defined analysis. We suggest possible tests for detecting significant relationships between demographic variables, and the genomic samples in the coordinate-space found using our method.

Using a custom data set of previously published human mtDNA, we show that the results obtained via our method yields meaningful results, consistent with the known relationships between human mitochondrial haplogroups. We also show that we may sensibly observe the complex relationships between these individuals, based on 281 single nucleotide polymorphisms (SNPs), in as few as two dimensions.

Finally, we analyse two further previously published data sets. First we show that the extinct thylacine (*Thylacinus cynocephalus*) formed genetically distinct groups in Tasmania, and Eastern to Western Australia. We then use the results of our method to identify a potential migration route from East to West Australia. Second we show that the distribution of ghost bat (*Macroderma gigas*) genetic variability is extremely well explained by the caves in which they roost, indicating that the ghost bat is particularly vulnerable to displacement by blast mining.

2.2 Statement of Authorship

Statement of Authorship

Title of Paper	Unsupervized Quantification of Demographic Structure for Single-copy Alignments
Publication Status	Under Review
Publication Details	Adam B Rohrlach, Nigel Bean, Gary Glonek, Barbara Holland, Raymond Tobler, Jonathan Tuke, and Alan Cooper. Unsupervized quantification of demographic structure for single-copy alignments. bioRxiv, page 338442, 2018.

Principal Authors

Adam Rohrlach (Candidate)			
Contribution to the Paper	Developed method, wrote code implementation of the method, designed analyses, performed analyses, wrote paper with the help of other authors.		
Overall percentage (%)	80		
Signature		Date	19/10/18

Co-Author Contributions

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

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

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

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

Name of Co-Author	Nigel Bean		
Contribution to the Paper	Helped to develop method. Wrote the paper with help from all co-authors.		
Signature		Date	18/10/2018

Name of Co-Author	Jono Tuke		
Contribution to the Paper	Helped to develop method. Wrote the paper with help from all co-authors.		
Signature		Date	18/10/2018

Name of Co-Author	Gary Glonek		
Contribution to the Paper	Key to developing proof in appendix. Wrote the paper with help from all co-authors.		
Signature		Date	19/10/18

Name of Co-Author	Barbara Holland		
Contribution to the Paper	Helped to develop method. Wrote the paper with help from all co-authors.		
Signature		Date	19/10/18

Name of Co-Author	Ray Tobler		
Contribution to the Paper	Helped to design experiments. Wrote the paper with help from all co-authors.		
Signature		Date	18/10/18

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Helped to design experiments. Wrote the paper with help from all co-authors.		
Signature		Date	18/10/18

Unsupervized Quantification of Demographic Structure for Single-copy Alignments.

AB Rohrlach[*,1,2],**Nigel Bean**,[1,2], **Gary Glonek**[1], **Barbara Holland**[3], **Ray Tobler**[4], **Jonathan Tuke**[1,2], **Alan Cooper**[4]
[1]School of Mathematical Sciences, University of Adelaide, Adelaide,South Australia, 5005, Australia.
[2]ARC Centre of Excellence for Mathematical and Statistical Frontiers, University of Adelaide, Adelaide, South Australia, 5005, Australia.
[3]School of Natural Sciences (Mathematics), University of Tasmania, Hobart, Tasmania 7001, Australia.
[4]Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia, 5005, Australia.

Abstract

Single-copy sequence alignments have been a valuable source of information for genetic studies; their lack of recombination makes phylogenetic analyses tractable [1]. Specifically, mitochondrial DNA will continue to play an important role in genetic studies due to its high mutation rate and high copy per cell count of the molecule [2]. In this paper we develop a new method for the analysis of single-copy sequence data that simultaneously considers the relationships between sequenced individuals and positions of interest in the genome. We then show that tests for relationships between genetic information and qualitative and quantitative characteristics can be calculated. We motivate the use of our method with examples from empirical data.

1 Introduction

An important feature of any genetic analysis can be detecting whether a sample comes from a structured population. Demographic structure can take many forms. For example, samples may be taken from geographically isolated subpopulations [3], from subpopulations along a migration route [4] or from temporally separated population replacement events [5]. In some cases it can be of interest to discover that no geographic structure exists at all, leading to the exploration of social structure [6].

A popular form of unsupervized data exploration is principal components analysis (PCA) [7]. PCA is a dimension reduction technique that takes n p-dimensional vectors and, using linear combinations of the original vectors, finds $min(n-1, p)$ p-dimensional basis vectors. The new vectors are ordered by the amount of variability explained by each 'principal dimension'. Often the first few dimensions are used to visualise points in the new transformed space.

PCA is a non-parametric, hypothesis-free exploratory technique, making it a particularly attractive analytical tool. However, PCA does require that the vectors of information are quantitative variables. Clearly sequence characters are not quantitative random variables, and so a transformation must be applied to raw sequence data before PCA can be directly applied [8]. However, we are aware of no such suitable transformation for DNA sequences that are non-biallelic, and in particular, haploid DNA such as

mitochondrial DNA (mtDNA) or Y chromosome sequences. Instead, we suggest the application of multiple correspondence analysis (MCA) directly to the sequence characters.

MCA is an adaptation of PCA where categorical variables (in this case Single Nucleotide Polymorphisms: SNPs) are converted into binary variables denoting the presence or absence of each level of the variables (in this case alleles) [9]. Unlike PCA, MCA can be applied to any number of alleles. Our method makes the assumption that SNP inheritance is random, *i.e.* that the underlying phylogenetic tree is a star tree. One could test whether alleles appear to occur independently by investigating a contingency table of pairwise allele counts for an alignment, and then apply a chi-squared test. Since one would almost always overwhelmingly reject the null hypothesis, the result of a chi-squared test would be of no interest. However, the matrix of signed residuals under this assumption form the basis of the transformation from sequence data to continuous data.

In this paper we aim to show that MCA is a statistically powerful method for the analysis of non-autosomal DNA. We show that MCA has many properties that are analogous to PCA, and is hence immediately intuitive to researchers with experience using PCA. We demonstrate that PCA only quantifies the relationships between rows (individuals), while MCA quantifies the relationships between the rows (individuals) and also the columns (SNPs) simultaneously. For this reason we can quantify and visualise relationships between individuals as in PCA, and also quantify and visualise the relationship between SNPs, and between SNPs and individuals simultaneously in the same dimensions.

We show that results obtained from MCA correspond to the results obtained from mtDNA phylogenetic trees in a meaningful way, and that demographic structure can be detected using these results. We explore an alignment of African mtDNA from haplogroups L0, L1, L2, L4 and L5 to show that our method produces valid and easily interpretable results by reproducing mtDNA macro-haplogroups via clustering. We also explore an alignment of modern and ancient thylacine mtDNA from a mainland and island population. We show that thylacine genetic signals are highly correlated with longitude, and identify a possible ancestral migration route. Finally we explore an alignment of Western Australian Ghost Bat mtDNA to show that genetic diversity can be almost completely explained by discrete cave locations.

2 New Approaches

MCA is a generalised form of correspondence analysis that can be seen as the counterpart of PCA for categorical data analysis. Utilizing this powerful unsupervised data exploration method for genetic data yields a number of useful results and techniques that PCA does not allow.

First, the method may be applied directly to alignment data, forgoing the need for a transformation of sequence data to normalised allele frequency counts which inherently assumes data comes from a population at Hardy-Weinberg equilibrium [8]. Second, the method is able to calculate coordinates for individuals in genetic space (as in PCA) but can also simultaneously calculate coordinates for genetic markers (such as SNPs). Here we derive a multi-dimensional coordinate space scaling that allows the coordinates of individuals and SNPs to be directly visualized, and for demographic structure to be explored in both spaces simultaneously. Finally we define methods for exploring supplementary data in the case of both continuous and discrete variables, and show that the results of MCA can be used to identify SNPs of interest, leading to the detection of diagnostic SNPs, or potentially selective markers.

3 Results

Coordinates in Gene Space and Dissimilarity Matrices for Haplotype Identification

The L-haplogroups represent the earliest evolution in modern human history, with the most recent common ancestor (MRCA) of the L-haplogroups being the MRCA of all humans. Hence, our method should be able to recover structure in the form of clusters

of the major haplogroups L0, L1, L2, L4 and L5. To test this, we analysed a custom alignment from several published studies involving African sampled mtDNA [10, 11, 12, 13, 14]. We randomly chose sequences from these studies from sub-haplogroups L0d, L0k, L1c, L2a, L4 and L5a. We aimed to include 20 samples per haplogroup, although we included only 9 from L5a, as this was all that was available at the time of writing, and 10 from L4 to deliberately introduce further sampling asymmetry, resulting in an alignment of 79 individuals (see Table S1 for the file list of Genbank accession numbers and haplotype assignments).

We aligned our sequences to the revised Cambridge Reference Sequence [15] using `MAFFT v7.310` [16]. Haplogroups were determined using `Haplogrep v2.1.0` [17]. Aligned sequences were filtered to remove any homogeneous sites. MCA was performed on the remaining 281 SNPs. The first two principal dimensions captured 50.93% of the total inertia. That is, 50.93% of the variability in the 16,569 dimensional space (the number of base pairs in the sequences) can be observed in the first two principal dimensions.

We reconstructed a phylogenetic tree to compare the topology with our results. A Tamura-Nei model, with invariant sites and a gamma distribution with five classes was selected as the best model of sequence evolution using `ModelGenerator v0.85` [18]. We used `Beast v1.8.3` [19] to construct the phylogenetic tree using an MCMC chain of length 5×10^9, logging parameters every 10,000 states. The first 5×10^8 states were discarded as burn-in, and the remaining trees were used to find a consensus tree using `treeannotator v1.8.4` [19]. Convergence was assessed through trace plots of posterior distributions. The branches of the consensus tree are in evolutionary time (relative mutation rate $\mu = 1$) as we are only interested in the topology of the tree as a means of comparison with the results of the MCA (see Figure 1).

In the first two principal dimensions (Figure 1, panel A), L0 (bottom right quadrant) is visibly separated from the remaining haplogroups, and this makes sense, with L0 being the most divergent human mtDNA haplogroup. The deep split within L0, between L0d and L0k can be observed here, with the 10 furthest points representing the L0k sub-haplogroup.

L1 then separates (top left quadrant) from L2, L4 and L5 (bottom left quadrant), which is the next major split in the human mtDNA tree. L5 is also separated from L2 and L4, and this is the next major split. Finally, although it is not as pronounced as the previous separations, L2 and L4 separate, and this is the final major split.

The third dimension (Figure 1, panel B) shows a clear distinction between L5 (positive coordinates) and L2 and L4 (negative coordinates). The fourth dimension (Figure 1, panel B) separates L0d (positive) coordinates from L0k (negative coordinates). Dimension 5 (Figure 1, panel C) finds a separation between L2 (negative coordinates) and L4 (positive coordinates). Finally dimension 6 (Figure 1, panel C) separates L1c1 from the remaining L1c individuals. The remaining dimensions further identify splits in the tree, though this is not included in Figure 1. For this reason, when performing clustering we include all principal dimensions.

We performed hierarchical agglomerative clustering on the coordinates from the MCA using the R-package `cluster v2.0.6` [20]. The choice of termination point for identifying clusters is arbitrary, and in our case we cease identifying clusters when a cluster of size one is suggested.

The clustering algorithm respected the configuration of the points in the first two principal dimensions. The first cluster identified was L0, followed by L1 and L5. L0d and L0k are separated into two clusters, followed by the split between L2 and L4, which reflects the greater divergence time for the respective haplogroups [21]. The fifth cluster separation of L2 and L4 represents the final major haplogroup according to the current nomenclature.

The remaining clusters all respect the sub-haplogroup structure of the mtDNA tree, identifying sub-haplogroups for each of L0, L1, L2, L4 and L5. The clusters identified here are specific to this dataset, *i.e.* it may be the case that if more than one sequence from the haplogroup L1c2b2 were included, then we may have identified L1c2b2 as a cluster.

It is worth noting that our clustering suggests that the current nomenclature for human mtDNA may

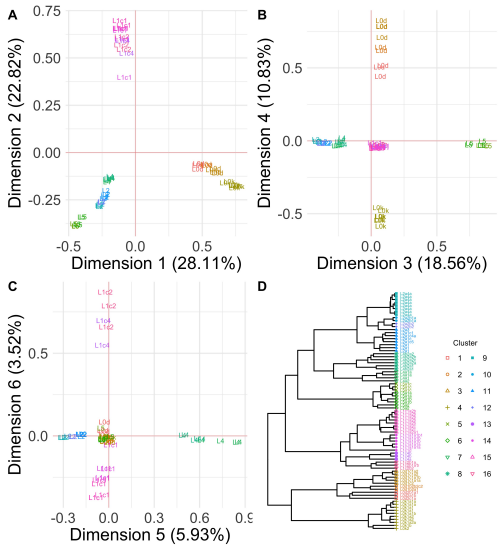


Figure 1: Scatter plots of the first six principal dimensions and phylogenetic reconstruction for the L-haplogroup alignment. Colors indicate cluster assignment via hierarchical agglomerative clustering.

not reflect statistically significant groups, but rather just the sequence of historically discovered diagnostic SNPs in densely sampled haplogroups. For example, the split between L0d and L0k appears more significant than the split between L2 and L4 in both the MCA and the phylogenetic tree. However, the sample sizes here are not large enough to refute the nomenclature, although the method provides a clear way forward to revise this.

Overall, the method has clearly shown that we can identify a tree like structure in the data, and that just the first two principal dimensions were able to visualise the haplotype structure in the data.

Application of method for continuous supplementary variables

The thylacine (*Thylacinus cynocephalus*) is an Australian marsupial carnivore most famous for its recent extinction due to human hunting [22, 23]. By the time of the arrival of Europeans to Australia, the thylacine had already undergone a significant population decline, was extinct on the mainland and was only found in Tasmania.

From museum samples we use sequence data from three samples from south-west Western Australia (WA), three samples from the Nullarbor Plain in WA, six samples from Tasmania (TAS) and one sample from New South Wales (NSW) (see Figure 2) [23]. Samples were removed if the longitude or latitude were unknown, or if the sampling age was unknown. To avoid artificial inflation of signal from geographical coordinates, for sequences found in the same location, a single representative was randomly selected. In total, 13 individuals were analysed (see Table S2 for supplementary variables and Genbank accession numbers).

Sequences were aligned using MAFFT v7.310 [16]. The alignment was filtered to remove homogeneous and missing sites, and a total of 113 SNPs were included in the MCA.

From the MCA row factor scores the first principal dimension, which captured 62.62% of the total inertia, correlates strongly with longitude ($r = 0.9467235$, $p = 9.517 \times 10^{-7}$), suggesting a possible migration gradient [24]. Gradients are not expected to be strictly linear for principal component maps, and the same can be assumed for MCA maps [4].

To investigate the relationship between geography and the MCA coordinates, a multi-response linear model was used. Multi-response linear models are similar to standard linear models, but allow for more than one response variable to be collectively modelled by the same set of explanatory variables [25]. A multi-response model was fitted to the data to predict latitude and longitude using principal dimension 1 (PD1). Polynomial models of varying degrees were fit and the best model was quadratic (using AIC), with R^2 values of 0.9334 and 0.9075 between longitude and latitude respectively.

A 'predicted' migration route can be projected onto the geographical map suggesting a coastal route was taken along the south of Australia (see Figure 2). However, the extremely small sample size means the results are limited as the MRCA of the sample is not necessarily closely related to the MRCA of the population, and there is little reason to believe that

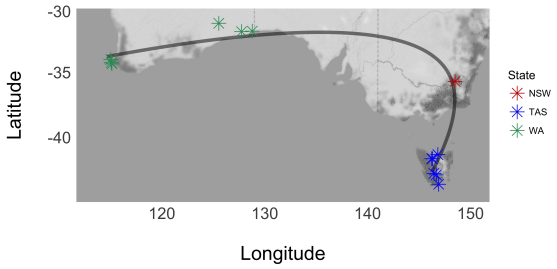


Figure 2: Sample location and sample IDs for thylacine mtDNA. The black line is the predicted geographic locations for thylacines given the observed range of principal dimension 1 coordinates. Colours indicate the location in which samples were found (Red=NSW, Blue=TAS, Green=WA).

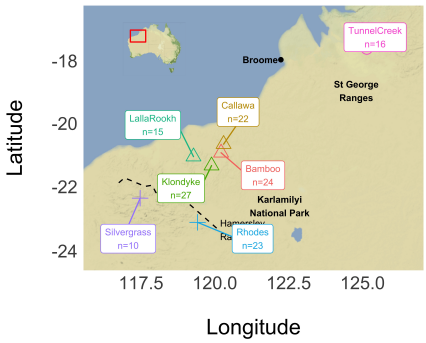


Figure 3: Sample location and sizes for ghost bat mtDNA. Colors indicate colonies and and shapes indicate

ancestral thylacine populations remained in the areas they originally inhabited.

Application of method for categorical supplementary variables

The ghost bat (*Macroderma gigas*) is a native Australian bat endemic to the Northern Pilbara and Kimberley in Western Australia, and in some regions of the Northern Territory and Queensland [26]. The ghost bat conservation status is currently listed as vulnerable by the International Union for Conservation of Nature.

Ghost bats are found in discrete populations within cave-based colonies. Blast mining disrupts, and in some cases destroys, cave complexes, displacing resident bat populations. Conservationists wish to understand the phylogeographic distribution of ghost bats to understand if the destruction of a single cave colony significantly reduces genetic diversity. Gene flow between colonies would indicate a reduced impact on ghost bat diversity, whereas a highly structured population would indicate a need to reduce the effects of blast mining and colony disruption.

We focus on samples collected in the Northern Pilbara found in four colonies on the northern side of the Hamersley Range (Bamboo, Callawa, Lalla Rookh, Klondyke), and two on the southern side (Rhodes, Silvergrass), and one colony in the Kimberley (Tunnel Creek) (see Figure 3). The Hamersley Range contains the twenty highest peaks in Western Australia, and so forms a significant geographical boundary for bats to cross. All colonies are represented by one sampled cave, with the Rhodes colony being the exception with four closely sampled caves (see Table S3 for supplementary variables and ID numbers).

We filtered an alignment of 257bp of the mtDNA HVR region, from 137 individuals, to remove homogeneous sites, and MCA was performed on the remaining 25 SNPs. For each individual, the colony and population (North, South, Kimberley) were recorded and treated as categorical supplementary variables. Longitude and latitude were also recorded and kept as quantitative supplementary variables.

In Figure 4 we present the squared-correlation plot for all supplementary variables and SNPs. The x and y coordinates of points in this plot give squared correlation values for the first two principal dimensions and each of the supplementary variables and SNPs. The further to the right of the plot a variable or SNP name is, the more highly correlated it is with Dimension 1. Similarly, the further to the top of the plot a variable or SNP is, the more highly correlated it is with Dimension 2. Immediately we see that latitude and longitude are not as strongly correlated with the two first principal dimensions as population and

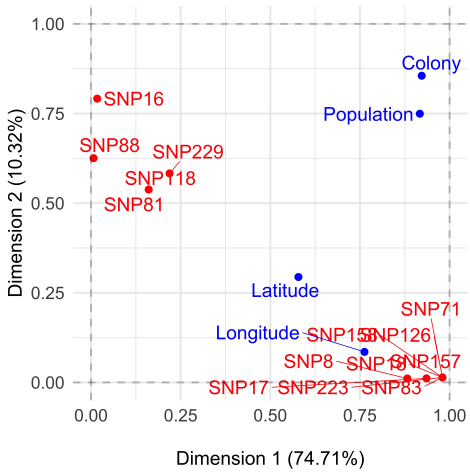


Figure 4: The correlation plot for all variables and SNPs for the Ghost Bat alignment.

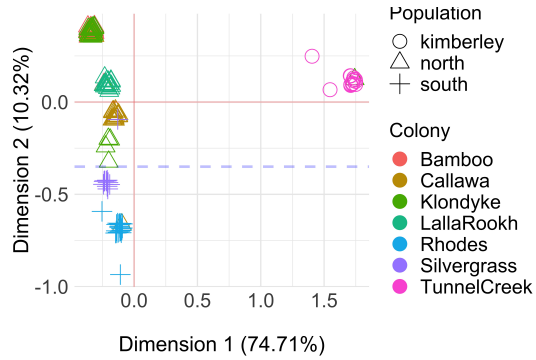


Figure 5: The scatter plot of the first two principal dimensions for the ghost bat alignment.

colony.

Calculating the η^2 values for population structure and colony structure yields $\eta^2_{pop} = 0.8332$ ($p < 9.999 \times 10^{-6}$) and $\eta^2_{col} = 0.8888$ ($p < 9.999 \times 10^{-6}$) respectively.

Clearly then, population explains a large proportion of the variability of the points in genetic space, however colony explains a greater proportion of the total variance, and thus had a larger η^2 value. Clearly colony explains a significant proportion of the structure of the individuals in genetic space since the first two principal dimensions explained a total of 85.03% of the total inertia.

The first principal dimension visualises the split between the Kimberley and Pilbara colonies (except for one Pilbara individual within the Kimberley samples), and the second principal dimension visualises the divide between the North and South colonies within the Pilbara region. In fact, if one places a boundary representing the Hamersley Range, on the y-axis at -0.35 (the dashed line in Figure 5), only one individual from the Northern Pilbara sample lies below the boundary, and only one Southern Pilbara sample lies above the boundary.

Since $\eta^2_{col} > \eta^2_{pop}$, this suggests that colony better explains the structure of the genetic coordinates. This can be seen in Figure 5 where we observe that five of the seven colonies form distinct clusters, with the exception of: a single Klondyke individual in the Tunnel Creek colony, a Callawa individual in the Rhodes colony, and a Silvergrass individual in the Callawa colony. The remaining two colonies, Klondyke and Bamboo, cluster together in the top left of the plot. A further four individuals from Klondyke form a separate cluster, genetically nearer the southern colonies. This may represent a recent migration from the southern colonies, or potentially members of the founding population for the southern colonies.

Without further information we cannot explain why these two colonies cluster together. It is worth noting that the Bamboo Creek site is a recently abandoned complex of mines (dismantled in 1962), whereas mining operations in the Klondyke mine area have increased drastically since 1955. It is possible that there has been a recent blending of the two colonies as bats have been displaced from places like Klondyke, and found refuge in caves left from abandoned mining efforts, like Bamboo Creek. However, our results still indicate a significant colony-based structure outside of these two colonies. This colony-based structure further strengthens the argument for more protection for roosting colonies from blast mining.

Finally we give an example of identifying potential diagnostic SNPs using our method. From Figure

4 we also see that we can identify potentially diagnostic SNPs. Given that Dimension 1 is explained by the geographical separation of the Kimberley and Pilbara populations, as all but one of the individuals with a positive first principal dimension coordinate are from the Kimberley, SNPs that are highly correlated with this dimension may be diagnostic for one of the regions. For example, SNP18 can be found in only Kimberley ghost bats (with the exception of the one Klondyke individual). If we had decided to use clustering to identify haplogroups, then SNP18 could be considered a diagnostic SNP, from this limited sample.

4 Materials and Methods

The transformation of genomic data to continuous coordinates

Consider an $n \times p$ alignment A of mtDNA, where $A_{ij} \in \{A, C, G, T\}$, filtered to remove homozygous sites. The n rows represent sequenced individuals, denoted $\{a_1, \cdots, a_n\}$ and the p columns represent single nucleotide polymorphisms (SNPs), denoted $\{s_1, \cdots, s_p\}$. Note that each of the SNPs can take between two to four forms, and we say that s_j has $|s_j|$ levels.

Consider each of the $Q = \sum_{j=1}^{p} |s_j|$ different allelic forms of the p SNPs, ordered (without loss of generality) numerically by position, then within SNPs, lexicographically by nucleotide. We can define an $n \times Q$ indicator matrix X, such that X_{ik} equals one if individual a_i has the allele at the position indicated by the kth column name, for $k = 1, \cdots, Q$ (see Figure 6). Note that for a SNP with $|s_j|$ levels, there are only $|s_j| - 1$ linearly independent columns of information in the X matrix (since if an individual does not have any of the first $|s_j| - 1$ forms of the allele, they must have the remaining allele). Hence, in total there are only $Q - p$ linearly independent columns. Finally, we can also calculate a contingency table of pairwise marker combinations $B = X^T X$ (see Figure 6), this matrix is discussed later in the process.

A =

	$SNP1$	$SNP2$	$SNP3$
a_1	A	G	C
a_2	A	T	C
a_3	C	T	G
a_4	C	T	G

X =

	$SNP1_A$	$SNP1_C$	$SNP2_G$	$SNP2_T$	$SNP3_C$	$SNP3_G$
a_1	1	0	1	0	1	0
a_2	1	0	0	1	1	0
a_3	0	1	0	1	0	1
a_4	0	1	0	1	0	1

B =

	$SNP1_A$	$SNP1_C$	$SNP2_G$	$SNP2_T$	$SNP3_C$	$SNP3_G$
$SNP1_A$	2	0	1	1	2	0
$SNP1_C$	0	2	0	2	0	2
$SNP2_G$	1	0	1	0	1	0
$SNP2_T$	1	2	0	3	1	2
$SNP3_C$	2	0	1	1	2	0
$SNP3_G$	0	2	0	2	0	2

Figure 6: A transformation from raw sequence alignment A, to an indicator matrix X, and a Burt table $B = X^T X$.

Let $N = \sum_{i=1}^{n} \sum_{j=1}^{Q} x_{ij}$, $\mathbf{r} = \frac{1}{N} X \mathbf{1}_Q$ and $\mathbf{c} = \frac{1}{N} X^T \mathbf{1}_n$, where $\mathbf{1}_k$ is a $k \times 1$ vector of ones, and define $D_r = \text{diag}(\mathbf{r})$ and $D_c = \text{diag}(\mathbf{c})$. We can define a new $n \times Q$ matrix, as a function of X,

$$f(X) = D_r^{-1/2} \left(\frac{1}{N} X - \mathbf{r}\mathbf{c}^T \right) D_c^{-1/2}. \qquad (1)$$

On $f(X)$ we perform a compact singular value decomposition (SVD) so that $f(X) = U \Sigma V^T$. Due to the above number of linearly independent columns, the diagonal matrix of singular values, Σ, will only have $J = Q - p$ non-zero entries, and we need only consider these dimensions. Following this reasoning, U and V are truncated to be matrices of dimensions $n \times J$ and $Q \times J$, respectively. From the diagonal matrix Σ we may also obtain the percentage of inertia (analogous to variability in PCA) explained by each of the first J principal dimensions, which are proportional to the singular values.

The *standard* row and column coordinates, defined as $F^* = {D_r}^{-1/2} U$ and $G^* = D_c^{-1/2} V$ respectively, are the unscaled row and factor scores that do not account for the proportion of inertia in principal dimensions. A natural choice for scaling the standard

row coordinates is to post multiply each dimension by the associated singular value. Hence the relative spread of points in each dimension is proportional to the amount of inertia captured by each dimension.

From the standard row scores we obtain the transformed coordinates, also called the 'row factor scores', and denoted F, of the individuals in the alignment A in 'genetic space' via

$$F = F^{*}\Sigma. \tag{2}$$

The distances between individuals calculated form these coordinates will respect three properties:

1. If two individuals have the same DNA sequence, they will have identical coordinates.
2. If two individuals share many alleles, they will be closer than two individuals that do not.
3. Individuals that share rare alleles will be closer still.

It is important to note that the pairwise distances between individuals calculated from the matrix F differ from classical pairwise genetic differences in two important ways. First, one need not assume a model of sequence evolution to find the matrix F. Second, classical pairwise genetic distances are calculated on only two sequences at a time, and so do not take into account the rarity of alleles. Our method uses the complete alignment to calculate the matrix F, and gives greater weight to rarer alleles.

The choice of rescaling for the standard column coordinates, with respect to the standard row coordinates, depends on the desired properties of the resulting column factor scores. We propose rescaling the standard column coordinates by the squares of the singular values, such that the column factors scores are

$$G = G^{*}\Sigma^{2}.$$

This rescaling of the standard column coordinates yields a desirable property for comparing the coordinates of individuals and alleles. The coordinates for any allele can be found at the centroid of the coordinates of the individuals that carry that allele (proof given in Appendix A). A special case of this property is that if an individual uniquely carries an allele, then that allele shares exactly the same coordinates as the individual (proof given in Appendix A).

There is a second, equivalent way to consider the method we have proposed. It is known that the row and column factor scores from $B = X^{T}X$ will be the same as the factor scores obtained from X [27]. The transformation $f(B)$ (found in a similar same way as in Equation 1, but with appropriate dimensions for $\mathbf{r}$, $\mathbf{c}$ and a recalculated normalising constant) would yield a $Q \times Q$ matrix, of the form $R = [\rho_{ij}]$. R is a matrix of the correlations for detecting linkage disequilibrium with multiple alleles, where if $\rho_{ij} \neq 0$, then the loci associated with alleles i and j are in linkage disequilibrium [28]. While mtDNA does not undergo recombination, our method also attempts to identify groups of alleles that occur together more than expected just by random chance, and hence the individuals that carry these alleles.

As with PCA, we can use the principal coordinates to visualize the relationships between individuals. However, our method also allows us to visualize the relationships between SNPs, and between individuals and SNPs. We can also look at the pairwise distances between individuals, and SNPs, in genetic space.

Figure 7 shows the relationship between the sequences as shown in Figure 6. Since a_3 and a_4 have identical sequences, they have the same coordinates in gene space. As a_1 shares no similarity with a_3 or a_4, they are the furthest apart. However, a_2 shares one SNP with a_3 and a_4, and two SNPs with a single individual a_1, and hence is more closely 'attracted' to a_1. Due to this 'attraction' to individuals with similar SNP profiles, the term 'inertia' is used in the place of 'variance'.

Note the relationship between individual coordinates, and SNP coordinates. Since a_3 and a_4 are the only individuals with 'SNP1_C' and 'SNP3_G', they share the same coordinates (the same can be said of a_1 and 'SNP2_G'). 'SNP1_A' is shared by a_1 and a_2, and so falls exactly at the mid-point of the two points. However, 'SNP2_T' is shared by a_2 and by both a_3 and a_4, and so lies only one-third the way along the line connecting a_3 and a_4 to a_2.

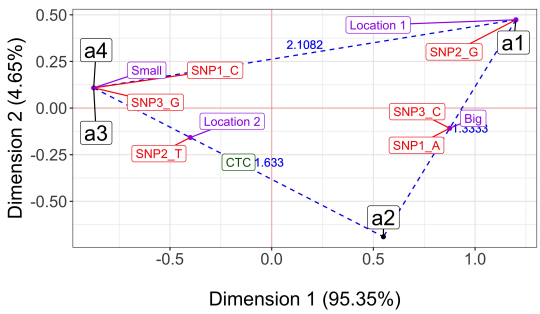


Figure 7: A biplot of the first two principal dimensions for the sequences as shown in Figure 6. Individuals are in black, SNPs are in red and projected coordinates for supplementary variable Size (Big and Small) and Location (Location 1 and Location 2) are shown in purple. The new sequence 'CTC' is projected onto the dimensions and given in green. Euclidean distances between individuals are given in blue.

Note that in Figure 6, if an individual has an 'A' at the first site, then they always have a 'C ' in the third position. Similarly, if an individual has a 'C' at the first site, then they always have a 'G ' in the third position. Hence the SNPs at the first and third sites provide no new information about the nature of the relationships between individuals since one can infer the third SNP, given the nature of the first SNP. For this reason, the first two principal dimensions capture 100% of the inertia, and reducing the dimensionality of the transformed genetic space results in no loss of information about the structure of the relationships between individuals.

It is possible to project new sequences onto the genetic space defined by an MCA. The new sequence must have one of the allelic forms for every SNP from the original alignment. For example, consider an alignment of new sequences of dimension $m \times p$ denoted H, with corresponding $m \times Q$ indicator matrix i_H (see Figure 8).

For the matrix $s_H = \text{diag}\,(i_H \mathbf{1}_Q)$, coordinates for the new sequences can be found [29] via

$$G_H = \left[s_H^{-1} i_H\right]^T D_c^{-1/2} V.$$

In Figure 7 we project the new sequence given in

H =

$SNP1$	$SNP2$	$SNP3$
C	T	C

$i_H =$

$SNP1_A$	$SNP1_C$	$SNP2_G$	$SNP2_T$	$SNP3_C$	$SNP3_G$
0	1	0	1	1	0

Figure 8: A transformation from a new raw sequence alignment H, to an indicator matrix i_H to be projected onto existing MCA dimensions.

Figure 8 onto the first two principal dimensions as given by the analysis of the alignment A from Figure 6. Note that this sequence is an equal 'mix' of the sequences a_2 and a_3, and so falls halfway along the line connecting the two sequences. While this makes mathematical and intuitive sense, this mixing of two sequences makes no sense for non-recombining DNA.

While this method could be used for pseudo-haploid DNA where this type of interpretation does make sense, there is value in projecting new sequences onto the principal dimensions for non-recombining DNA. For example, when data contains many SNPs, principal dimensions may represent haplogroups with a collection of diagnostic SNPs. Projecting ancient samples, for example, would include individuals ancestral to individuals from the alignment that have not acquired more recent SNPs. These projected points might fall along the line connecting the origin to the group, with individuals that carry fewer diagnostic SNPs closer to the origin.

Finally, we may project the 'average' coordinates of some qualitative supplementary variable. Imagine we have r such variables, with a total of R levels. Let W be the $n \times r$ matrix of supplementary information, with corresponding $n \times R$ indicator matrix j_W (see Figure 9).

Following a similar method for projecting new sequences, for the matrix $s_W = \text{diag}\,(\mathbf{1}_R j_W)$, coordinates for the average qualitative supplementary variables can be found via

$$F_W = \left[s_W^{-1} j_W\right]^T D_r^{-1/2} U \Sigma.$$

The projected coordinates for the supplementary variables are an estimate of the average coordinates for individuals with the given levels of the qualitative supplementary variables. For example, in Figure

W =

	Size	Location
a_1	Big	1
a_2	Big	2
a_3	Small	2
a_4	Small	2

j_W =

	Size_Big	Size_Small	Location_1	Location_2
a_1	1	0	1	0
a_2	1	0	0	1
a_3	0	1	0	1
a_4	0	1	0	1

Figure 9: A transformation from a matrix of supplementary qualitative information W, to an indicator matrix j_W to be projected onto existing MCA dimensions.

9, if you were to imagine that we could sample all individuals from the true population with level 'Big' for variable 'Size', then we believe that the centre of the cluster of points would fall at approximately $(0.875, -0.1083)$. Note that the calculated coordinates are based on inertia, and called 'barycentres' rather than centroids.

In Figure 7 we see that only sequences a_3 and a_4 have Size 'Small', and since they share coordinates, the barycentre for 'Small' also shares this coordinate. Sequences a_1 and a_2 both have Size 'Big', and so the barycentre for 'Big' falls halfway along the line connecting their coordinates. Similarly, a_1 is the only individual found at 'Location 1', and so the barycentre for 'Location 1' shares the coordinates of a_1. However, a_2, a_3 and a_4 were all found at 'Location 2', and so the barycentre for 'Location 2' can be found two thirds of the way along the line connecting a_2 to a_3 and a_4. Notice also that 'Location 2' is found exclusively when there is a T at SNP2, and so they also share coordinates.

Once a coordinate representation of the relationship between individuals has been constructed, we can examine relationships between individuals in this 'genetic space', and compare them to characteristics that have been recorded for individuals. These 'supplementary variables' are anything recorded about sampled individuals that were not used in the alignment table (*i.e.* any non-SNP data). Of particular interest are demographic variables, such as country of origin, spatial coordinates on a landscape or morphological characters, for example.

Here we give three examples that illustrate the ability of the method to produce biologically meaningful and intuitive results in a rigorous statistical framework, using previously published empirical data sets.

Correlation tests for continuous supplementary variables

Identifying relationships between coordinates in genetic space and continuous supplementary variables is intuitively simple. One could simply calculate the Pearson correlation coefficient for each continuous supplementary variable, followed by an exact test for a significantly non-zero coefficient.

It should be noted for a principal components analysis of spatially structured sequence data that has undergone recombination, that the top two principal components are expected to be highly correlated with perpendicular geographic axes [4]. In the case of mtDNA, or any other recombination-free sequence data, this assumption cannot be made. More extreme axis values can be interpreted as the accumulation of more and more of a unique set of SNPs that characterize some partition of the most-related tips of a tree.

Correlation tests for categorical supplementary variables

Supplementary categorical variables which explain significant proportions of the structure of individuals in gene space can be identified, and their effect quantified, using the correlation ratio η^2 [30]. The correlation ratio η^2 can be thought of as the proportion of variability explained by a qualitative variable. For a one-dimensional response variable Y, η^2 is equivalent to R^2, the coefficient of determination for a linear model with the qualitative variable as the sole predictor variable. In the case of multiple dimensional response variables, an analogous η^2 may be calculated [31].

A permutation test can be used to find if η^2 is significantly greater than for a random relabelling of the population. For each of the T permutations of the group labellings, we calculate η_t^2, the correlation ratio calculated for the tth permutation. An empirical

p-value of the form $(r+1)/(T+1)$ is calculated, where r is the total number of permuted samples yielding a greater correlation ratio than the observed sample [32].

Discussion

MCA provides a powerful method for unsupervized exploration of single-copy DNA. Our method is analogous to a PCA analysis of classical allele correlation values for detecting linkage disequilibrium. We have shown that $p-$dimensional single-copy DNA can be transformed into coordinates in genetic space, analogous to the way in which diploid DNA is transformed via PCA in many genetic studies. One of the attractive features of our approach is the parallel with PCA, making the interpretation of results natural for researchers experienced with PCA.

Our method allows for the coordinates of supplementary variables to be calculated and visualized in the same coordinate space as for individuals and SNPs, and for the relationships between the supplementary variables and principal dimensions to be quantified. Like PCA, additional sequences can be projected onto the coordinate space that has been calculated from an alignment of interest.

Dimension reduction can be performed, reducing potentially massive numbers of SNPs into far fewer dimensions with potentially little reduction in information, leading to informative visualization of high-dimensional data. Unlike PCA, our method is able to simultaneously investigate the relationships between individuals and SNPs. This extra information can lead to the detection of diagnostic SNPs, and potentially SNPs that are correlated with supplementary variables of interest such as habitat or phenotypic traits.

Our method was able to detect known haplotype structure, showing that the results of MCA are biologically meaningful. Similarly, the fact that our method can be reformulated as a PCA of the linkage disequilibrium table for multiple loci indicates that applications to recombining DNA may also be useful for detecting population structure.

Using techniques from classical statistics, our method was also able to efficiently visualise the strength of the relationships between supplementary information and empirical sequence data. Finally, using standard polynomial regression techniques, our method was able to identify a possible migration route for geographically distributed sequence data.

References

[1] Cann, R. L., Stoneking, M., and Wilson, A. C., 1987. "Mitochondrial DNA and human evolution". *Nature,* **325**(6099), pp. 31–36.

[2] Leonardi, M., Librado, P., Der Sarkissian, C., Schubert, M., Alfarhan, A. H., Alquraishi, S. A., Al-Rasheid, K. A., Gamba, C., Willerslev, E., and Orlando, L., 2017. "Evolutionary patterns and processes: lessons from ancient DNA". *Systematic Biology,* **66**(1), pp. e1–e29.

[3] Tobler, R., Rohrlach, A., Soubrier, J., Bover, P., Llamas, B., Tuke, J., Bean, N., Abdullah-Highfold, A., Agius, S., O'Donoghue, A., et al., 2017. "Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia.". *Nature,* **544**(7649), p. 180.

[4] Novembre, J., and Stephens, M., 2008. "Interpreting principal component analyses of spatial population genetic variation.". *Nature Genetics,* **40**(5), pp. 646–649.

[5] Posth, C., Renaud, G., Mittnik, A., Drucker, D. G., Rougier, H., Cupillard, C., Valentin, F., Thevenet, C., Furtwängler, A., Wißing, C., et al., 2016. "Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe.". *Current Biology,* **26**(6), pp. 827–833.

[6] Van Gremberghe, I., Leliaert, F., Mergeay, J., Vanormelingen, P., Van der Gucht, K., Debeer, A.-E., Lacerot, G., De Meester, L., and Vyverman, W., 2011. "Lack of phylogeographic structure in the freshwater cyanobacterium Microcystis aeruginosa suggests global dispersal.". *PloS One,* **6**(5), p. e19561.

[7] Pearson, K., 1901. "LIII. On lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* **2**(11), pp. 559–572.

[8] Patterson, N., Price, A., and Reich, D., 2006. "Population Structure and Eigenanalysis.". *PLoS Genetics,* **2**(12), p. e190.

[9] Jolliffe, I., 2002. *Principal Component Analysis.* Wiley Online Library.

[10] Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.-J., 2006. "Harvesting the fruit of the human mtDNA tree.". *TRENDS in Genetics,* **22**(6), pp. 339–345.

[11] Behar, D. M., Metspalu, E., Kivisild, T., Rosset, S., Tzur, S., Hadid, Y., Yudkovsky, G., Rosengarten, D., Pereira, L., Amorim, A., et al., 2008. "Counting the founders: the matrilineal genetic ancestry of the Jewish Diaspora". *PloS One,* **3**(4), p. e2062.

[12] Costa, M. D., Cherni, L., Fernandes, V., Freitas, F., el Gaaied, A. B. A., and Pereira, L., 2009. "Data from complete mtDNA sequencing of Tunisian centenarians: testing haplogroup association and the "golden mean" to longevity.". *Mechanisms of Ageing and Development,* **130**(4), pp. 222–226.

[13] Batini, C., Lopes, J., Behar, D. M., Calafell, F., Jorde, L. B., Van der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G., and Comas, D., 2011. "Insights into the demographic history of African Pygmies from complete mitochondrial genomes.". *Molecular Biology and Evolution,* **28**(2), pp. 1099–1110.

[14] Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S. W., Stoneking, M., and Pakendorf, B., 2013. "Ancient substructure in early mtDNA lineages of southern Africa.". *The American Journal of Human Genetics,* **92**(2), pp. 285–292.

[15] Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N., 1999. "Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA.". *Nature Genetics,* **23**(2), pp. 147–147.

[16] Katoh, K., and Standley, D. M., 2013. "MAFFT multiple sequence alignment software version 7: improvements in performance and usability.". *Molecular Biology and Evolution,* **30**(4), pp. 772–780.

[17] Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.-J., Kronenberg, F., Salas, A., and Schönherr, S., 2016. "HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing". *Nucleic Acids Research*, pp. W58–W63.

[18] Keane, T., Naughton, T., and McInerney, J., 2004. "Modelgenerator: amino acid and nucleotide substitution model selection". *National University of Ireland, Maynooth, Ireland,* **34**.

[19] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J., 2014. "BEAST 2: a software platform for Bayesian evolutionary analysis.". *PLoS Computational Biology,* **10**(4), p. e1003537.

[20] Kaufman, L., and Rousseeuw, P. J., 2009. *Finding Groups in Data: an Introduction to Cluster Analysis.*, Vol. 344. John Wiley & Sons.

[21] Gonder, M. K., Mortensen, H. M., Reed, F. A., de Sousa, A., and Tishkoff, S. A., 2007. "Whole-mtDNA genome sequence analysis of ancient African lineages.". *Molecular Biology and Evolution,* **24**(3), pp. 757–768.

[22] Harris, G. P., 1808. "XI. Description of two new Species of Didelphis from Van Diemen's Land.". *Transactions of the Linnean Society of London,* **9**(1), pp. 174–178.

[23] White, L. C., Mitchell, K. J., and Austin, J. J., 2017. "Ancient Mitochondrial Genomes Reveal the Demographic History and Phylogeography

of the Extinct, Enigmatic Thylacine (Thylacinus Cynocephalus)". *Journal of Biogeography.*

[24] Menozzi, P., Piazza, A., and Cavalli-Sforza, L., 1978. "Synthetic Maps of Human Gene Frequencies in Europeans.". *Science,* **201**(4358), pp. 786–792.

[25] Berridge, D. M., and Crouchley, R., 2011. *Multivariate generalized linear mixed models using R.* CRC Press.

[26] Armstrong, K. N., and Anstee, S. D., 2000. "The ghost bat in the Pilbara: 100 years on.". *Australian Mammalogy,* **22**(2), pp. 93–101.

[27] Greenacre, M., 2007. *Correspondence Analysis in Practice.* CRC Press.

[28] Zaykin, D. V., Pudovkin, A., and Weir, B. S., 2008. "Correlation-based inference for linkage disequilibrium with multiple alleles.". *Genetics,* **180**(1), pp. 533–545.

[29] Abdi, H., and Valentin, D., 2007. "Multiple Correspondence Analysis.". *Encyclopedia of Measurement and Statistics.*, pp. 651–657.

[30] Brown, J. D., 2008. "Effect Size and Eta Squared.". *JALT Testing & Evaluation SIG News.*

[31] Breiman, L., and Friedman, J. H., 1997. "Predicting multivariate responses in multiple linear regression.". *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **59**(1), pp. 3–54.

[32] Davison, A. C., and Hinkley, D. V., 1997. *Bootstrap Methods and their Application.*, Vol. 1. Cambridge University Press.

4.1 Appendix A

We aim to show that our choice of scaling for the row and column factor scores yields the property that if an individual uniquely carries an allele, then the individual and the allele share the same coordinates. To do this we investigate properties of the indicator matrix X, where the columns have been permuted to make the first column the identifying allele and to make the first row the identified individual (without loss of generality). To avoid carrying constants, we assume that X has already been normalized to have grand sum one.

Result 1: If there is an allele that uniquely identifies an individual, then the individual and the allele have the same coordinates if the standard row factor scores are scaled by the singular values, and the standard column factor scores are scaled by the squared singular values.

Proof: Let X be an $n \times Q$ matrix such that

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1Q} \\ 0 & x_{22} & \dots & x_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n2} & \dots & x_{nQ} \end{bmatrix}, \tag{3}$$

such that $\sum_{i=1}^{n} \sum_{j=1}^{Q} x_{ij} = 1$ and $x_{ij} \geq 0 \, \forall i, j$.

Let $\mathbf{1}_k$ be a $k \times 1$ vector of ones, and let

$$\mathbf{r} = X\mathbf{1}_Q = (r_1, \cdots, r_n)^T$$

and

$$\mathbf{c} = X^T\mathbf{1}_n = (c_1, \cdots, c_Q)^T$$

be the strictly positive row and column sums of X respectively, and let

$$D_r = \text{diag}(\mathbf{r}) \text{ and } D_c = \text{diag}(\mathbf{c}).$$

We begin by showing that the SVD of the matrix

$$A = D_r^{-1/2} X D_c^{-1/2}, \tag{4}$$

has a particular structure. We then exploit this to show the required result for the matrix

$$C = D_r^{-1/2} \left(X - \mathbf{r}\mathbf{c}^T \right) D_c^{-1/2}, \tag{5}$$

namely that for our choice of row and column standard coordinate scaling, the identifying allele and the identified individual share the same scaled factor scores.

Let A have a compact singular value decomposition (SVD) of the form

$$A = U_A \Sigma_A V_A^T. \tag{6}$$

Consider the matrix product

$$M = D_c^{-1/2} A^T U_A. \tag{7}$$

Since $U_A \Sigma_A V_A^T$, is a SVD, then U_A is a unitary matrix, and so $U_A^T U_A = I_n$ (where I_k is the $k \times k$ identity matrix), and substituting Equation (6) into Equation (7) gives,

$$\begin{aligned} M &= D_c^{-1/2} A^T U_A \\ &= D_c^{-1/2} \left(U_A \Sigma_A V_A^T\right)^T U_A \\ &= D_c^{-1/2} V_A \Sigma_A^T U_A^T U_A \\ &= D_c^{-1/2} V_A \Sigma_A. \end{aligned}$$

Hence, the first row of M is $c_1^{-1/2} \mathbf{v}_1 \Sigma_A$, where $\mathbf{v}_1$ is the first row of V_A.

However, instead substituting Equation (4) into Equation (7) yields

$$\begin{aligned} M &= D_c^{-1/2} A^T U_A \\ &= D_c^{-1/2} \left(D_r^{-1/2} X D_c^{-1/2}\right)^T U_A \\ &= D_c^{-1} X^T D_r^{-1/2} U_A. \end{aligned}$$

Note that since $x_{21} = \cdots = x_{n1} = 0$, then $c_1 = x_{11}$ and hence the first row of $A = D_c^{-1} X^T D_r^{-1/2}$ will equal

$$c_1^{-1} r_1^{-1/2} \left(x_{11}, 0, \cdots, 0\right) = \left(r_1^{-1/2}, 0, \cdots, 0\right).$$

So the first row of M is also $r_1^{-1/2} \mathbf{u}_1$, where $\mathbf{u}_1$ is the first row of U_A.

This shows that

$$r_1^{-1/2} \mathbf{u}_1 = c_1^{-1/2} \mathbf{v}_1 \Sigma_A, \tag{8}$$

which are the first rows of the row and column scores of the SVD of $D_r^{-1/2} X D_c^{-1/2}$, where X is of the form given in Equation (3).

To extend this result to the SVD of the matrix C in Equation 5, first note that

$$\begin{aligned} A^T \left(D_r^{-1/2} \mathbf{r}\right) &= D_c^{-1/2} X^T D_r^{-1} \mathbf{r} \\ &= D_c^{-1/2} X^T \mathbf{1}_n \\ &= D_c^{-1/2} \mathbf{c} \end{aligned}$$

and

$$\begin{aligned} A \left(D_c^{-1/2} \mathbf{c}\right) &= D_r^{-1/2} X D_c^{-1} \mathbf{c} \\ &= D_r^{-1/2} X \mathbf{1}_n \\ &= D_r^{-1/2} \mathbf{r}. \end{aligned}$$

Now this shows that the SVD of A has a singular value of 1, with left and right singular vectors $D_r^{-1/2} \mathbf{r}$ and $D_c^{-1/2} \mathbf{c}$, respectively.

We now show that a SVD of C, denoted $C = U_C \Sigma_C V_C^T$ can be augmented by these singular vectors and the singular value 1 to construct a SVD for A.

Consider new matrices

$$U_* = \left[U_C \middle| D_r^{-1/2} \mathbf{r}\right], V_* = \left[V_C \middle| D_c^{-1/2} \mathbf{c}\right],$$

and

$$\Sigma_* = \left[\begin{array}{c|c} \Sigma_C & \mathbf{0}_n \\ \hline \mathbf{0}_Q^T & 1 \end{array}\right],$$

where $\mathbf{0}_k$ is a $k \times 1$ vector of zeros. Next we show that U^*, $V*$ are unitary matrices and Σ^* is a rectangular matrix diagonal matrix with non-negative real numbers on the diagonal.

$$\begin{aligned} &\mathbf{r}^T D_r^{-1/2} U_C \Sigma_C V_C^T \\ =&\mathbf{r}^T D_r^{-1/2} D_r^{-1/2} \left(X - \mathbf{r}\mathbf{c}^T\right) D_c^{-1/2} \\ =&\mathbf{r}^T D_r^{-1} \left(X - \mathbf{r}\mathbf{c}^T\right) D_c^{-1/2} \\ =&\mathbf{1}_n \left(X - \mathbf{r}\mathbf{c}^T\right) D_c^{-1/2} \\ =&\left(\mathbf{1}_n X - \mathbf{1}_n \mathbf{r}\mathbf{c}^T\right) D_c^{-1/2} \\ =&\left(\mathbf{c}^T - \mathbf{c}^T\right) D_c^{-1/2} \\ =&\mathbf{0}_Q D_c^{-1/2} \\ =&\mathbf{0}_Q. \end{aligned}$$

Since, Σ_C is a diagonal matrix with positive diagonal entries, we know that Σ_C^{-1} exists. Further, $V_C^{-1} = V_C^T$, so this implies that

$$\begin{aligned} &\mathbf{r}^T D_r^{-1/2} U_C \Sigma_C V_C^T = \mathbf{0}_Q \\ \implies &\mathbf{r}^T D_r^{-1/2} U_C = \mathbf{0}_Q. \end{aligned}$$

Similarly, it can be shown that

$$\mathbf{c}^T D_c^{-1/2} V_C = \mathbf{0}_n.$$

It follows then that

$$\begin{aligned} U_*^T U_* &= \left[\begin{array}{c|c} U_C^T U_C & U_C^T D_r^{-1/2}\mathbf{r} \\ \hline \mathbf{r}^T D_r^{-1/2} U_C & \mathbf{r}^T D_r^{-1}\mathbf{r} \end{array}\right] \\ &= \left[\begin{array}{c|c} I_n & \mathbf{0}_Q \\ \hline \mathbf{0}_Q^T & 1 \end{array}\right] \\ &= I_{n+1}, \end{aligned}$$

and that

$$\begin{aligned} V_*^T V_* &= \left[\begin{array}{c|c} V_C^T V_C & V_C^T D_c^{-1/2}\mathbf{c} \\ \hline \mathbf{c}^T D_c^{-1/2} V_C & \mathbf{c}^T D_c^{-1}\mathbf{c} \end{array}\right] \\ &= \left[\begin{array}{c|c} I_Q & \mathbf{0}_n \\ \hline \mathbf{0}_n^T & 1 \end{array}\right] \\ &= I_{Q+1}. \end{aligned}$$

Note that

$$\begin{aligned} &U_* \Sigma_* V_*^T \\ =&U_C \Sigma_C V_C^T + D_r^{-1/2}\mathbf{r}\mathbf{c}^T D_c^{-1/2} \\ =&D_r^{-1/2}\left(X - \mathbf{r}\mathbf{c}^T\right) D_c^{-1/2} + D_r^{-1/2}\mathbf{r}\mathbf{c}^T D_c^{-1/2} \\ =&D_r^{-1/2} X D_c^{-1/2} \\ =&A. \end{aligned}$$

Therefore, since Σ_* is a rectangular diagonal matrix, with positive diagonal entries, and U_* and V_* are unitary matrices, it must be that $U_*\Sigma_*V_*^T$ is a SVD of A.

Thus we have two representations for the compact SVD of A. Hence they are equivalent. However, from Equation 8, we know that the first row and column factor scores for the SVD of A are equal, and are given by

$$r_1^{-1/2}\mathbf{u}_1^* = c_1^{-1/2}\mathbf{v}_1^*\Sigma_*.$$

Hence any sub vectors that are constructed from removing corresponding elements from the vectors $r_1^{-1/2}\mathbf{u}_1^*$ and $c_1^{-1/2}\mathbf{v}_1^*\Sigma_*$ will also be equal, specifically

$$r_1^{-1/2}\mathbf{u}_1^C = c_1^{-1/2}\mathbf{v}_1^C\Sigma_C,$$

where $\mathbf{u}_1^C$ is the first column of U_C, and $\mathbf{v}_1^C$ is the first column of V_C.

□.

Result 2: If a single allele identifies a group of m individuals, the the column factor score for the allele is the centroid of the row factor scores of the identified individuals, if the standard row factor and the standard column factor scores are scaled by the squared singular values.

Proof: If

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1Q} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mQ} \\ 0 & x_{(m+1)2} & \cdots & x_{(m+1)Q} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & x_{n2} & \cdots & x_{nQ} \end{bmatrix},$$

where $x_{11} = \cdots = x_{m1} > 0$, and $x_{(m+1)1} = \cdots = x(n1) = 0$, and $\sum_{i=1}^{n}\sum_{j=1}^{Q} x_{ij} = 1$ and $x_{ij} \geq 0\ \forall i, j$. As previously, the first column of $M = D_c^{-1/2}A^T U_A$ is $c_1^{-1/2}\mathbf{v}\Sigma_A$. However, the sum of the first column of X is mx_{11}, and hence the first row of $D_c^{-1}X^T D_r^{-1/2}$ is now

$$\begin{aligned} &\frac{1}{mx_{11}}\left(x_{11}, \cdots, x_{m1}, 0, \cdots, 0\right) \\ =&\left(1/m, \cdots, 1/m, 0, \cdots, 0\right), \end{aligned}$$

and it follows that the first column of M is also

$$\frac{1}{m}\sum_{i=1}^{m} r_i^{-1/2}\mathbf{v}_i,$$

yielding that

$$c_1^{-1/2}\mathbf{v}\Sigma_A = \frac{1}{m}\sum_{i=1}^{m} r_i^{-1/2}\mathbf{v}_i.$$

Following the same argument as before, it is also true that this is the case for the SVD of $C = D_r^{-1/2}\left(X - \mathbf{r}\mathbf{c}^T\right)D_c^{-1/2}$. Hence the identifying allele has column factor score equal to the centroid of the row factor scores for the identified individuals.

$\square$

Chapter 3

An Application of Unsupervized Quantification of Demographic Structure for Single-copy Alignments

3.1 Introduction

In this chapter we present the paper titled "Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia" for which I was a joint first author. This paper was published in *Nature* on the 8th of March, 2017 and is an application of the work presented in Chapter 2.

In this work we investigate the details of the arrival, and first colonisation of Australia, by the Aboriginal Australian peoples. We were interested in not only the timing of these events, but also whether or not there was any remaining structure to the genetic diversity that persisted from the original colonisation.

This work was a unique opportunity to study the geographic distribution of Aborig-

inal Australian genetic diversity that existed before the arrival of European settlers. The data came in the form of museum hair samples (collected as early as 1928), for which informed consent was obtained from the original donors, or from their family. Family history (pre-dating the European resettling of Aboriginal Australian peoples) was recorded from donors at the time of collection, hence hair samples could reliably be provenanced to their ancestral lands. From the hair samples, complete mitogenomes were sequenced. All of the work to collect the samples, obtain informed consent from donors (or their families) and the sequencing of the mtDNA was performed by our collaborators at the Australian Centre for Ancient DNA, and the South Australian Museum.

Our contribution to this research was the complete development of a method to construct a coordinate representation of the aligned sequence data. Once we had constructed this coordinate representation, we designed statistical tests for detecting geographic structure in the coordinates we had calculated.

We found an extremely strong positive relationship between the distances between individuals in geographical space, and the distances between individuals in genetic space. That is, individuals that were more closely related genetically were also closer geographically. We also found a significant relationship between longitude and latitude, and many of the important principal dimensions in the genetic space. That is, we were able to show that genetic markers that catergorise certain haplogroups could be strongly linked to specific locations on the map of Australia. These two findings, along with full phylogenetic reconstructions of the alignment data, allowed us to infer an entry time of approximately 50,000 years before present, with a two-pronged stepping-stone migration around the coast of Australia, meeting in South Australia. We found evidence to support that these original settlements persisted from the original migration until the arrival of European settlers. This was in agreement with the notable Aboriginal Australian cultural attachment to their country, and is further evidenced by Songlines and Dreaming narratives [47].

Here we include the main publication, but we also include the supplementary information for completeness. We direct the reader to our contribution in the methods section on Pages 1 and 2 of the supplementary information, although this methodology is fully described in Chapter 2.

3.2 Statement of Authorship

Statement of Authorship

Title of Paper	Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia
Publication Status	Published
Publication Details	Tobler, Ray, Adam Rohrlach, Julien Soubrier, Pere Bover, Bastien Llamas, Jonathan Tuke, Nigel Bean et al. "Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia." *Nature* 544, no. 7649 (2017): 180.

Principal Authors

Adam Rohrlach (Candidate)			
Contribution to the Paper	Designed statistical methods for analysing phylogeography and spectral decomposition of single-copy DNA. Wrote the paper with help from all co-authors.		
Overall percentage (%)	40		
Signature		Date	24/10/2018

Raymond Tobler			
Contribution to the Paper	Designed experiments. Performed bioinformatics analyses: processed and analysed mtDNA data, phylogenetics. Analysed and interpreted results. Wrote the paper with help from all co-authors.		
Overall percentage (%)	40		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am one of two primary authors of this paper.		
Signature		Date	10/10/18

Co-Author Contributions

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

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

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

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

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Designed experiments, provided samples, interpreted results. Wrote the paper with help from all co-authors.		
Signature		Date	10/10/18

Name of Co-Author	Alan Williams		
Contribution to the Paper	Provided archaeological input and advice in development of the paper. Contributed to the writing and development of the final manuscript. I support Adam's use of this paper in the submission of his PHD thesis.		
Signature		Date	10.10.18

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Performed laboratory work to generate mitochondrial data, performed data processing and analyses, edited manuscript.		
Signature		Date	10/10/2018

Name of Co-Author	Matthew Williams		
Contribution to the Paper	I helped generate the ancient DNA libraries and helped construct the genealogical relationships.		
Signature		Date	14.10.18

Name of Co-Author	Emma Kowal		
Contribution to the Paper	Contributed ethical and social science perspectives to the paper.		
Signature		Date	10/10/18

Name of Co-Author	Fran Zilio		
Contribution to the Paper	Contributed to the writing and development of the final manuscript.		
Signature		Date	11/10/2018

Name of Co-Author	Wolfgang Haak		
Contribution to the Paper	Conceived the concept of the study and wrote grant application, liaised and consulted with Aboriginal communities, collected and processed hair samples, established protocols, conducted genetic, genealogical, and phylogeographic analyses, interpreted the results, reported back to Aboriginal communities, wrote the paper with help from all co-authors.		
Signature		Date	09/10/2018

Name of Co-Author	Peter Sutton		
Contribution to the Paper	Initiated possibility of project by informing Prof. Cooper of details of the Aboriginal hair collection of the SA Museum upon his arrival in Adelaide in 2005. Provided anthropological input on the nature of classical Aboriginal societies and their territorial arrangements, and on the linguistic prehistory of Australia. Made editorial suggestions on drafts.		
Signature		Date	10/10/2018

Name of Co-Author	Julien Soubrier		
Contribution to the Paper	Performed phylogenetic analyses for divergence time analyses.		
Signature		Date	09/10/18

Name of Co-Author	Pere Bover		
Contribution to the Paper	Contributed to the writing and development of the final manuscript.		
Signature		Date	10/10/2018

Name of Co-Author	Nigel Bean		
Contribution to the Paper	Interpreted results and helped to design spectral decomposition methodology. Wrote the paper with help from all co-authors.		
Signature		Date	09/10/2018

Name of Co-Author	Jonathan Tuke		
Contribution to the Paper	Interpreted results and helped to design spectral decomposition methodology. Wrote the paper with help from all co-authors.		
Signature		Date	09/10/18

Name of Co-Author	Stephen Richards		
Contribution to the Paper	Developed the in-house mitogenome hybridization capture method		
Signature		Date	10/10/18

Name of Co-Author	Chris S. M. Turney		
Contribution to the Paper	Age modelling of early occupation sites across the Sahul for comparison to the genetic data.		
Signature		Date	18 October 2018

Name of Co-Author	Robert Mitchell		
Contribution to the Paper	Contributed to the writing and development of the final manuscript.		
Signature		Date	23/10/2018

Name of Co-Author	Amy O'Donoghue		
Contribution to the Paper	Community consultation and archival research.		
Signature		Date	23/10/2018

Name of Co-Author	Lesley Williams		
Contribution to the Paper	Community consultation and archival research.		
Signature		Date	23/10/2018

Name of Co-Author	Ali Abdullah-Highfold		
Contribution to the Paper	Community consultation and archival research.		
Signature		Date	23/10/2018

Name of Co-Author	Shane Agius		
Contribution to the Paper	Community consultation and archival research.		
Signature		Date	23/10/2018

Name of Co-Author	Isabel O'Laughlin		
Contribution to the Paper	Community consultation and archival research.		
Signature		Date	23/10/2018

To whom it may concern,

As the Director of the Australian Centre for Ancient DNA, the lab at which the significant proportion of the work for the publication "*Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia*" was performed, I certify that the candidate Adam Benjamin Rohrlach completed the work as indicated in the Statement of Authorship.

Unfortunately Adam was unable to obtain the signatures of Keryn Walsche and John R. Stephen. However, I can confirm that Adam made significant efforts to try and obtain statements from all authors.

Sincerely,

Professor Alan Cooper

ARTICLE

doi:10.1038/nature21416

Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia

Ray Tobler[1]*, Adam Rohrlach[2,3]*, Julien Soubrier[1,4], Pere Bover[1], Bastien Llamas[1], Jonathan Tuke[2,3], Nigel Bean[2,3], Ali Abdullah-Highfold[5], Shane Agius[5], Amy O'Donoghue[5], Isabel O'Loughlin[5], Peter Sutton[5,6], Fran Zilio[5], Keryn Walshe[5], Alan N. Williams[7], Chris S.M. Turney[7], Matthew Williams[1,8], Stephen M. Richards[1], Robert J. Mitchell[9], Emma Kowal[10], John R. Stephen[11], Lesley Williams[12], Wolfgang Haak[1,13]§ & Alan Cooper[1,14]§

Aboriginal Australians represent one of the longest continuous cultural complexes known. Archaeological evidence indicates that Australia and New Guinea were initially settled approximately 50 thousand years ago (ka); however, little is known about the processes underlying the enormous linguistic and phenotypic diversity within Australia. Here we report 111 mitochondrial genomes (mitogenomes) from historical Aboriginal Australian hair samples, whose origins enable us to reconstruct Australian phylogeographic history before European settlement. Marked geographic patterns and deep splits across the major mitochondrial haplogroups imply that the settlement of Australia comprised a single, rapid migration along the east and west coasts that reached southern Australia by 49–45 ka. After continent-wide colonization, strong regional patterns developed and these have survived despite substantial climatic and cultural change during the late Pleistocene and Holocene epochs. Remarkably, we find evidence for the continuous presence of populations in discrete geographic areas dating back to around 50 ka, in agreement with the notable Aboriginal Australian cultural attachment to their country.

At the time of initial human colonization (around 50 ka)[1,2], Australia and New Guinea were connected as a single landmass (termed Sahul) that remained contiguous until separated by rising sea levels around 9 ka (ref. 3). Despite this, the initial Sahul colonists appear to have rapidly diverged into distinct New Guinean and Australian populations, with limited signs of subsequent gene flow[4–12]—although genetic data remains sparse. Little is known about the post-colonization diversification of Australian lineages or the effects of major environmental and cultural changes over the last 50 thousand years (kyr). Palaeoclimatically, these include continental-scale aridification and cooling of Australia during the Last Glacial Maximum (21 ± 3 ka), warming in the early Holocene (9–6 ka), and intensification of the El Niño/Southern Oscillation during the mid-to-late Holocene (4–2 ka)[13,14]. Substantial changes in the cultural record are not observed until the terminal Pleistocene and Holocene, and include the formation of the Panaramittee art style, the spread of the Pama–Nyungan group of languages across most of the continent, and the increase in diversity and complexity of technology and resource exploitation[15,16]. Aboriginal history is inextricably interwoven with the Australian landscape and is culturally expressed through the central importance of kin group attachment to 'country', and further reinforced through Songlines and Dreaming narratives[17]. Close relationships to the landscape are likely to have played an important role in surviving the extreme environmental changes of late Pleistocene Australia.

Reconstructing the genetic history of Aboriginal Australia is greatly complicated by past government policies of enforced population relocation and child removal that have eroded much of the physical connection between groups and geography in modern Australia. However, a unique opportunity is provided by a remarkable set of hair samples and detailed ethnographic metadata collected with permission from more than 5,000 Aboriginal Australians during expeditions run by the Board for Anthropological Research (BAR) from the University of Adelaide between the 1920s and 1970s (Supplementary Information). The extensive genealogical and geographical information collected with the samples allows detailed reconstruction of the genetic and historical relationships between Aboriginal Australian groups before the effects of European colonization.

Dataset

We obtained informed consent from hair donors or their families (Supplementary Information) to perform genetic analyses and sequenced complete mitogenomes from hair samples of 111 individuals across three different Aboriginal communities (Point Pearce, South Australia; Cherbourg, Queensland; Koonibba, South Australia; Supplementary Information). Using the genealogical and cultural metadata, we traced the geographic origin of each individual (referred to as BAR samples) as far back as possible along the ancestral maternal lineage. The resulting broad geographic range is shown in Extended Data Fig. 1. We identified 54 unique mtDNA haplotypes, which fell into the five major mitochondrial haplogroups S, O, M, P and R that have been described previously for Aboriginal Australia[9,10,12] (Supplementary Information). Phylogenetic relationships were

[1]Australian Centre for Ancient DNA, School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia. [2]School of Mathematical Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia. [3]ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide, South Australia 5005, Australia. [4]Genetics and Molecular Pathology, SA Pathology, Adelaide, South Australia 5000, Australia. [5]South Australian Museum, Adelaide, South Australia 5005, Australia. [6]School of Biological Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia. [7]Palaeontology, Geobiology and Earth Archives Research Centre, and Climate Change Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia. [8]School of Archaeology and Anthropology, College of Arts and Social Sciences, Australian National University, Canberra, Australian Capital Territory 0200, Australia. [9]Department of Biochemistry and Genetics, La Trobe University, Melbourne, Victoria 3086, Australia. [10]Alfred Deakin Institute, Deakin University, Melbourne, Victoria 3125, Australia. [11]Australian Genome Research Facility, The Waite Research Precinct, Adelaide, South Australia 5064, Australia. [12]Community Elder and Cultural Advisor, Cherbourg, Queensland, Australia. [13]Department of Archeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany. [14]Environment Institute, The University of Adelaide, Adelaide, South Australia 5005, Australia.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

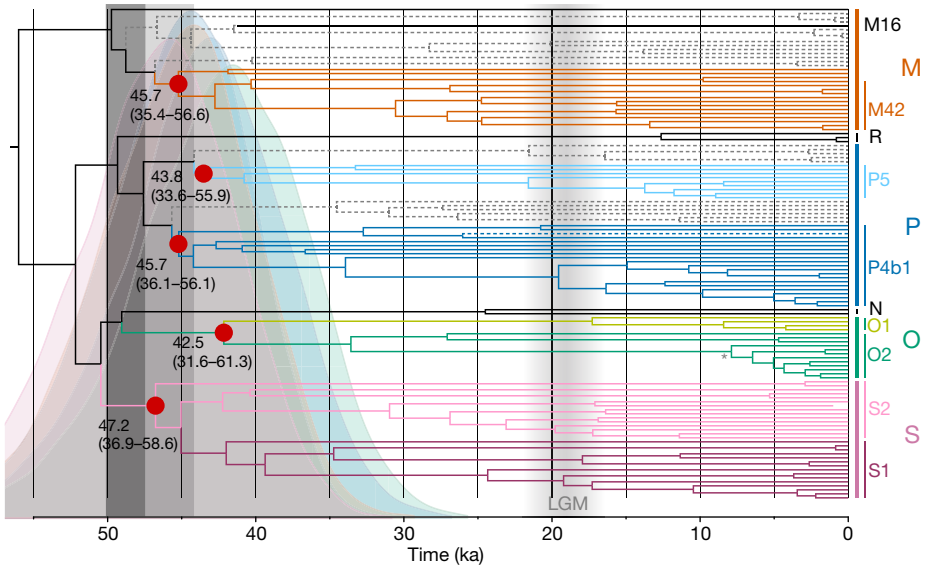


Figure 1 | Australian mtDNA phylogeny. Phylogenetic analysis of Aboriginal Australian and Melanesian (dashed grey lines) mitogenomes using BEAST[31], showing the four major haplogroups detected in Australia (in colour), along with other Aboriginal Australian lineages not used in dating analyses (solid black lines). The age of the most recent common ancestor (TMRCA) and 95% highest posterior density intervals were calculated for each Aboriginal-Australian-only clade (red dots) using human mitochondrial evolutionary rates calibrated with Palaeolithic European and Asian mitogenomes[18,32] to minimize the effects of rate temporal dependency[33,34] (see Methods). The posterior distributions for each TMRCA are shown behind the phylogeny, in matching colours. The dark grey box represents the initial colonization of Australia indicated by archaeological evidence at 48.8 ± 1.3 ka (see Methods). The light grey box indicates the period when mitochondrial lineages were still sorting into Australia or New Guinea/Melanesia, which occurred during the initial colonization of Sahul. Genetic divergences during this time (for example, between M16 and M42, or O and N) might have occurred outside Australia, and were excluded from TMRCA calculations. The short branch length of an ancient S2 sequence[14] reflects the radiocarbon-dated age of the specimen. The early Holocene diversification of lineages within haplogroup O2 is indicated with an asterisk. LGM, Last Glacial Maximum.

analysed with other full mtDNA haplotypes from Aboriginal Australians and Melanesians (44 and 25 samples, respectively, 123 unique mtDNA lineages in total).

Dating the colonization of Sahul

The timing of human arrival in Australia was estimated using the age of the most recent common ancestor (TMRCA) for the different Australian-only haplogroups, calculated using a molecular clock with substitution rates calibrated with ancient European and Asian mitogenomes[18]. Although these TMRCA values are likely to be minimal estimates given the limited sampling, they group in a narrow window of time from approximately 43–47 ka (Fig. 1 and Extended Data Figs 2, 3), consistent with previous studies (Supplementary Information). To examine the accuracy of this molecular age estimate we re-analysed a comprehensive suite of radiocarbon and optically stimulated luminescence ages from early archaeological sites across Sahul using currently available calibration datasets[19] and the phase function in OxCal 4.2.4. The resulting independent estimate for initial colonization of Sahul, 48.8 ± 1.3 ka, is a close match to the genetic age estimates (Fig. 1 and Supplementary Table 4). Indeed, the basal splits between haplogroups O, S and N13, P and R, M16 and M42 (Fig. 1) might reflect the initial within-Australia events, around 50 ka. However, we have taken a conservative approach and assumed these reflect lineages present in the initial population colonizing Sahul, as suggested by the presence of basal sister clades of Melanesian and Aboriginal Australian lineages within haplogroups M and P (Fig. 1).

Aboriginal Australian phylogeography

Phylogenetic analysis of all Aboriginal Australian samples with reliable geographical information (74 BAR samples and two from previous mtDNA studies[8,14], 76 lineages in total; see Methods), revealed large-scale phylogeographic patterns for each major haplogroup (Fig. 2). For example, none of the haplogroup O lineages were found in eastern Australia, which was dominated by haplogroups P, S and M42a. Within the two main Australian P-clades (based around P5 and P4b1) there was a clear split between northeastern and Riverine/South Australia (Fig. 2). Similar patterns are observed in the other major haplogroups, indicating that Aboriginal Australian mitochondrial lineages have undergone limited amounts of dispersal over time, and related lineages are grouped geographically. Furthermore, the basal lineages within each major haplogroup were mostly in northern Australia, presumably reflecting early divergences as members of the founding populations remained while others moved south where more derived lineages were observed. Together with the deep divergences among the mtDNA lineages, these results suggest that populations were structured by the initial major population movements following colonization around 50 ka (Fig. 1).

To verify that the small sample sizes are not biasing the phylogeographic patterns, we used a novel correlation test based on the results of a multiple correspondence analysis to examine the 76 mtDNA lineages with reliable provenance. This method is a generalization, for individual haplotypes, of the principal component analysis used for population genetic analyses of diploid genotypes. The major axes of variation among the pooled haplotype data are determined and then used to test for significant correlations with supplementary variables of interest. The test showed strong phylogeographic clustering among Aboriginal Australian mtDNA lineages, and a significant correlation between the phylogenetic structure between and within each haplogroup and both the latitudinal and longitudinal origin of the samples (Table 1 and Extended Data Table 1). As a second test for relative geographic structure, we applied a Mantel test to find correlations between pairwise distances for individuals calculated from geographic and genetic coordinates (from the multiple correspondence analysis). We also found significant correlations between these distances, both within and between haplogroups, indicating (geographically) neighbouring individuals were closely related genetically (Table 1 and Extended Data Table 1). These findings confirm that there was strong phylogeographic clustering among Aboriginal Australian mtDNA lineages before European colonization, differentiated along latitudinal and longitudinal gradients, indicating that there were very limited amounts of geographic

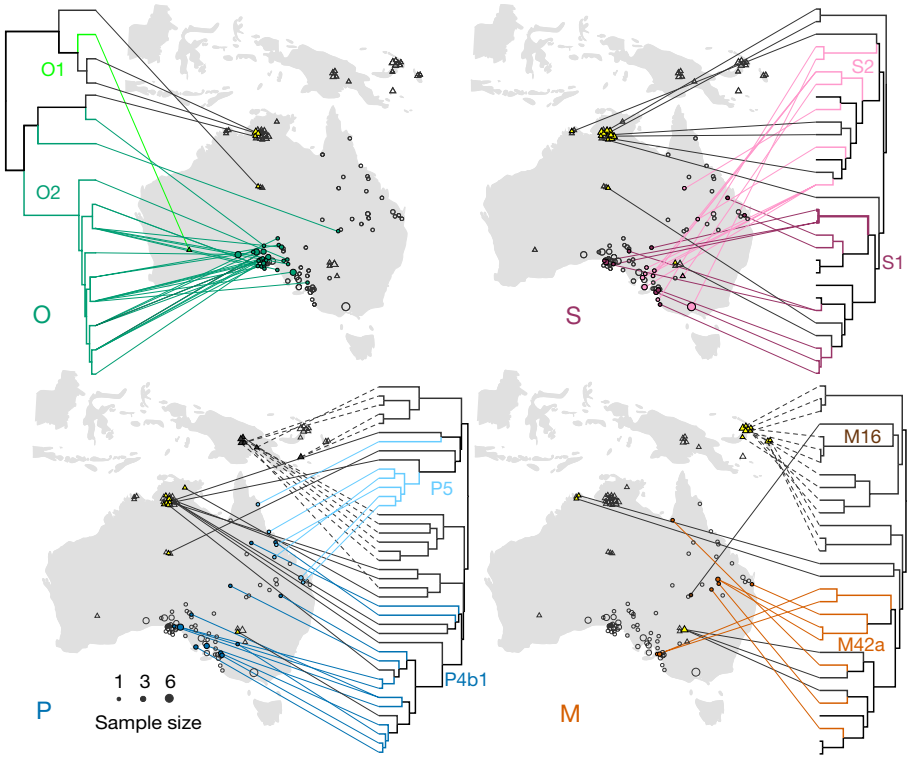


Figure 2 | Australian mtDNA phylogeography. Phylogeographic distributions of Aboriginal Australian mitogenome haplotypes, grouped into the four major haplogroups O, S, P and M with timescales calculated using an ancient-DNA-calibrated molecular clock (see Methods). Lineages from samples in the current study (circles) are shown at the location of the oldest known maternal ancestor recorded in genealogical and geographic data, generally before the effects of European colonization. Triangles represent data from modern samples reported in previous studies. The size of the symbols reflects the number of identical haplotypes as indicated in the figure. Identical sequences from the same location were pruned, whereas those from multiple locations were only used where they could not be explained through genealogical records. Coloured circles and lines represent haplotypes with known geographical provenance, with colours matching the cluster assignments of the multiple correspondence analysis (Supplementary Table 3), whereas grey (empty) circles represent the geographic distribution of samples not falling within each specific haplogroup. Previously published haplotypes that lack detailed geographic data histories are shown with yellow triangles (and black lines) for each haplogroup, whereas those with no associated locations are shown on the tree as black branches alone. Map data was sourced from the Oak Ridge National Laboratory Distributed Active Archive Center (https://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10003_1).

dispersal given the long time periods involved. Similarly, an additional set of Aboriginal Australian mtDNA genomes recently generated as part of a genomic study[12] show a concordant phylogeographic distribution to the patterns in our data (Extended Data Fig. 4). However, these sequences are not available and the samples lack information about pre-European distributions, complicating historical analysis.

Migratory patterns and regionalism within Australia

The phylogeographic distribution of the major Aboriginal haplogroups are consistent with coastal colonization models of Australia[20,21] where the initial Sahul colonizers spread across northern Australia, and then south along the east (haplogroups P, S, M42a) and west (haplogroups O, R) coasts in parallel clockwise and counter-clockwise movements (Fig. 3). The disjunction between haplogroups O and S in central southern Australia (Fig. 2) potentially reflects a meeting of the two movements. Limited genetic surveys in Tasmania are consistent with this model, because haplogroups P, S and M were detected, but not haplogroup O or R (ref. 22). A major migration corridor is also apparent between northeastern and southern Australia, potentially along the Murray–Darling River[23].

The 49–45 ka age range recently reported from Warratyi rock shelter[24], Flinders Ranges, South Australia is close in age to the earliest sites reported from northern Australia[1]. To similarly constrain the timing of human arrival in the far southwest of Australia, we re-examined the multi-dated sequence of Devil's Lair, southwestern Australia (Extended Data Fig. 5) along with continental-wide earliest occupation ages (Supplementary Table 4). The resulting age estimate (47.8 ± 1.5 ka), together with multiple early occupation sites across southern Australia (Fig. 3 and Extended Data Fig. 6) suggest the initial expansion around Australia was very rapid, perhaps taking only a few thousand years. The initial human colonization considerably preceded the extinction of the last megafauna[25], as indicated by the presence of the Diprotodont *Zygomaturus* at 42 ka just south of the Flinders Ranges[26], and this temporal overlap is similar to the pattern recently reported for South America[27].

The marked population structure of deeply diverged Aboriginal Australian mitogenomes appears to date back to the original arrival of people on the Australian part of Sahul. These patterns are surprising given the pronounced environmental changes that have occurred since

Table 1 | Australian phylogeography test results

Haplogroup	O	S	M (without M16)	P
Longitude	−0.6395 (0.0629)*	0.3351 (0.0016)***	0.642 (0.0929)*	0.7796 (0.0002)***
Latitude	0.5010 (0.0083)***	0.5977 (0.0006)***	0.8560 (0.0055)***	0.8690 (4×10^{-6})***
Mantel test	0.3352 (0.0176)**	0.2695 (0.0374)**	0.3273 (0.0953)*	0.4488 (3×10^{-6})***

Tests based on multiple correspondence analysis of phylogeographic structure within the major Aboriginal Australian haplogroups reveal significant correlations with latitude and longitude, implying lineages are likely to be found in certain geographic locations. Mantel tests confirm the lineages are grouped geographically on the landscape, implying that neighbouring individuals are expected to share common ancestry (see Methods). For each haplogroup, the correlation coefficient is given for the dimension with the most significant correlation in the case of longitude and latitude, along with the *P* value in brackets (*$P < 0.1$; **$P < 0.05$; ***$P < 0.01$). Although not every principal dimension is significantly correlated with geography, we would not expect that this is the only driver for lineage distribution.

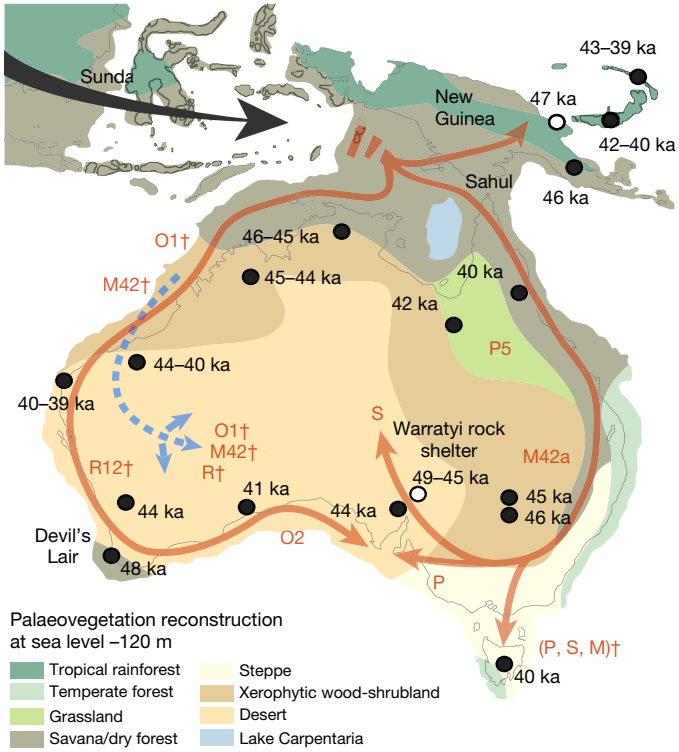


Figure 3 | The peopling of Australia. Model of the peopling of Australia combining genetic and archaeological data, showing approximate, and stylised, coastal movements of haplogroups O and R (west) and P, S, and M (east). The inferred movement of S into the interior is influenced by the path of a recent study on water sources and human movement[21]. Data from other studies where pre-European distributions are unclear are indicated with a dagger (†), and include a potential late-glacial movement into the western central desert region (blue dashed arrows; see Methods). Early archaeological sites in Australia and New Guinea (black dots) are given with mean ages for earliest occupation of sites in each region (Supplementary Table 4). Insufficient data were available for sites with white dots, which were not used in the age model for the initial Sahul colonization date but provide independent age controls. Ages in southwestern and south central Australia, at Devil's Lair (49–46 ka) and Warratyi rock shelter (49–45 ka), suggest that the overall population movements were rapid and that the coastal regions of Australia were colonized within a few thousand years. Approximate late Pleistocene vegetation reconstructions are shown (from ref. 35). The map was adapted from the figure in ref. 36, originally constructed by J.S.

initial colonization. The most extreme example of this is the widespread aridification and cooling of the Last Glacial Maximum, during which archaeological models suggest pronounced geographic contraction of populations and abandonment of large parts of the continent[28]. The diversity and grouping of Aboriginal Australian mitogenome data indicate that Aboriginal Australian populations survived these changes without large-scale movements, although there is potential evidence for a late-glacial (approximately 15 ka) re-expansion into the Western central desert (Extended Data Fig. 4 and Supplementary Information). Notably, both the diversity of mitochondrial lineages and population size estimates during this time period do not suggest severe population bottlenecks (Fig. 1 and Extended Data Fig. 7), indicating that many populations survived in local refugia that may have been cryptic to the archaeological record[29].

Holocene intensification

The rapid diversification of derived haplotypes within haplogroup O2 is indicative of a population expansion around 7 ka in southern Australia (Fig. 1), but this is the only obvious genetic signal that coincides with the mid-Holocene climatic optimum (9–6 ka) and the increasing accessibility of the arid interior to hunter–gatherer groups[13,15]. The above suggests that the extensive cultural changes evident during the Holocene, including the establishment of Panaramittee rock art, spread of the Pama–Nyungan languages, adoption of complex and diversified technologies (for example, seed grinding, wooden toolkits), advanced food-processing techniques (of, for example, *Macrozamia* plants), and greater reliance on marine resources, may have been the result of demographic change and/or cultural transmission, rather than population movement or replacement[15]. In this regard, recent archaeological models propose that rapid demographic growth during the Holocene led to reduced mobility and a consequent greater investment in technology[15]. It is also possible that some cultural changes were entirely male-mediated, and therefore not apparent in mtDNA data. Recent genomic data from modern Aboriginal Australians has been used to tentatively link the spread of the Pama–Nyungan languages to an early Holocene population expansion in northeast Australia, and limited gene flow to the rest of Australia[12]. However, the strength of the genetic signal for both the population expansion and movement remains ambiguous at best (Supplementary Information).

Discussion

The long-standing and diverse phylogeographic patterns documented here are remarkable given the timescale involved, and raise the possibility that the central cultural attachment of Aboriginal Australians to 'country' may reflect the continuous presence of populations in discrete geographic areas for up to 50 kyr. The very limited geographical movement of populations over time is consistent with observations of nomadic sedentism in recent Aboriginal Australian societies, where ranging was anchored in localized, collective and stable land/language ownership units, and occurred within a broad environmental region[17] (Supplementary Information). This form of subsistence (and territoriality) might also explain the notable lack of exchange between New Guinea and Australian mitochondrial lineages, despite a land bridge between the two until about 9 ka. Overall, these patterns are similar to recent reports of marked mitochondrial phylogeography in early South American populations[30], and raise the possibility that hunter–gatherer groups were capable of exhibiting pronounced regionalism, or at least female philopatry, over prolonged time periods.

The mitochondrial dates reported here for Aboriginal Australian arrival and dispersal appear considerably older than recent estimates from nuclear-genomic data[12] that suggest a single ancestral population started to differentiate as recently as 10–32 ka, following an admixture event with Denisovans around 43 ka. The latter event, at least, is inconsistent with the Australian archaeological record that does not support the presence of Denisovans, indicating that any admixture must have occurred before the colonization of Sahul around 50 ka. This raises the possibility that the molecular-dating analyses of the nuclear-genomic data have been confounded by complex population histories, including multiple hominin introgressions[12] and/or patterns of selection (Supplementary Information). By contrast, when combined with detailed phylogeographical data, mitogenome dating may provide a less complex alternative to reconstructing human colonization patterns in situations such as Australia.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 August 2016; accepted 24 January 2017.
Published online 8 March 2017.

1. Roberts, R. G., Jones, R. & Smith, M. A. Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. *Nature* **345,** 153–156 (1990).
2. O'Connell, J. F. & Allen, J. The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. *J. Archaeol. Sci.* **56,** 73–84 (2015).

3. Lewis, S. E., Sloss, C. R., Murray-Wallace, C. V., Woodroffe, C. D. & Smithers, S. G. Post-glacial sea-level changes around the Australian margin: a review. *Quat. Sci. Rev*. **74,** 115–138 (2013).
4. Redd, A. J. & Stoneking, M. Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations. *Am. J. Hum. Genet*. **65,** 808–828 (1999).
5. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
6. Fehren-Schmitz L. *et al.* A re-appraisal of the early Andean human remains from Lauricocha in Peru. *PloS One*. **10,** e0127141 (2015).
7. Bergström, A. *et al.* Deep roots for Aboriginal Australian Y chromosomes. *Curr. Biol*. **26,** 809–813 (2016).
8. Nagle, N. *et al.* Antiquity and diversity of aboriginal Australian Y-chromosomes. *Am. J. Phys. Anthropol*. **159,** 367–381 (2016).
9. Hudjashov, G. *et al.* Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc. Natl Acad. Sci. USA* **104,** 8726–8730 (2007).
10. van Holst Pellekaan, S. Genetic evidence for the colonization of Australia. *Quat. Int.* **285,** 44–56 (2013).
11. Heupink, T. H. *et al.* Ancient mtDNA sequences from the First Australians revisited. *Proc. Natl Acad. Sci. USA* **113,** 6892–6897 (2016).
12. Malaspinas, A.-S. *et al.* A genomic history of Aboriginal Australia. *Nature* **538,** 207–214 (2016).
13. Reeves, J. M. *et al.* Palaeoenvironmental change in tropical Australasia over the last 30,000 years—a synthesis by the OZ-INTIMATE group. *Quat. Sci. Rev*. **74,** 97–114 (2013).
14. Fitzsimmons, K. E. *et al.* Late Quaternary palaeoenvironmental change in the Australian drylands. *Quat. Sci. Rev*. **74,** 78–96 (2013).
15. Williams, A. N., Ulm, S., Turney, C. S. M., Rohde, D. & White, G. Holocene demographic changes and the emergence of complex societies in prehistoric Australia. *PLoS One* **10,** e0128661 (2015).
16. Ulm, S. 'Complexity' and the Australian continental narrative: themes in the archaeology of Holocene Australia. *Quat. Int.* **285,** 182–192 (2013).
17. Sutton, P. *Native title in Australia: an ethnographic perspective.* (Cambridge Univ. Press, 2003).
18. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol*. **23,** 553–559 (2013).
19. Reimer, P. J. *et al.* IntCal13 and marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55,** 1869–1887 (2013).
20. Bowdler, S. in *Sunda and Sahul: Prehistoric Studies in Southeast Asia, Melanesia and Australia* (eds Allen, J., Golson, J. & Jones, R.) 205–246 (Academic Press, 1977).
21. Bird, M. I., O'Grady, D. & Ulm, S. Humans, water, and the colonization of Australia. *Proc. Natl Acad. Sci. USA* **113,** 11477–11482 (2016).
22. Mcallister, P., Nagle, N. & Mitchell, R. J. Brief communication: the Australian Barrineans and their relationship to Southeast Asian negritos: an investigation using mitochondrial genomics. *Hum. Biol*. **85,** 485–502 (2013).
23. White, J. P. & O'Connell, J. F. *A Prehistory of Australia, New Guinea, and Sahul.* (Academic Press, 1982).
24. Hamm, G. *et al.* Cultural innovation and megafauna interaction in the early settlement of arid Australia. *Nature* **539,** 280–283 (2016).
25. Saltré, F. *et al.* Climate change not to blame for late Quaternary megafauna extinctions in Australia. *Nat. Commun*. **7,** 10511 (2016).
26. Roberts, R. G. *et al.* New ages for the last Australian megafauna: continent-wide extinction about 46,000 years ago. *Science* **292,** 1888–1892 (2001).
27. Metcalf, J. L. *et al.* Synergistic roles of climate warming and human occupation in Patagonian megafaunal extinctions during the last deglaciation. *Sci. Adv*. **2,** e1501682 (2016).
28. Williams, A. N., Ulm, S., Cook, A. R., Langley, M. C. & Collard, M. Human refugia in Australia during the Last Glacial Maximum and terminal Pleistocene: a geospatial analysis of the 25–12ka Australian archaeological record. *J. Archaeol. Sci.* **40,** 4612–4625 (2013).
29. Smith, M. *The Archaeology of Australia's Deserts* (Cambridge Univ. Press, 2013).
30. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv*. **2,** e1501385 (2016).
31. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol*. **7,** 214 (2007).
32. Posth, C. *et al.* Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr. Biol*. **26,** 827–833 (2016).
33. Ho, S. Y. W. *et al.* Time-dependent rates of molecular evolution. *Mol. Ecol*. **20,** 3087–3101 (2011).
34. Rieux, A. & Balloux, F. Inferences from tip-calibrated phylogenies: a review and a practical guide. *Mol. Ecol*. **25,** 1911–1924 (2016).
35. Balme, J., Davidson, I., McDonald, J., Stern, N. & Veth, P. Symbolic behaviour and the peopling of the southern arc route to Australia. *Quat. Int*. **202,** 59–68 (2009).
36. Cooper, A. & Stringer, C. B. Paleontology. Did the Denisovans cross Wallace's Line? *Science* **342,** 321–323 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge the support and involvement of the Point Pearce, Cherbourg and Koonibba communities and the individual families. We also acknowledge the work of N. Tindale, J. Birdsell and members of the original Board for Archaeological Research expeditions collecting the specimens. We thank the South Australian Museum, Australian Research Council, University of Adelaide Environment Institute, the Genographic Project and Bioplatforms Australia for support, and S. Ulm, G. Gower, I. Mathieson, L. O'Brien, S. Easteal, M. Vilar, C. Stringer and ACAD colleagues for helpful comments and advice. The Aboriginal Heritage Project webpage is https://www.adelaide.edu.au/acad/ahp/, and this work was carried out under the auspices of the University of Adelaide Human Research Ethics Committee, project approval H-2014-252.

Author Contributions The project was conceived by A.C., W.H. and P.S. and directed by A.C. and W.H. Archival research and community outreach was led by I.O., A.A-H., S.A., A.O., F.Z. and L.W. with A.C., W.H., R.T. and R.J.M. The genetic sequencing was performed and coordinated by W.H., P.B., M.W., S.R. and J.R.S., and the genetic analysis by W.H., R.T., A.R., J.S., J.T., N.B., B.L. and A.C. Archaeological and anthropological interpretations were provided by P.S., C.T., A.N.W. and K.W. The manuscript was written by A.C. and R.T., with critical input from P.S., C.T., A.N.W., A.R., J.S., W.H. and all other co-authors. R.T., J.S., A.N.W. and A.R. compiled the Supplementary Information.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.C. (alan.cooper@adelaide.edu.au).

Reviewer Information *Nature* thanks P. Bellwood, C. Lalueza-Fox and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Samples. The 111 hair samples used in the present study were originally collected during anthropological expeditions to one of the following communities: Cherbourg, Queensland (23 samples), Point Pearce, South Australia (41 samples) and Koonibba, South Australia (47 samples) (Extended Data Fig. 1 and Supplementary Table 1). Consent was obtained from the original donors, or their descendants, according to protocols detailed in the Supplementary Information. Six of the Koonibba samples were collected during an expedition to the area between 13 and 25 August 1928, all remaining samples were obtained from the extensive Harvard and Adelaide Universities Anthropological Expeditions lead by N. B. Tindale and J. B. Birdsell that took place from 13 May 1938 to 30 June 1939. Hair was collected from different parts of the body, but all samples used in the current study consist of small locks of hair that were cut with permission from the head of participants. Since the initial collection date, the hair samples have been stored in sealed paper envelopes. The envelopes are currently secured in a restricted-access storage room maintained by the South Australian Museum. For each sample, a portion of the hair (between 20–190 mg) was removed from each envelope for use in the present study.

Ancient DNA analysis. The hair samples from Cherbourg and Point Pearce were soaked in 3.5 ml of 1% bleach, rinsed in 7 ml of water, and subsequently 3.5 ml of 100% ethanol and before being air-dried. For the Koonibba samples, we applied 2 washes in 3 ml of water, a subsequent wash in 3 ml of 100% ethanol, followed by air-drying. Each sample was digested for 1 h under constant rotation at 55 °C in 4 ml of a digestion buffer containing 75 mM Tris pH 8.0, 50 mM NaCl (Sigma-Aldrich), 0.5 mg ml^{-1} Proteinase K (Life Technologies), 50 mM DTT (Promega) and 0.75% SDS (Life Technologies). After lysis, samples were centrifuged at 4,600 r.p.m. for 1 min and the supernatant was pipetted into 100 μl silica suspension and 16 ml modified binding buffer (90% QG Buffer (Qiagen), 1.3% Triton X-100 (Sigma-Aldrich), 25 mM NaCl (Sigma-Aldrich) and 0.2 M sodium acetate (Sigma-Aldrich)), and left for 1 h at room temperature under constant rotation. Silica suspensions subsequently pelleted using a centrifuge at 4,600 r.p.m. for 5 min, and the supernatant was discarded. The silica pellet was washed three times in 80% ethanol and centrifugation at 13,000 r.p.m. for 1 min. After the last wash, the pellet was air dried for 30 min and resuspended twice in 120 μl of a pre-warmed (at 50 °C) mix of EB buffer (Qiagen) and 0.05% Tween 20, and incubated for 10 min. After centrifugation at 13,000 r.p.m. for 1 min, a final 240 μl extract was obtained. Subsequently, 60 μl extract was purified using a MinElute Reaction Cleanup Kit (Qiagen) following the manufacturer's protocol.

Double-stranded libraries were prepared following standard protocols[30,37,38], using short Illumina adapters with dual 5-mer (non-Koonibba samples) or 7-mer (Koonibba samples) internal barcodes. For the Koonibba samples partial uracil-DNA-glycosylase (UDG) treatment[39] was performed for DNA repair in the first step of library construction. Libraries for the Koonibba sample extracts were amplified using Platinum Taq HiFi (Invitrogen), whereas the Cherbourg and Point Pearce samples were amplified using isothermal amplification (TwistAmp Basic kit, TwistDx Ltd). The latter were enriched by hybridization using mitochondrial RNA baits prepared in-house and finally amplified using full-length 7-mer indexed Illumina adapters (see ref. 6 for a full explanation of the protocol). Libraries were pooled and sequenced in a HiSeq 2×100 PE run. The Koonibba libraries were amplified using full-length 7-mer indexed Illumina adapters and shotgun sequenced in MiSeq (2×150 PE) and NextSeq (2×150 PE) Mid Output runs at the Australian Genome Research Facility.

Mapping and consensus calling. Raw Illumina reads were processed using the PaleoMix v1.0.1[40] pipeline. AdapterRemoval v2 (ref. 41) was used to trim adaptor sequences, merge the paired reads, and eliminate all reads shorter than 25 bp. Filtered reads were then mapped to the Reconstructed Sapiens Reference Sequence (RSRS) mitochondrial reference genome[42] with BWA v0.6.2 (ref. 43). The minimum mapping quality was set to 25, seeding was disabled and the maximum number or fraction of open gaps was set to 2. MapDamage v2 (ref. 44) was used to check that the expected mapping and damage patterns were observed for each library and re-scale base qualities for the non-repaired libraries (see Supplementary Table 2 for library statistics).

All mtDNA genome consensus sequences were called using Geneious v9.1.3 (ref. 45). For each sample, reads were remapped to the RSRS reference using the Geneious mapper (default settings, serial mapping iterated five times). To call a base, each region required a coverage ≥3, with a majority allele frequency ≥0.75. The resulting consensus sequences were then inspected by eye, with particular attention being paid to the hypervariable regions and nucleotide positions previously identified as being problematic on the phylotree website (http://www.phylotree.org/)[46]. All ambiguous sites were called as 'N'.

Identical haplotypes were collapsed into a single haplotype sequence. Individuals with genealogical information that indicated a shared common maternal ancestor were checked for sequence similarity, and were identical in all but two cases where they differed by a single nucleotide. These cases were subsequently maintained as separate mtDNA haplotypes. For all individuals where identity by maternal descent was unknown, two sequences were deemed as identical if their sequences shared all diagnostic variants for a given haplogroup. After combining all common haplotypes, a total of 54 non-redundant consensus sequences were determined (from 111 original samples; Supplementary Table 1). The resulting consensus haplotypes cover all the major mtDNA haplogroups previously described for Australia (Supplementary Information).

Phylogenetics. To help determine the timing of the split between Melanesian and Australian populations, and the colonization history of Australia, the phylogenetic software BEAST (v1.8.3)[31,47] was used on 123 complete (or mostly complete) mtDNA genomes (54 unique Aboriginal Heritage Project (AHP) consensus samples combined with 44 Australian and 25 Melanesian publicly available sequences; see Supplementary Table 1). The non-AHP sequences were obtained from the mitochondrial database mtDB[48] and two recently published papers[5,11]. Before analysis, all 123 mtDNA genomes were aligned to the RSRS with BLAT[49] and then analysed with a custom R script, so that indels were removed and only point mutations relative to the RSRS were used in the subsequent analyses.

The TN93+G6 model of nucleotide substitution was selected through comparison of BIC scores using ModelGenerator v0.85 (ref. 50), a GMRF skyride model[51] was used to allow for a complex population history, with a relaxed uncorrelated log-normal clock[52] to account for rate heterogeneity between lineages (a strict clock was empirically rejected as ucld.stdev posterior distribution did not include zero). Monophyly was constrained for all major haplogroups and the ancient sequence hap97 was given a tip date log-normal prior distribution with a mean of 1,250 years and a standard deviation of 0.7 (95% of the dates fall between 500 and 3,000 years; based on estimates from ref. 11). Two mutation rates with normally distributed priors were applied, using the values from ref. 18 (mean $= 2.67 \times 10^{-8}$ substitutions per site per year, s.d. $= 2.6 \times 10^{-9}$) and from ref. 32 (mean $= 2.74 \times 10^{-8}$ substitutions per site per year, s.d. $= 2 \times 10^{-9}$). These two rate estimates were chosen as they both use state of the art tip-dating calibration methods to infer mutation rates, thereby providing inferences that minimise the effects of rate temporal dependency on late Palaeolithic events[33,34]. In particular, the mutation-rate estimates reported in refs 18,32 are based on 10 and 66 radiocarbon-dated ancient sequences, respectively. Notably, the calibration dates for these ancient sequences are distributed across 46,000–4,000 ka and cover both haplogroups M and N, a scenario that is well–suited for comparison with Australia, both in terms of temporal coverage and mtDNA diversity. Separate BEAST phylogenies were inferred for the combined set of Melanesian and Australian lineages using the mutation rate from ref. 18 (Fig. 1 and Extended Data Fig. 2) and ref. 32 (Extended Data Fig. 3). A phylogeny based on Australian lineages only was also inferred using the mutation rate from ref. 18 and used to determine the palaeodemography of Australia (Extended Data Fig. 7).

All parameters showed sufficient sampling (indicated by effective sample sizes above 200) after 20,000,000 steps, with the first 10% of samples discarded as burn-in. Notably, the two different mutations rates produced TMRCA estimates for the major haplogroups within 1.5 kyr of each other (Extended Data Figs 2, 3), with posterior mutation-rate estimates that were also highly similar (mean rate $= 2.70 \times 10^{-8}$ (ref. 18), mean rate $= 2.74 \times 10^{-8}$ (ref. 32)), indicating that the choice of prior distribution for the mutation rate had little effect on our dating.

Multiple correspondence analyses. A useful tool for detecting and analysing demographic structure in genetic data is principal components analysis (PCA)[53]. When working with non-autosomal data, PCA cannot be applied to any (satisfactory) recoding of sequence data (unless it is manually, that is, subjectively, sorted into haplogroups). Multiple correspondence analysis (MCA) is an analysis technique for data exploration and dimension reduction for categorical data. MCA is a generalization of PCA to categorical variables and can therefore be applied to raw sequence data. MCA has been independently rediscovered many times since its original development, and as such can also be found under titles including 'optimal scaling', 'dual scaling' and 'homogeneity analysis'[54]. MCA was originally developed for the analysis of survey data, so that responses that were commonly (or rarely) reported together could be efficiently identified. We apply the same notion but treat single nucleotide polymorphisms (SNPs) as survey questions, and observed SNP markers as responses.

We restricted the MCA to AHP samples and two Australian mtDNA haplotypes derived from ancient samples whose origin was assumed to be the area in which the specimen was collected[5,11] (Supplementary Table 3). Unfortunately, we have been unable to obtain the mtDNA data from a recent Aboriginal genomics study[12] to use in the MCA analyses, although these samples may have had limited utility for phylogeographic analysis given the large-scale relocation of Aboriginal Australians after European arrival. However, we have included the reported sample locations and mtDNA lineages in geographic plots to examine the consistency with our results (Extended Data Fig. 4). For the AHP samples, geographic locations were

determined for each individual using the relevant genealogies to trace maternal ancestry as far back as the archival information allowed. Importantly, the broad distribution of the female ancestors for the AHP samples collected from each of the three sampling locations (Extended Data Fig. 1) reflects the forced relocation of Aboriginal Australians from their traditional territories, and highlights the difficulties associated with obtaining valid phylogeographic information using only modern samples.

Identical samples were treated separately if they came from different geographical locations, as these most likely represented more distant family relationships not captured in the genealogical information. This resulted in 76 unique sequences (Supplementary Table 3). Restricting the analyses to these samples ensured that the underlying phylogeographic signal was not diluted by the addition of sequences from modern individuals that are likely to have been affected by forced-displacement or child-removal policies and typically lack genealogical information. Independent MCA analyses were run for all samples combined and for each haplogroup separately. We excluded the M16 lineage from the M haplogroup tests, because this was a deeply divergent Australian lineage that clusters among Melanesian samples and thereby most likely represents a pre-Sahul split (Fig. 1).

We cleaned the aligned sequence data by removing any homogeneous (uninformative) sites, and any sites containing missing data. Unlike PCA analyses, we are not forced to filter out triallelic SNPs and thereby can retain the information contained within these sites[53]. For M sequences in an alignment, the MCA analysis will return $M-1$ principal dimensions of length $J-Q$, where Q is the number of cleaned SNPs of interest, and J is given by,

$$J=\sum_{i=1}^{Q} J_i$$

where J_i is the number of alleles observed at SNP i, for $i=1, \ldots, Q$. These principal dimensions are analogous to the principal components returned from PCA analyses, and the dimensions are ordered by the amount of inertia (analogous to variability in PCA) that they explain. Dimensions with associated eigenvalues less than $1/Q$ are discarded as they explain less variation than expected (analogous to the threshold of 1 for the eigenvalues in PCA)[55]. The retained coordinates are then used for the visualization of the relationships between individuals, investigation of correlation between the dimensions and geographic variables, and clustering for genetic similarity. We carried out our MCA analysis using the FactoMineR package[56].

Clustering via *k* medoids. Identifying points in n-dimensional space based on similarity inferred through Euclidean distance is not a new problem. By far the most popular clustering algorithm is the k-means clustering algorithm[57]. We used the closely related k-medoids algorithm instead, which, instead of using a centroid for each cluster, forces one of the observed data points to be the centre of the cluster. In doing so, the inter- and intra-cluster distances are more robust to noise and outliers[58]. We consider an exhaustive range of values for k, and a 'best' number of clusters must be chosen. Unlike the possibly subjective 'elbow method', used in PCA through scree plots, we instead calculate $\bar{s}_k$, called the 'average silhouette'[59], for each value of k. The value of k that maximises $\bar{s}_k$ is chosen. However to avoid 'over-fitting' the number of clusters, we apply a leave-one-out jack-knife approach to both identify if influential individuals exist in the data and to obtain some measure of variability for the values of $\bar{s}_k$. We carried out our clustering methodology using the cluster package[60], in the R statistical programming language[61].

Testing for correlation. We tested for geographic correlation through two methods that seem similar, but are subtly different in their interpretation. First we applied the Mantel test, which is a test for correlation between two distance matrices[62]. One distance matrix contains the pairwise Euclidean distances between individuals with respect to their geographic location, and the second distance matrix contains the genetic distances, calculated from the coordinates of the MCA. The null hypothesis is that there is a perfectly mixed population (that is, pan-mixia), so that rejection of the null hypothesis indicates some genetic clustering on the landscape. At the cost of statistical power, we use the Spearman correlation coefficient, as it is unreasonable to assume strictly linear relationships. We used 10^5 permutations for each test. Second, we calculated the correlation between the longitude and latitude of individuals, and the retained principal dimensions. We perform a standard test for correlation under Spearman's ρ (for the same reasons indicated in the Mantel test). All tests of correlation use the AS 89 algorithm for calculation of P values[63].

Although these tests may appear similar, the Mantel test is a relative test of geographic correlation that simply tests for some clustering with respect to local geographic location. In essence, the Mantel test investigates if certain combinations of SNP markers are often found within close proximity. With the test of significant correlations between the principal dimensions and either longitude or latitude, not only distance, but also direction is important. Hence, the correlation tests are absolute tests for identifying if combinations of certain SNP markers can be linked to certain geographical locations on the landscape, with respect to the entire sampling region. We performed the Mantel test using the vegan package[64], and the standard correlation tests in the R statistical programming language[61]. The full list of P values from Spearman's correlation and the Mantel tests are shown in Extended Data Table 1.

The Australian archaeological record. *Devil's Lair, southwest Australia.* To more precisely constrain the time of arrival of modern humans in southwest Australia, we analysed a comprehensive multi-dating suite of ages for Devil's Lair, one of the earliest archaeological sites in southwestern Australia[65]. The dates comprise radiocarbon dating (pretreated using acid–base–acid or ABA, and acid–base–acid stepped combustion, or ABOX–SC, pretreatment), optically stimulated luminescence, electron-spin resonance (derived using an early uptake model) and U-series dating. Devil's Lair (34° 9′ S, 115° 4′ E) is a single-chamber cave (floor area 200 m^2) formed in the Quaternary dune limestone of the Leeuwin–Naturaliste Ridge, 5 km from the modern coastline and approximately 250 km south of Perth (Western Australia). Archaeological investigation over the past four decades has identified a stratigraphic sequence in the cave floor deposit that consists of 660 cm of sandy sediments, with >100 distinct layers, intercalated with flowstone and other indurated deposits[65–67]. Archaeological evidence for intermittent human occupation extends down to layer 30 (around 350 cm depth), with hearths, bone and stone artefacts found throughout. The lower part of layer 30 represents a fan of redeposited topsoil that accumulated rapidly after widening of the cave mouth, and contains the earliest evidence for occupation of the cave. Below layer 30, six stone artefacts have been identified, including a single specimen each from layers 32–35, 37 and 38. No artefacts have been found below layer 38.

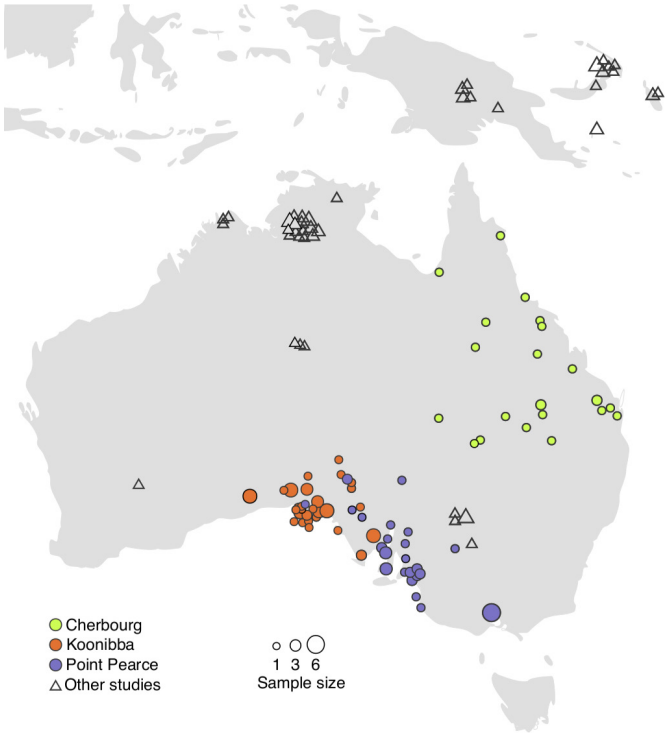
The age model was created with OxCal v.4.2.4 using a Poisson process deposition model (P_sequence)[68] with the 'general outlier' analysis option[69] of all ages as reported in ref. 65. The outlier option was used to detect ages that fall outside the calibration model for the sequence and, if necessary, down-weight their contribution to the final age estimates. Radiocarbon ages were calibrated using the SHCal13 calibration dataset[70]. Taking into account the deposition model and the actual age measurements, the posterior probability densities quantify the most probable age distributions. Notably, the lowest artefact in the sequence is constrained by age estimates obtained using all four dating techniques (but excluding the ABA radiocarbon (^{14}C) ages, which reached background levels around 40 ka)[65], providing confidence in the calculated age for this level. Using this approach we derive an age for layer 30 (lower) for cave occupation of 47.1 ± 0.8 ka and the lowest artefact (layer 38) of 49.5 ± 1.1 ka (Extended Data Fig. 5).

Early colonization of Australia. We extended this approach across Australia, and examined radiocarbon and optically stimulated luminescence ages associated with the lowest cultural horizons in early Australian archaeological sites (Extended Data Fig. 6 and Supplementary Table 4) to estimate the timing of colonization across the continent. Here we used the Phase model option in OxCal v.4.2.4 (ref. 68) with general outlier analysis detection (probability = 0.05)[69]. Notably, the Phase option is a grouping model which assumes no geographic relationship between samples (in contrast to the P_sequence used above, which assumes a stratigraphic relationship between dated levels). The model simply assumes that the ages represent a uniform distribution between a start and end boundary[68]. Terrestrial samples were calibrated using the SHCal13 dataset[70]; marine ages were converted to calendar ages using the Marine13 calibration dataset[19] and corrected for regional ΔR (marine reservoir age) with reported values for Papua New Guinea (372 ± 64 years)[71] and the east Indian Ocean (43 ± 81 years)[72,73]. Using this approach, and incorporating the age calculated above from Devil's Lair, we derive an age estimate for human arrival in Australia (the start of continental occupation) as 48.8 ± 1.3 ka (Extended Data Fig. 3). Notably, this age estimate includes the luminescence-dated Northern Territory sites of Malakunanja II and Nauwalabila I (ref. 74–76), which are statistically indistinguishable from the timing of occupation continent-wide. Our estimated timing of human arrival is consistent with the minimum age obtained from the Huon Peninsula[77,78] and the recently reported ages obtained from Warratyi Rockshelter in the Flinders Ranges[24].

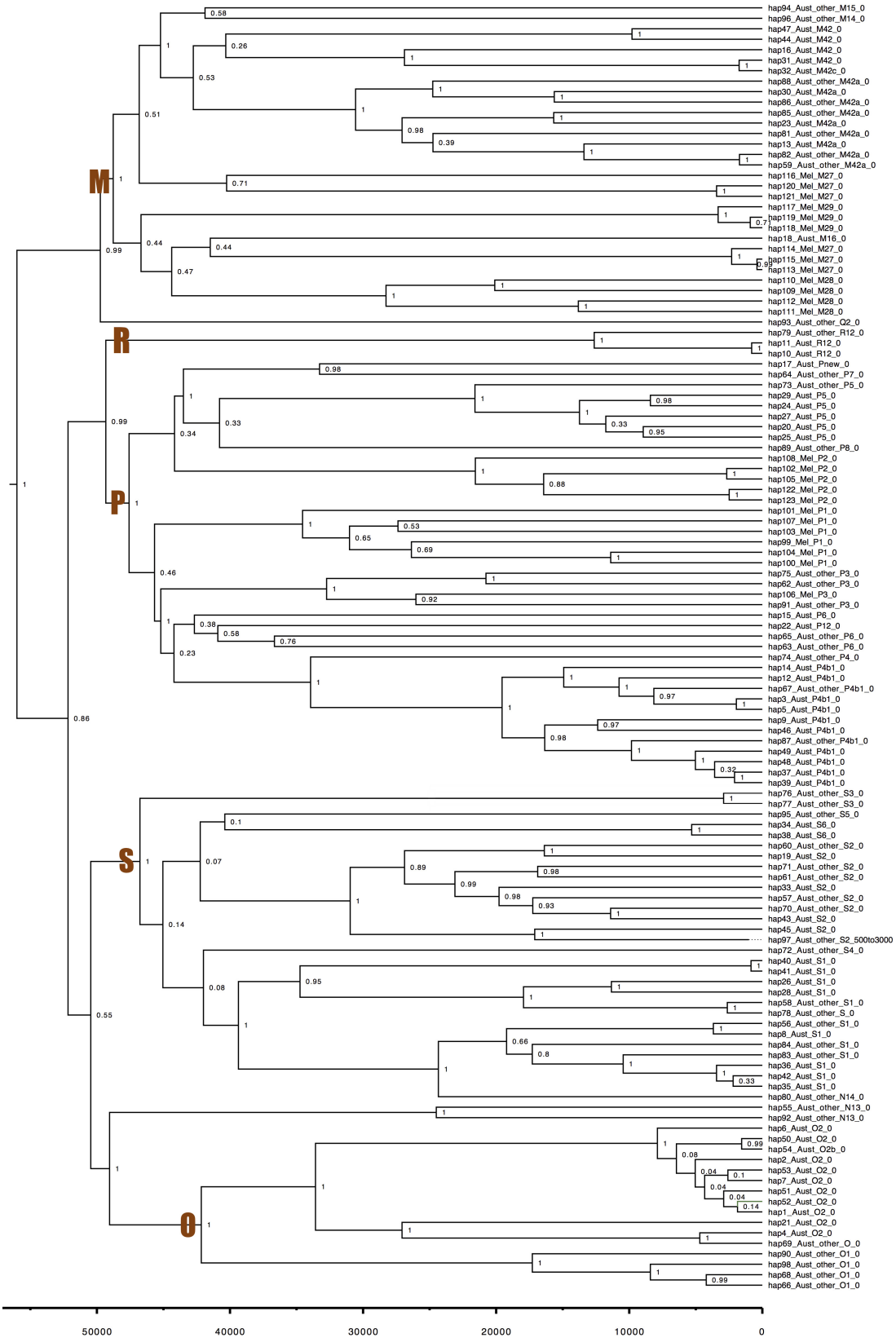
Data availability. The datasets generated and analysed during the current study are available in the European Nucleotide Archive repository, and are accessible through accession number PRJEB15344. Additional data related to this paper may be requested from the authors.

37. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448 (2010).
38. Knapp, M., Stiller, M. & Meyer, M. Generating barcoded libraries for multiplex high-throughput sequencing. *Methods Mol. Biol.* **840,** 155–170 (2012).
39. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Phil. Trans. R. Soc. Lond. B* **370,** 20130624 (2015).

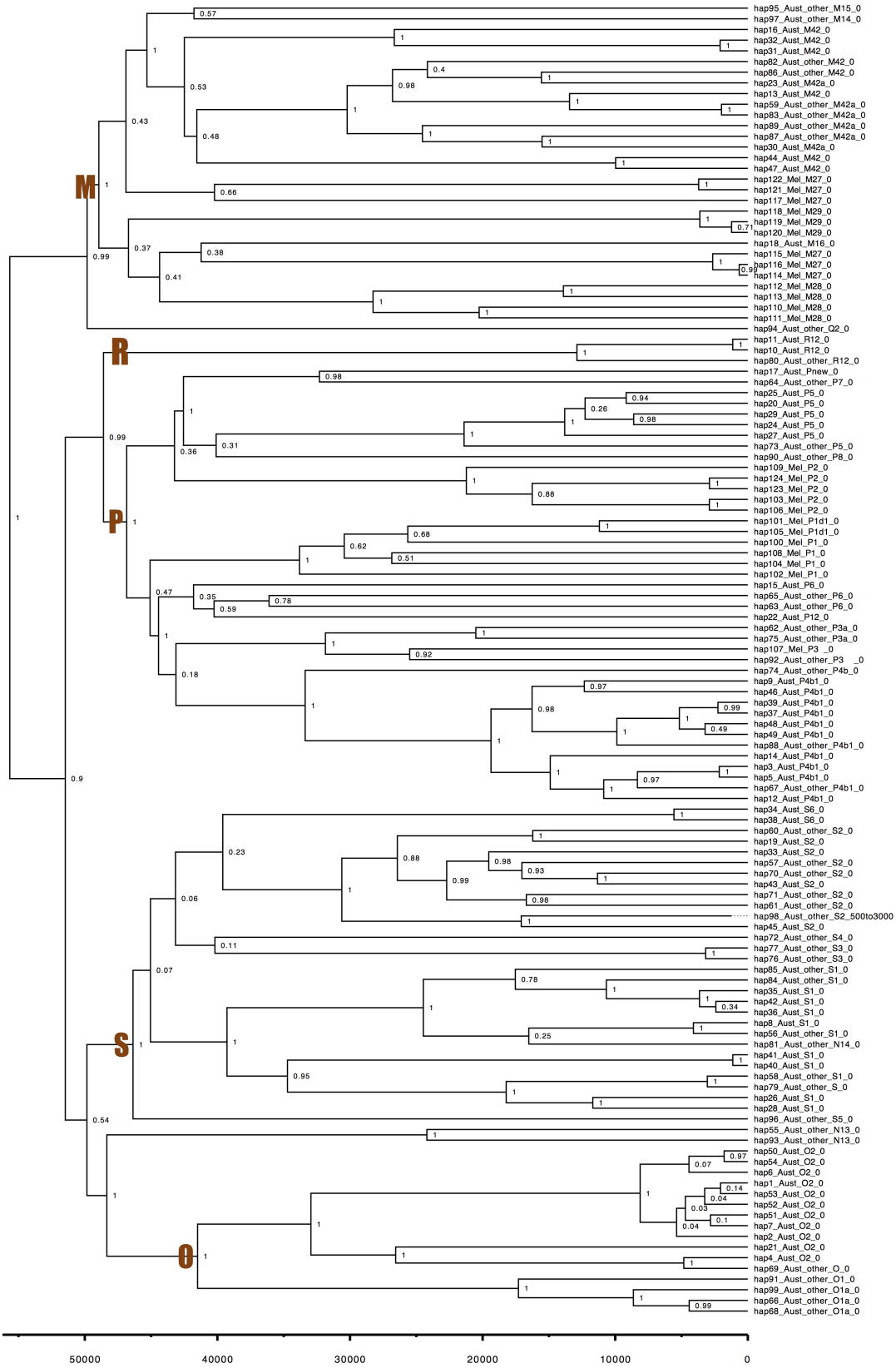
40. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protocols* **9,** 1056–1082 (2014).
41. Lindgreen, S. AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC Res. Notes* **5,** 337 (2012).
42. Behar, D. M. *et al.* A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.* **90,** 675–684 (2012).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
44. 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).
45. Kearse, M. *et al.* Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
46. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).
47. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol*. **29,** 1969–1973 (2012).
48. Ingman, M. & Gyllensten, U. mtDB: Human mitochondrial genome database, a resource for population genetics and medical sciences. *Nucleic Acids Res.* **34,** D749–D751 (2006).
49. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
50. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. 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 Evol. Biol*. **6,** 29 (2006).
51. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol*. **25,** 1459–1471 (2008).
52. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. **4,** e88 (2006).
53. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
54. Abdi, H. & Valentin, D. in *Encyclopedia of Measurement and Statistics* (ed. Salkind, N.) 651–657 (Thousand Oaks, 2007).
55. Greenacre, M. *Correspondence Analysis in Practice.* (CRC Press, 2007).
56. Lê, S., Josse, J. & Husson, F. FactoMineR : an R package for multivariate analysis. *J. Stat. Softw.* **25,** 1–18 (2008).
57. Lloyd, S. P. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28,** 129–137 (1982).
58. Kaufman, L. & Rousseeuw, P. J. Clustering by means of medoids. *Statistical Data Analysis Based on the L 1-Norm and Related Methods. First International Conference* 405–416416 (1987).
59. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20,** 53–65 (1987).
60. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. Cluster Analysis Basics and Extensions. R package version 2.0.4. *CRAN* (2016).
61. R Development Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria* http://www.R-project.org (2013).
62. Legendre, P. & Fortin, M. J. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Mol. Ecol. Resour*. **10,** 831–844 (2010).
63. Best, D. J. & Roberts, D. E. Algorithm AS 89: The upper tail probabilities of Spearman's *Rho*. *Appl. Stat.* **24,** 377–379 (1975).
64. Dixon, P. VEGAN, a package of R functions for community ecology. *J. Veg. Sci*. **14,** 927–930 (2003).
65. Turney, C. S. M. *et al.* Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. *Quat. Res.* **55,** 3–13 (2001).
66. Dortch, C. E. & Dortch, J. Review of Devil's Lair artefact classification and radiocarbon chronology. *Aust. Archaeol*. **43,** 28–32 (1996).
67. Dortch, C. Devil's Lair, an example of prolonged cave use in South-Western Australia. *World Archaeol*. **10,** 258–279 (1979).
68. Bronk Ramsey, C. & Lee, S. Recent and planned developments of the program OxCal. *Radiocarbon* **55,** 720–730 (2013).
69. Bronk Ramsey, C. Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon* **57,** 1023–1045 (2009).
70. Hogg, A. *et al.* SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP. *Radiocarbon* **55,** 1889–1903 (2013).
71. Petchey, F., Phelan, M. & White, J. P. New ΔR values for the southwest Pacific Ocean. *Radiocarbon* **46,** 1005–1014 (2004).
72. O'Connor, S., Ulm, S., Fallon, S. J., Barham, A. & Loch, I. Pre-bomb marine reservoir variability in the Kimberley region, Western Australia. *Radiocarbon* **52,** 1158–1165 (2010).
73. Bowman, G. M. Oceanic reservoir correction for marine radiocarbon dates from northwestern Australia. *Aust. Archaeol*. **20,** 58–67 (1985).
74. Roberts, R. G. *et al.* The human colonisation of Australia: optical dates of 53,000 and 60,000 years bracket human arrival at Deaf Adder Gorge, Northern Territory. *Quat. Sci. Rev*. **13,** 575–583 (1994).
75. Roberts, R. G. *et al.* Single-aliquot and single-grain optical dating confirm thermoluminescence age estimates at Malakunanja II rock shelter in northern Australia. *Anc. TL* **16,** 19–24 (1998).
76. Bird, M. I. *et al.* Radiocarbon dating of organic- and carbonate-carbon in *Genyornis* and *Dromaius* eggshell using stepped combustion and stepped acidification. *Quat. Sci. Rev*. **22,** 1805–1812 (2003).
77. Groube, L., Chappell, J., Muke, J. & Price, D. A 40,000 year-old human occupation site at Huon Peninsula, Papua New Guinea. *Nature* **324,** 453–455 (1986).
78. Roberts, R. G. Luminescence dating in archaeology: from origins to optical. *Radiat. Meas*. **27,** 819–892 (1997).



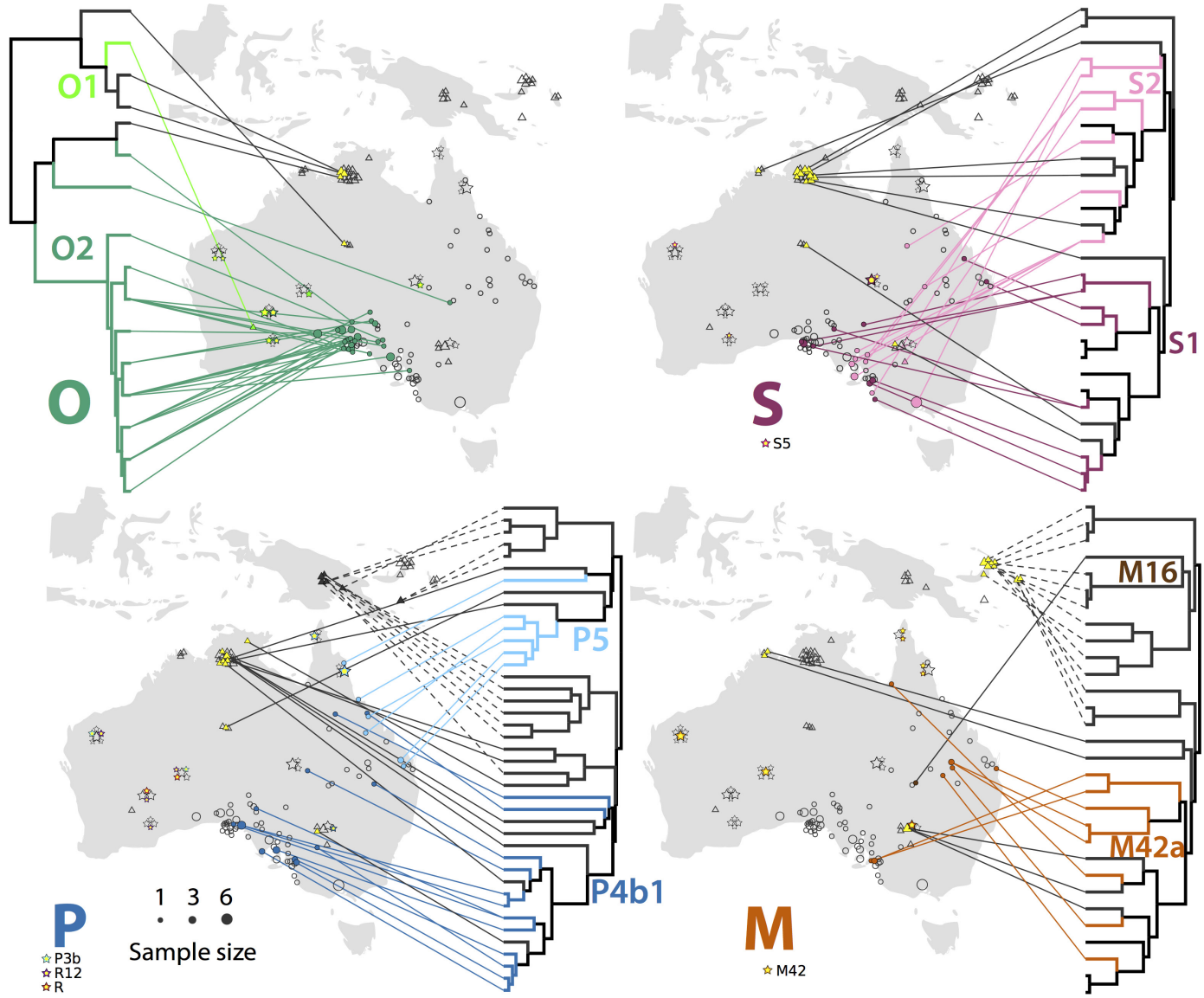
Extended Data Figure 1 | The geographical distribution of the oldest recorded maternal ancestors for the hair sample donors. Despite being collected from three different historical locations—Cherbourg (Queensland), Point Pearce and Koonibba (both South Australia)—the broad distribution of the maternal ancestors of the hair sample donors demonstrates the massive displacement experienced by Aboriginal Australians after European colonization. This pattern illustrates why the accurate reconstruction of Aboriginal Australian genetic history ultimately relies upon samples or genealogical records that capture patterns prior to this displacement. Map data was sourced from the Oak Ridge National Laboratory Distributed Active Archive Center (https://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10003_1).



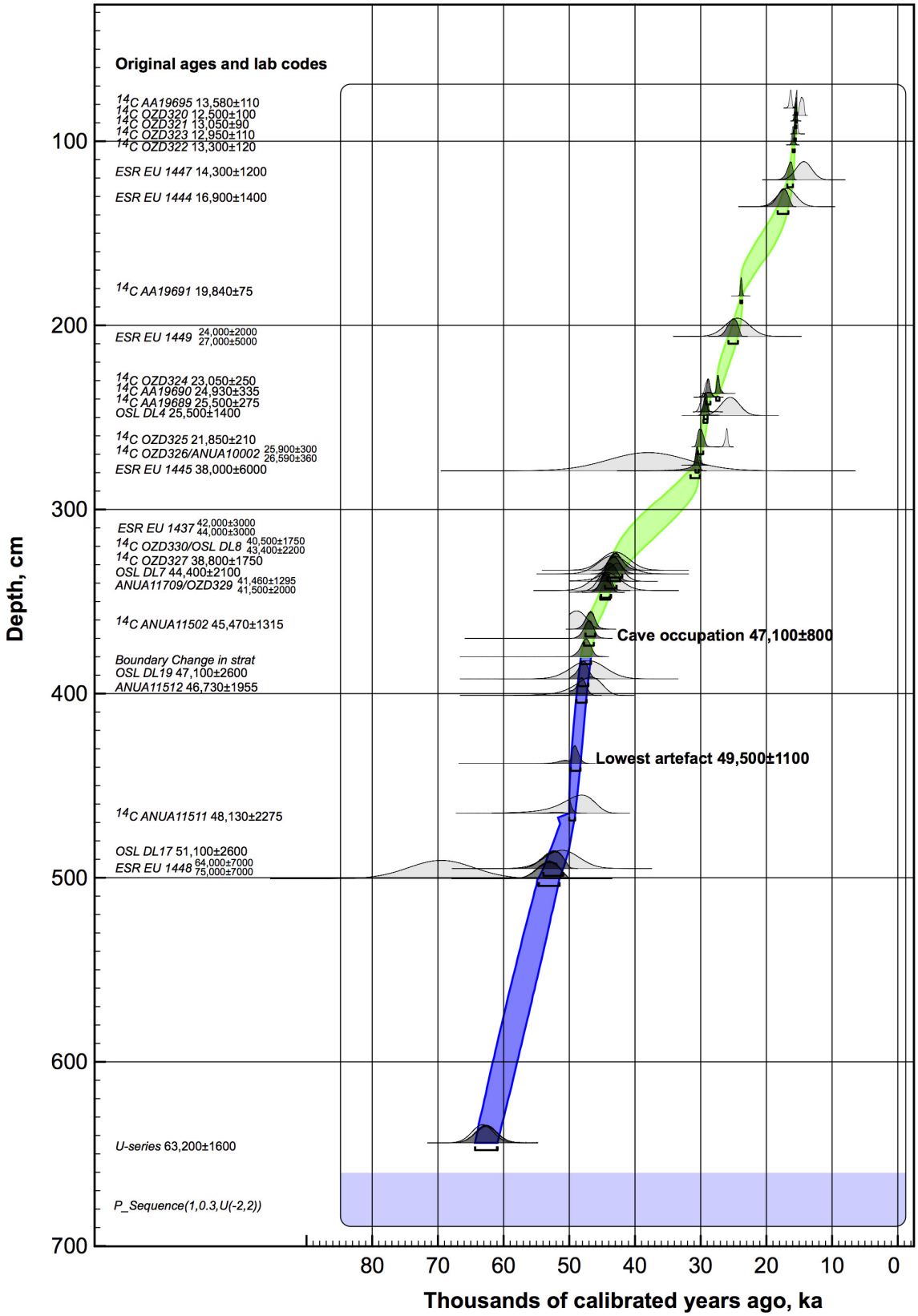
Extended Data Figure 2 | Sahul phylogenetic tree calibrated using the mitogenome rate from ref. 18. BEAST[31] phylogenetic tree of 123 Australian and Melanesian mtDNA lineages, which was calibrated using the ancient mitogenome rate in ref. 18 to minimize the impacts of temporal dependency[33,34] and improve estimation of the timing of the founding migrations. The major mitogenome haplogroups are shown at the base of each clade, and posterior support values are provided for all nodes.



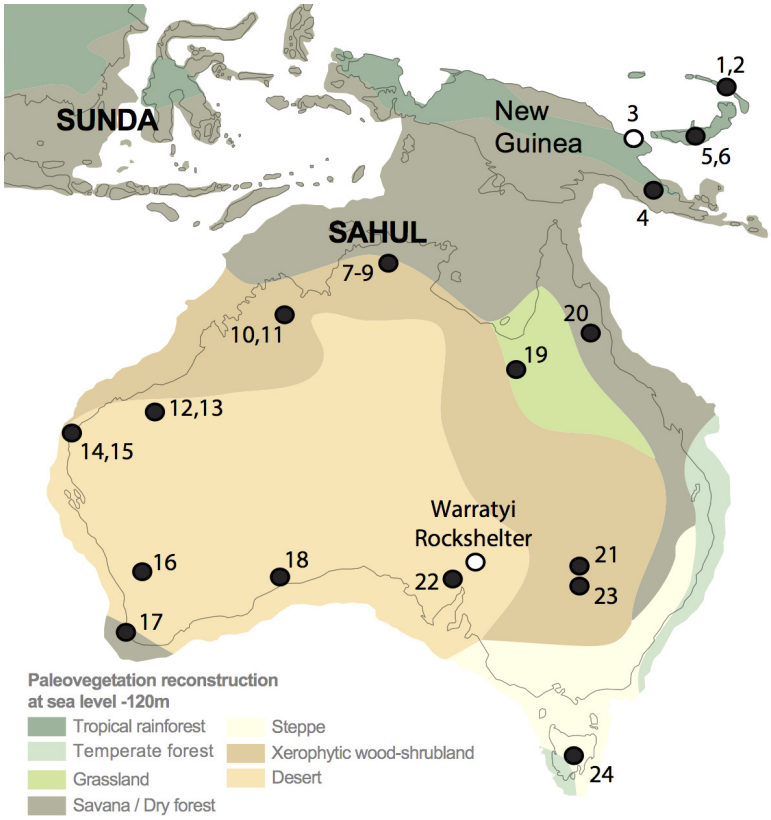
Extended Data Figure 3 | Sahul phylogenetic tree calibrated using mitogenome rate from ref. 32. As for Extended Data Fig. 2, except that rate calibration used the mitogenome rate from ref. 32.



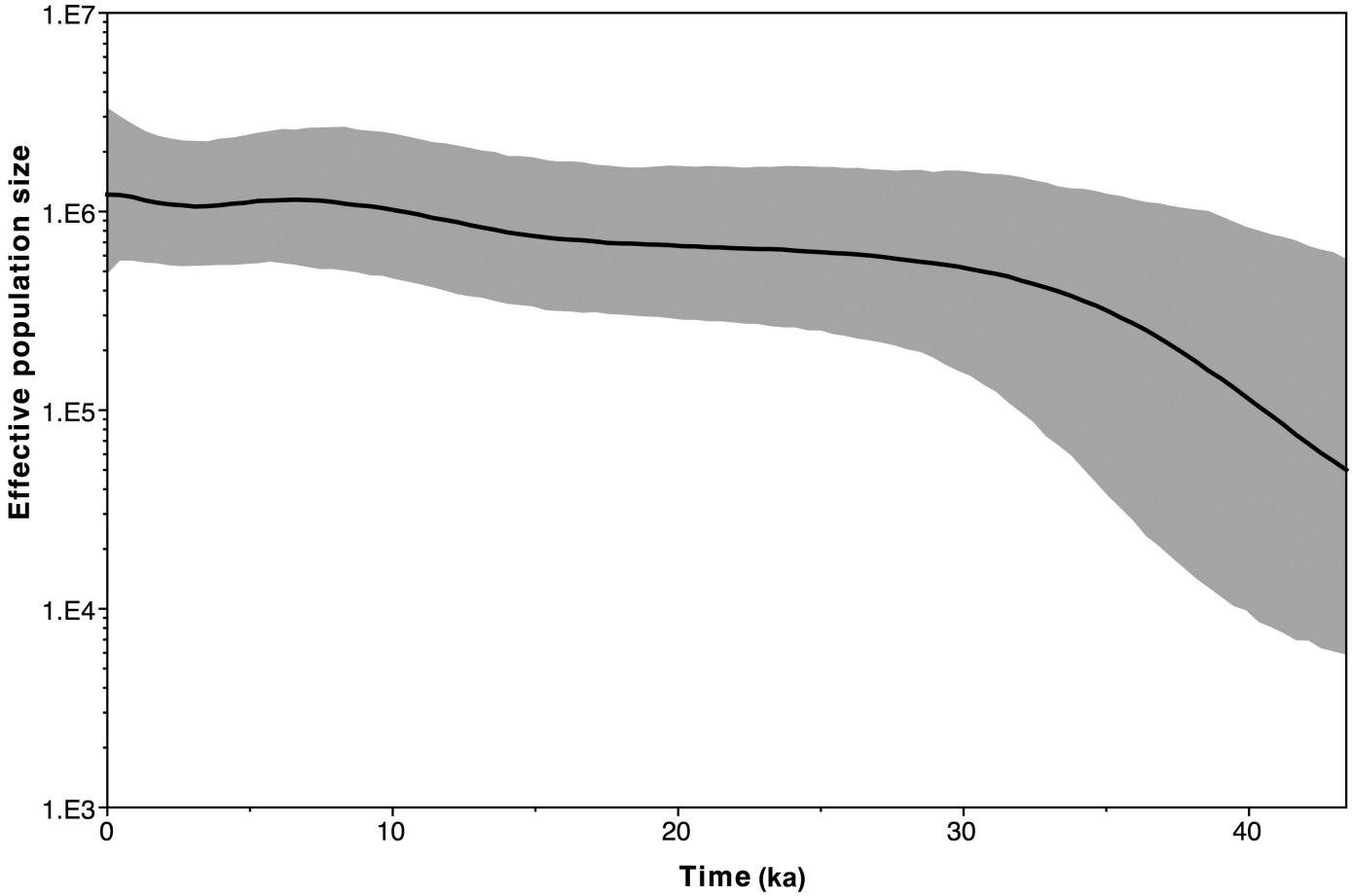
Extended Data Figure 4 | Australian phylogeography incorporating mtDNA lineage information from modern samples reported in ref. 12. The additional samples from ref. 12 are shown as stars and are distributed according to their reported locations of collection, all other sample information is presented in an identical manner to Fig. 2. The mtDNA haplogroups from ref. 12 are coloured according to the system used in Fig. 2, with haplogroups not previously shown (that is, R, R12, M42, P3b and S5) indicated with new colours that are described beneath the relevant haplogroup map (we have added the two R haplogroups on the P haplogroup map, as this is the closest sister clade). As in Fig. 2, mtDNA samples from other studies are shown in yellow, with the samples from ref. 12 having a yellow dot to indicate this status. Map data was sourced from the Oak Ridge National Laboratory Distributed Active Archive Center (https://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10003_1).



Extended Data Figure 5 | Age-depth model for Devil's Lair, south-western Australia. The age-depth model was generated with OxCal v.4.2.4 (ref. 68) using the Poisson process (outlier) deposition model. Original ages with 68% uncertainty (prior to modeling) with laboratory codes shown on left hand side. Prior (light grey) and posterior (dark grey) probability distributions are plotted. The blue and green envelopes describe the 68% confidence interval for the sedimentary units below and above layer 30 (lower) respectively.



Extended Data Figure 6 | Locations of the early occupation sites used to estimate the timing of the colonization of Sahul. Sites used for colonization time estimation are shown as black dots, with white dots indicating sites that were used to provide independent age controls. Sites names: 1, Buang Merabak; 2, Matenkupkum; 3, Huon Peninsula; 4, Ivane; 5, Kupona na Dari; 6, Yombon; 7, Nawarla Gabarnmang; 8, Malakunanja II; 9, Nauwalabila I; 10, Carpenter's Gap; 11, Riwi; 12, Djadjiling; 13, Ganga Mara; 14, Jansz; 15, Mandu Mandu; 16, Upper Swan; 17, Devil's Lair; 18, Allen's Cave; 19, GRE8; 20, Ngarrabullgan; 21, Menindee; 22, Cooper's Dune (PACD H1); 23, Lake Mungo; and 24, Warreen Cave. Additional information for these sites including phase calibrated age ranges for initial occupation is provided in Supplementary Table 4. Phase calibrations were performed using OxCal v.4.2.4 (ref. 68) and resulted in an estimate of the initial colonization of Sahul at 48.8 ± 1.3 ka. The map was adapted from the figure in ref. 36, originally constructed by J.S.



Extended Data Figure 7 | Palaeodemography of Australian mitogenomes. GMRF Skyride[51] analysis of the 98 Australian-only mtDNA lineages showing the estimated effective maternal population size since the initial colonization of Sahul around 50 ka (see Methods). Owing to the lack of available calibration points, the palaeodemographic curve should be considered relatively approximate. Nonetheless, there is no obvious indication of a major population bottleneck during the Last Glacial Maximum (around 21–18 ka). Line, median and grey shading, 95% highest posterior densities.

Extended Data Table 1 | Complete Australian phylogeography test results

Haplogroup	Metric	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
M	Longitude	-0.3194 (0.3474)	0.642 (0.0929)*	-0.32 (0.2476)	0.6072 (0.1615)	NA
	Latitude	0.2310 (0.5444)	0.8560 (0.0055)***	0.4444 (0.4118)	0.0970 (0.6314)	NA
	CI%	42.30	70.30	84.76	93.96	NA
	Mantel Test			0.3273 (0.0953)*		
O	Longitude	0.0429 (0.9258)	-0.6395 (0.0629)*	0.5677 (0.3067)	NA	NA
	Latitude	0.5010 (0.0083)***	0.0002 (0.3935)	0.2923 (0.2174)	NA	NA
	CI%	47.86	75.10	98.18	NA	NA
	Mantel Test			0.3352 (0.0176)***		
P	Longitude	0.7796 (0.0002)***	-0.0703 (0.0053)***	-0.2378 (0.0048)***	-0.0360 (0.0169)**	NA
	Latitude	0.8690 (4e-6)***	0.0260 (0.9190)	-0.0101 (0.1155)	0.2300 (0.9811)	NA
	CI%	38.02	63.18	82.91	90.45	NA
	Mantel Test			0.4488 (3e-6)***		
P4b1	Longitude	-0.1940 (0.2557)	-0.1639 (0.1457)	NA	NA	NA
	Latitude	0.4826 (0.0035)***	0.2675 (0.0587)*	NA	NA	NA
	CI%	42.97	74.44	NA	NA	NA
	Mantel Test			-0.08687 (0.6008)		
P5	Longitude	0.08780 (0.0417)**	0.0556 (0.6083)	NA	NA	NA
	Latitude	-0.7540 (0.1167)	-0.2400 (0.3417)	NA	NA	NA
	CI%	41.49	67.07	NA	NA	NA
	Mantel Test			0.1152 (0.3750)		
S	Longitude	-0.1578 (0.5242)	0.1689 (0.1097)	0.3301 (0.1597)	0.6798 (0.0404)**	0.3351 (0.0016)***
	Latitude	-0.0797 (0.6019)	-0.2512 (0.8633)	0.5977 (0.0006)***	0.2175 (0.8720)	0.1020 (0.1780)
	CI%	26.87	50.25	68.84	81.60	89.56
	Mantel Test			0.2695 (0.0374)**		
All	Longitude	0.3807 (0.0003)***	0.2482 (0.0137)**	NA	NA	NA
	Latitude	0.1748 (0.0617)*	0.1059 (0.1765)	NA	NA	NA
	CI%	12.87	21.43	NA	NA	NA
	Mantel Test			0.2827 (7e-5)***		

Spearman's ρ for correlation with longitude, latitude (with associated P value in parentheses; $^{*}P<0.1$; $^{**}P<0.05$; $^{***}P<0.01$) and the cumulative percentage of inertia (CI%, confidence interval) captured for each principal dimension (first three rows for each haplogroup), along with Spearman's ρ for the Mantel test (with associated P value), for haplogroups M (without M16), O, P (including P4b1 and P5 separately) and S, and the pooled samples (All). Analyses were performed on the 76 samples with reliable provenance (see Methods and Supplementary Table 3).

Chapter 4

Development of Modelling Admixture via Site Pattern Distributions

4.1 Introduction

Speciation is the process by which populations evolve to become distinct species. The process was first described by Charles Darwin in his highly influential book, *The Origin of Species* [14]. In some cases speciation occurs due to populations splitting and inhabiting different geographical locations with different environmental pressures, called allopatric speciation, such as the famous case of Darwin's finches [13]. Other cases may include a subform of allopatric speciation, peripatric speciation, where new sub-populations are made up of a very small number of founding individuals, or parapatric speciation where sub-populations are not completely separated from one another, allowing for very limited interbreeding between populations [26, 27].

Hybridisation is the production of a 'hybrid' offspring from two phylogenetically distinct populations that have undergone partial speciation, but are still able to

interbreed [1]. The terms 'admixture' and 'introgression' both describe types of hybridisation, and are seemingly used interchangeably [2, 42].

Most often, 'introgression' is used to describe the introduction of specific genes into a target population. For example, man-made hybridisation has occurred in modified crops, such as the Flavr Savr tomatoes, which are modified tomatoes that are more resistant to rot [9]. A naturally occurring example is the natural adaption of Tibetans to high altitude breathing through the introgression of the EPAS1 gene from archaic hominids [21].

Conversely, 'admixture' is often used to describe the mixing of whole genomes in population histories. For example the admixture event that occurred between European bison and Aurochs, producing a morphologically distinct hybrid offspring which was captured in cave paintings 21-18 thousand years before present [46].

Hybdridisation plays a key role in evolution, and can act to rapidly adapt a species to a given environment or can act against divergence by allowing continuing gene flow [1, 28]. In some cases detecting gene flow may help in identifying the correct model for the inference of population histories [5]. In other cases the proportion of ancestral admixture in modern populations may be the focus of the research itself [40].

Detecting gene flow, and identifying the proportion of ancestry from ancestral populations, is a difficult problem. The problem received significant interest when the first genome of a Neanderthal individual was sequenced and compared to anatomically modern humans. This allowed researchers to investigate the shared ancestry of the two species [43, 51, 16].

Current methods such as LAMP, HAPMIX and PCADMIX are local ancestry-based methods, and look to infer recent history by investigating patterns of linkage disequilibrium. Despite these methods being extremely powerful for detecting relatively recent admixture events, these methods lack power to detect ancient admixture events, and require sample sizes in the thousands [43, 36, 8].

Global ancestry-based methods are more powerful tools for detecting the sort of population substructure involved in older admixture events. Global ancestry-based methods which employ the use of PCA, and the model-based clustering methods such as STRUCTURE, similarly require large sample sizes in the thousands [33]. Hence these methods are of little use when very few samples are available, such as in ancient DNA studies.

The global ancestry-based methods such as the method implemented in the ADMIXTURE software package and the ratio of the so-called f_4-statistics as employed in the ADMIXTOOLS software package are more in line with our method [37, 34]. However the maximum-likelihood solutions to the model implemented in ADMIXTURE can be shown to be non-identifiable, a problem we also encounter in Section 4.3 [10]. This leaves estimates of mixing proportions via the ratio of f_4 statistics as implemented in the ADMIXTOOLS package as the only method against which we can compare our results. However, it should be noted that ADMIXTOOLS accounts for incomplete lineage sorting (ILS). This makes the method potentially more accurate over shorter time scales where ILS may make a significant impact on parameter estimates, but as a consequence increases model complexity and computational run time.

Here we are specifically interested in the problem of estimating the proportion of ancestry in a hybrid species from two source populations, when only one sample from each population is available, and the effects of ILS can be ignored. Due to the limited information available under these sampling conditions, we aim to develop a relatively assumption-free model.

We begin by deriving the statistical properties for a three-taxon alignment, before adapting the model to include a parameter for admixture. We then discuss parameter estimation through two Bayesian methodologies. Through simulation studies we show that our method performs well under a range of biologically reasonable conditions to produce estimates of admixture proportions, and follow with a discussion of the limitations of our method.

We conclude by using our method to approximate proportions of admixture for two species. Our first data set contains an alignment of ancient European human from before the last glacial maximum. We estimate the proportion of Neanderthal ancestry for each individual, and compare our findings to previously published results. Our second data set contains the late Pleistocene wisent, which we show is an hybrid offspring of the extinct Steppe Bison (*Bison priscus*), and the ancestors of modern cattle, aurochs (*Bos primigenius*).

4.2 A Simple Three-taxon Tree

Consider a simple three-taxon tree with Species A and B more closely related to one another than Species C (see Figure 4.1). Branch lengths are denoted such that $\frac{1}{2}t_m$ units of evolutionary time have passed between the most recent common ancestor (MRCA) of Species A and C, and the hybridisation event (that produced Species B). Similarly, t_a, t_b and t_c units of evolutionary time have passed since the hybridisation event, and the sampling times of A, B and C respectively. Let

$$t_\ell = t_a + t_b + t_c + t_m$$

be the total evolutionary time under consideration. Note that evolutionary time is not measured in calendar units, but rather in units of $4N_e\mu$, where N_e is the effective population size, and μ is the substitution rate per site per generation.

We begin by assuming that we have three aligned sequences of length $N_u >> 0$, one each for Species A, B and C, and that the branch lengths $\boldsymbol{t}^* = (t_a, t_b, t_c, t_m)$ are known. We also assume that only bialellic sites (sites for which there are only two variants) have been kept, such that each site must contain at least one substitution, and that sites have been thinned (by some factor Δ) to reduce the effect of linkage disequilibrium. One possible way to achieve this filtering is given in Algorithm 1.

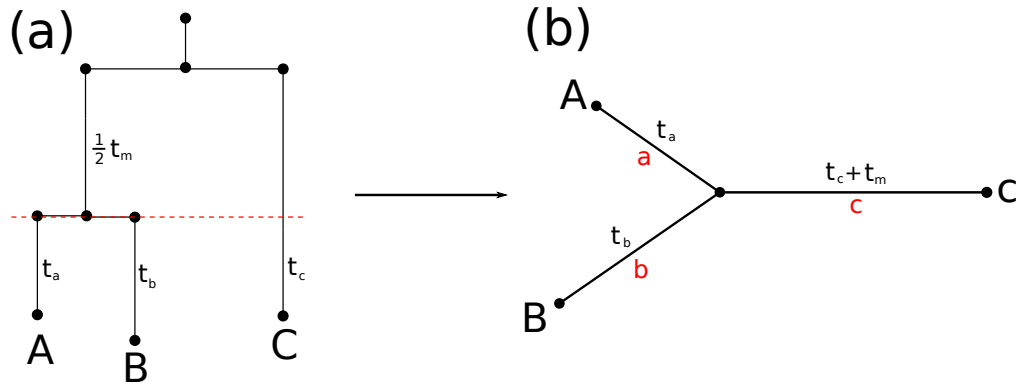


Figure 4.1: (a) A rooted three-taxon tree with (b) the associated unrooted tree. Branch lengths are denoted in black, and branch names are denoted in red. The hybridisation even is depicted by the dashed red line.

Algorithm 1: An algorithm for thinning data to reduce the effect of linkage disequilibrium for an alignment of length N_u, where Δ is the distance between two sites such that the expected coefficient of linkage disequilibrium is sufficiently small.

Require: The position of the first heterogeneous site, j

1: Set $i = j$, and $t = 0$.
2: **while** $i \leq N_u$ **do**
3: Record $s_t = i$.
4: Go to site $i + \Delta$.
5: Find the first site k such that $k \geq i + \Delta$, and k is a biallelic site.
6: $t = t + 1$.
7: $i = k$.
8: **end while**
9: **return** $\boldsymbol{s} = (s_0, s_1, \cdots, s_{t-1})$

We only consider trees on a scale of time such that: the time since the MRCA of A, B and C is relatively short, and hence the probability of observing more than one substitution at any given site is negligible, and yet enough time has passed so that the effect of incomplete lineage sorting is also negligible. This calendar time

must be reasonably derived from external sources, such as from paleontological or environmental evidence. All selected sites may now be considered independent, and are assumed to have undergone exactly one substitution.

Without loss of generality, we relabel the site positions in the thinned alignment $\{1, \cdots, N\}$, such that $N < N_u$. A site pattern P_i (the ordered sequence of nucleotides at site i) must now be of one of the following forms (see Figure 4.2), for X,Y$\in$ $\{A, C, G, T\}$, X$\neq$Y, 'YXX', 'XYX' or 'XXY'.

We define a variable S_i such that

$$S_i = \begin{cases} 1, & \text{if } P_i = \text{“}YXX\text{”}, \\ 2, & \text{if } P_i = \text{“}XYX\text{”}, \\ 3, & \text{if } P_i = \text{“}XXY\text{”}, \end{cases}$$

and let $\boldsymbol{S} = (S_1, S_2, \cdots, S_N)$.

From the thinned alignment we can now count the observed number of site pattern types 1, 2 and 3, and denote these counts

$$\boldsymbol{n} = (n_1, n_2, n_3),$$

such that

$$n_1 + n_2 + n_3 = N.$$

Consider a single site, and let Y_k be the number of substitutions that occurred on branch $k \in \{a, b, c\}$, and define

$$Y = Y_a + Y_b + Y_c.$$

Given that we have filtered sites such that exactly one substitution has occurred at each site, we have that $Y = 1$ with probability one. If we assume a Markov model of nucleotide substitution, then mutations occur according to a Poisson process with rate $4N_e\mu$ along each branch. Since the branches are non-overlapping, the probability of mutations occurring on any individual branch is proportional to length of the

Site	1	2	3	4	5	6
A	A	G	G	C	A	C
B	A	G	A	T	G	C
C	C	T	G	T	G	G

$\longrightarrow$

Site	1	2	3	4	5	6
A	X	X	X	Y	Y	X
B	X	X	Y	X	X	X
C	Y	Y	X	X	X	Y

Figure 4.2: An example of an alignment of length $N = 6$ where the site patterns are recoded in terms of the two similar (X) nucleotides and the unique (Y) nucleotide. In this case $\boldsymbol{S} = (3, 3, 2, 1, 1, 3)$ and $\boldsymbol{n} = (2, 1, 3)$.

branches. That is,

$$P(S_i = j \big| \boldsymbol{t}^*) = \begin{cases} \dfrac{t_a}{t_\ell}, & j = 1, \\ \dfrac{t_b}{t_\ell}, & j = 2, \\ \dfrac{t_c + t_m}{t_\ell}, & j = 3. \end{cases}$$

Since there are finitely many independent sites, with only three possible observable states (with a constant probability of being observed across the alignment), the site pattern counts $\boldsymbol{n}$ can be modelled by a multinomial distribution

$$\boldsymbol{n} \sim \mathrm{MN}\left(N, \frac{t_a}{t_\ell}, \frac{t_b}{t_\ell}, \frac{t_c + t_m}{t_\ell}\right).$$

Note that for $\boldsymbol{t} = (kt_a, kt_b, kt_c, kt_m)$, where $k \in \mathbb{R}^+ \backslash \{0\}$,

$$\begin{aligned} P(S_i = 3 \big| \boldsymbol{t}) &= \frac{kt_c + kt_m}{kt_a + kt_b + kt_c + kt_m} \\ &= \frac{k(t_c + t_m)}{kt_\ell} \\ &= \frac{t_c + t_m}{t_\ell} \\ &= P(S_i = 3 \big| \boldsymbol{t}^*). \end{aligned}$$

Hence we cannot discern between scalar multiples of sets of branch lengths, and so non-dimensionalise by using the constant $k = \frac{1}{t_a + t_b + t_c}$. This rescaling of the branch lengths yields the interpretation that the branch lengths are now the relative amount

of evolutionary time along each branch compared to the total amount of ancestry since the hybridisation event.

This results in a parameter space of reduced dimension, with relative branch lengths

$$\begin{aligned}\boldsymbol{t} &= \left(\frac{t_a}{t_a+t_b+t_c}, \frac{t_b}{t_a+t_b+t_c}, \frac{t_c}{t_a+t_b+t_c}, \frac{t_m}{t_a+t_b+t_c}\right) \\ &= (\alpha_1, \alpha_2, \alpha_3, \beta)\,,\end{aligned}$$

however, since $\alpha_2 = 1 - \alpha_1 - \alpha_3$ we may reduce the parameter space to

$$\boldsymbol{t} = (\alpha_1, \alpha_3, \beta)\,. \tag{4.1}$$

Hence, the site pattern counts have the multinomial distribution

$$\boldsymbol{n} \sim \mathrm{MN}\left(N, \frac{\alpha_1}{1+\beta}, \frac{1-\alpha_1-\alpha_3}{1+\beta}, \frac{\alpha_3+\beta}{1+\beta}\right),$$

with probability mass function

$$f(\boldsymbol{n}|N, \boldsymbol{\alpha}, \beta) = \frac{N!}{n_1!n_2!n_3!}\left(\frac{\alpha_1}{1+\beta}\right)^{n_1}\left(\frac{1-\alpha_1-\alpha_3}{1+\beta}\right)^{n_2}\left(\frac{\alpha_3+\beta}{1+\beta}\right)^{n_3}.$$

4.3 A Three-taxon Admixture Graph

Consider a three-taxon graph with two progenitor species, denoted A and C, and a hybrid species, denoted B (see Figure 4.3). Define $\gamma \in [0, 1]$ to be the proportion of the genome that B has inherited from A, and hence B has inherited a proportion $1 - \gamma$ of its genome from C.

Since the genome of B will be made of blocks of genetic information inherited from A and C, the admixture graph can be thought of as the linear combination of the two underlying phylogenetic trees with topologies denoted X_1^r and X_2^r (see Figure 4.3). For simplicity, we again consider the associated unrooted topologies of X_1^r and X_2^r, denoted X_1 and X_2 respectively (see Figure 4.4).

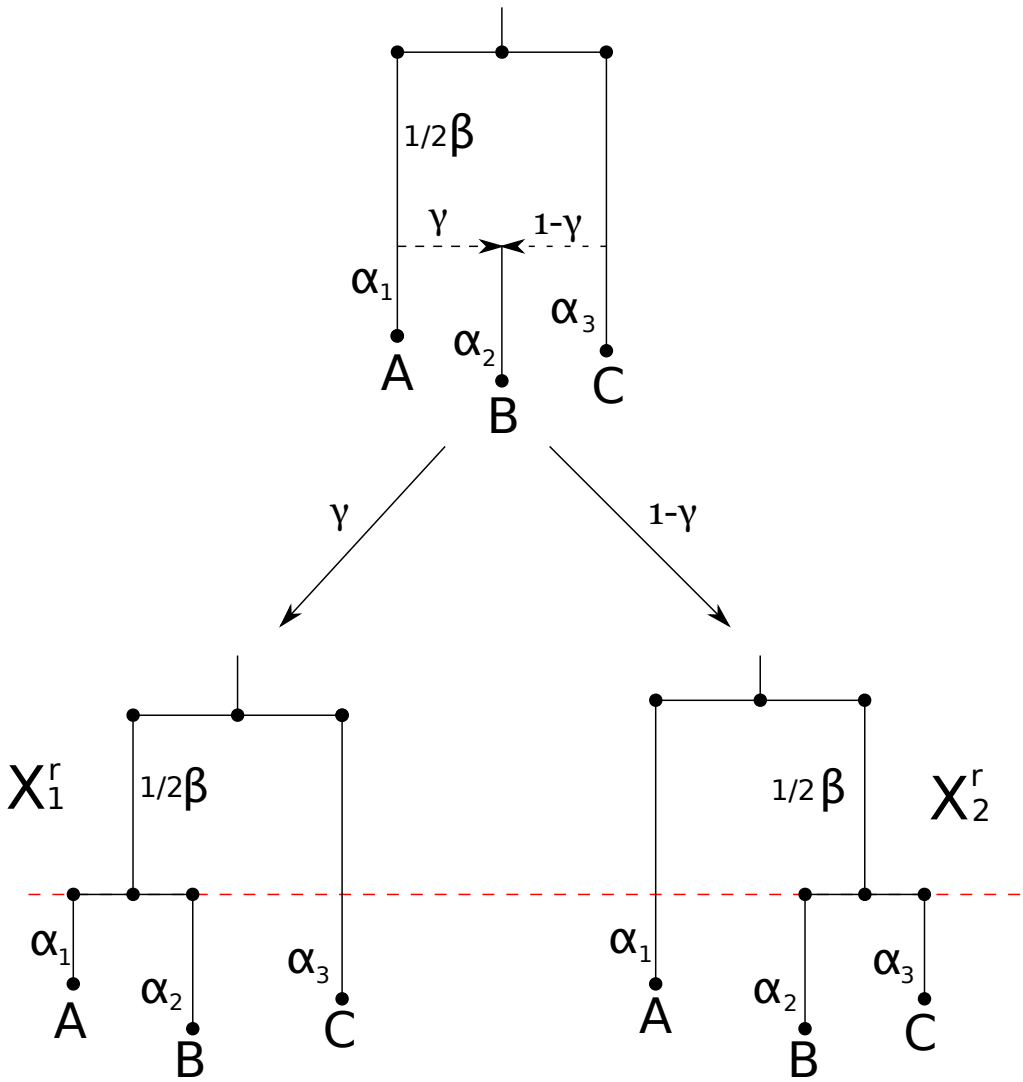


Figure 4.3: The simple three-taxon admixture graph with a hybrid species, denoted B. γ is the proportion of genetic information inherited by B from A. The graph can (site-by-site) be decomposed into a linear combination of the two underlying (rooted) phylogenetic trees X_1^r and X_2^r.

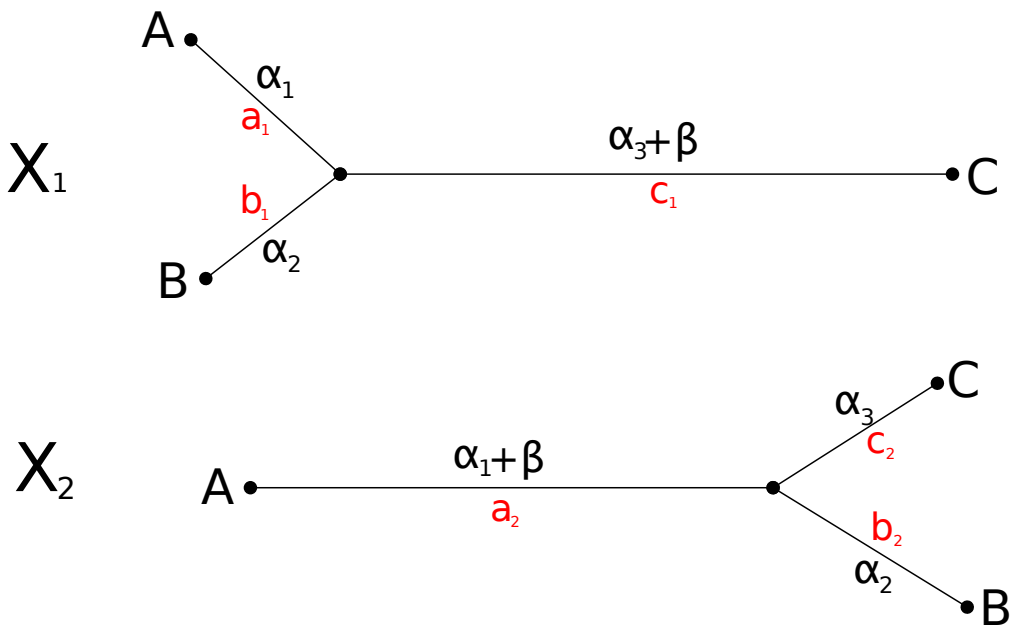


Figure 4.4: The two unrooted underlying topologies X_1 and X_2 for the simple three-taxon admixture graph in Figure 4.3. Branch names are given in red and branch lengths are given in black.

Let Z_i be an indicator variable that describes whether the genetic information at site i is inherited through topology X_1 or X_2. That is,

$$Z_i = \begin{cases} 1, & \text{if site } i \text{ comes from topology } X_1, \\ 2, & \text{if site } i \text{ comes from topology } X_2. \end{cases}$$

Let $\boldsymbol{n}^{X_k} = \left(n_1^{X_k}, n_2^{X_k}, n_3^{X_k}\right)$ denote the site pattern counts contributed by topology X_k, $k = 1, 2$. From Section 4.2, and using the rescaling from Equation (4.1), we have shown that the probability of independent site patterns, given topology X_1 has occurred is

$$P(S_i = j \big| Z_i = 1, \boldsymbol{\alpha}, \beta) = \begin{cases} p_{11} := \frac{\alpha_1}{1+\beta}, & j = 1, \\ p_{12} := \frac{1-\alpha_1-\alpha_3}{1+\beta}, & j = 2, \\ p_{13} := \frac{\alpha_3+\beta}{1+\beta}, & j = 3. \end{cases}$$

It can be shown by a similar argument that the analogous probabilities, given that X_2 has occurred are

$$P(S_i = j \big| Z_i = 2, \boldsymbol{\alpha}, \beta) = \begin{cases} p_{21} := \frac{\alpha_1+\beta}{1+\beta}, & j = 1, \\ p_{22} := \frac{1-\alpha_1-\alpha_3}{1+\beta}, & j = 2, \\ p_{23} := \frac{\alpha_3}{1+\beta}, & j = 3. \end{cases}$$

Hence we have that

$$\boldsymbol{n}^{X_i} \sim \mathrm{MN}\left(N_i, p_{i1}, p_{i2}, p_{i3}\right),$$

where

$$N_1 = \lfloor \gamma N \rfloor, \;\; N_2 = \lfloor (1-\gamma) N \rfloor, \;\; i = 1, 2.$$

Consider now the probability of observing site pattern j at site S_i, given that an admixture event has occurred. By the law of total probability, this site pattern can

come from either topology X_1 or X_2. That is,

$$\begin{aligned}
&P(S_i = 1\big|\boldsymbol{t}) \\
&\qquad =P(S_i = 1\big|Z_i = 1, \boldsymbol{t})P(Z_i = 1\big|\boldsymbol{t}) + P(S_i = 1\big|Z_i = 2, \boldsymbol{t})P(Z_i = 2\big|\boldsymbol{t}) \\
&\qquad =\gamma\left(\frac{\alpha_1}{1+\beta}\right) + (1-\gamma)\left(\frac{\alpha_1+\beta}{1+\beta}\right) \\
&\qquad =\frac{\alpha_1 + (1-\gamma)\beta}{1+\beta}.
\end{aligned}$$

Using the same argument as above for the remaining two cases, and defining

$$\pi_j = P(S_i = j\big|, \boldsymbol{t}),$$

it can be shown that

$$P(S_i = j\big|\boldsymbol{t}) = \begin{cases} \pi_1 := \frac{\alpha_1+(1-\gamma)\beta}{1+\beta}, & j = 1, \\ \pi_2 := \frac{1-\alpha_1-\alpha_3}{1+\beta}, & j = 2, \\ \pi_3 := \frac{\alpha_3+\gamma\beta}{1+\beta}, & j = 3. \end{cases}$$

The site pattern counts $\boldsymbol{n}$ from the admixture graph is then a linear combination of the contributions from the two topologies, where the mixing parameter is the proportion of the N total sites contributed by each topology.

Hence, if we denote $\mathbf{n} = (n_j, j = 1, 2, 3)$ to be the site pattern counts observed on a mixture of the topologies X_1 and X_2, with proportions γ and $(1-\gamma)$ contributed from the topologies respectively, and denote $\boldsymbol{T} = \{\gamma, \boldsymbol{t}\}$ then

$$\mathbf{n}\big|\boldsymbol{T} \sim \mathrm{MN}\left(N, \pi_1, \pi_2, \pi_3\right). \tag{4.2}$$

The expected site pattern counts yields an intuitive result.

$$\begin{aligned}
E\left[\boldsymbol{n}\right] =&E\left[\mathbf{n}^{X_1}\right] + E\left[\mathbf{n}^{X_2}\right] \\
=&\gamma\left(N\frac{\alpha_1}{1+\beta}, N\frac{1-\alpha_1-\alpha_3}{1+\beta}, N\frac{\alpha_3+\beta}{1+\beta}\right) \\
&+ (1-\gamma)\left(N\frac{\alpha_1+\beta}{1+\beta}, N\frac{1-\alpha_1-\alpha_3}{1+\beta}, N\frac{\alpha_3}{1+\beta}\right) \\
=&\frac{N}{1+\beta}\Big[\left(\alpha_1, \alpha_2, \alpha_3\right) + \beta\left(1-\gamma, 0, \gamma\right)\Big].
\end{aligned}$$

That is, we expect a number of site pattern counts that is proportional to the branch lengths from the admixture event until the sampling events. However, for some fixed t_m, and hence β, the expected number of site patterns of the type 'YXX' decreases, and the expected number of the type 'XXY' increases as $\gamma \to 1$. This makes intuitive sense since an increase in γ indicates an increased proportion of ancestry from topology X_1, meaning Species A and B share more ancestry, leading to a decrease in patterns of the form 'YXX', and an increase in patterns of the form 'XXY'. Conversely, the expected number of site patterns of the type 'YXX' increases, and the expected number of the type 'XXY' decreases as $\gamma \to 0$.

Note that in the formulation of this mixture model we have greatly reduced the complexity of the problem. Since the branches are measured in evolutionary time (in units of $4N_e\mu$), we need no knowledge of the demographic history of any of the populations of interest. By filtering for sites with at least (and by assumption, at most) one substitution, and recoding these in terms of common and unique nucleotides, we have removed the need to consider a substitution model. By assuming that the substitution rate μ has remained constant for all branches on the tree, the need to consider μ was ignored when probabilities were calculated as proportions of branch lengths. Finally, though it was necessary due to the aliasing of the multinomial probability parameters, by rescaling the branching lengths by $(t_a + t_b + t_c)^{-1}$, we had that $\alpha_2 = 1 - \alpha_1 - \alpha_3$, further reducing the dimension of the parameter space by one.

Recall that, in general, for a multinomially distributed vector of counts $\boldsymbol{n} = (n_1, n_2, n_3)$, with associated probability vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$, the probability mass function is given by

$$f(\boldsymbol{n}|\boldsymbol{\pi}) = \frac{(\sum_{i=1}^{3} n_i)!}{\prod_{i=1}^{3} n_i!} \prod_{i=1}^{3} \pi_i^{n_i}.$$

To find maximum-likelihood estimates for the π_i, we find the log-likelihood function, and include a Lagrange multiplier to account for the constraint of unity for the

probabilities,

$$\ell(\boldsymbol{\pi}, \lambda) = \log\left(\frac{(\sum_{i=1}^{3} n_i)!}{\prod_{i=1}^{3} n_i!}\right) + \sum_{i=1}^{3} n_i \log(\pi_i) + \lambda\left(1 - \sum_{i=1}^{3} \pi_i\right).$$

The first derivative of the log-likelihood function, with respect to π_j is

$$\frac{\partial \ell(\boldsymbol{\pi}, \boldsymbol{\lambda})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \lambda. \tag{4.3}$$

Setting Equation (4.3) to zero, and rearranging for π_j yields

$$\hat{\pi}_j = \frac{n_j}{\lambda}. \tag{4.4}$$

We know that the sum of the class probabilities must be one, and so

$$\begin{aligned} &\sum_{i=1}^{3} \pi_i = 1 \\ \implies &\sum_{i=1}^{3} \frac{n_i}{\lambda} = 1 \\ \implies &\lambda = \sum_{i=1}^{3} n_i = N. \end{aligned} \tag{4.5}$$

Hence, substituting Equation (4.5) into Equation (4.4) yields

$$\hat{\pi}_j = \frac{n_j}{N}, \ \ \forall j = 1, 2, 3,$$

which yields,

$$\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3) = \left(\frac{n_1}{N}, \frac{n_2}{N}, \frac{n_3}{N}\right).$$

However, we have four parameters, γ, α_1, α_2 and β in the original parametrisation of the model, but only three maximum-likelihood estimators for the $\hat{\pi}_i$ (and only two of which are linearly independent). Hence, while we have found the MLE for the probabilities of the 'mixture' model, we cannot back-transform these estimates to find maximum-likelihood estimates for γ, α_1, α_2 and β. Specifically, γ is not identifiable here using just the $\hat{\pi}_i$.

4.4 Parameter Estimation via Approximate Bayesian Computation

Given a specific vector of values $\boldsymbol{T} = (\gamma, \alpha_1, \alpha_2, \beta)$, it is a trivial computational task to calculate the probabilities of observing the site patterns, and hence to simulate multinomial counts. This ease of simulation lends itself naturally to the use of approximate Bayesian computation (ABC) for inference [4].

ABC is a likelihood-free method which, given some data $\boldsymbol{n}_{obs}$, obtains a finite sample from the (approximate) posterior distribution [4]. To do this, we require a prior distribution for the admixture parameters $\tau(\boldsymbol{T})$. From the prior distribution we sample a candidate set of parameter values $\boldsymbol{T}^{(j)} = \left(\gamma^{(j)}, \alpha_1^{(j)}, \alpha_2^{(j)}, \beta^{(j)}\right)$, and simulate a data set $\boldsymbol{n}^{(j)}$ from the multinomial distribution defined in Equation (4.2). The Euclidean distance between the jth simulated data set and the observed data can be calculated, denoted $\rho^{(j)} = \rho(\boldsymbol{n}^{(j)}, \boldsymbol{n}_{obs})$, indicating how similar the simulated data is to the observed data. If the data is 'similar enough' such that $\rho^{(j)} \leq \epsilon$, for some predefined tolerance parameter ϵ, the candidate value of $\gamma^{(j)}$ is added to the posterior sample, otherwise it is discarded.

The algorithm terminates when the total number of accepted samples reaches a predefined sample size, N_P. Alternatively, to reduce computational run-time, some algorithms begin by simulating N_M candidate parameter and data sets, and retain the $\lfloor \xi N_M \rfloor$ closest data sets, where $\xi \in (0, 1)$.

It is possible to use regression-based correction methods for the sampled posterior density. That is, although we choose to retain only values of $\gamma^{(j)}$ such that $\rho^{(j)} \leq \epsilon$, we may wish to weight more posterior density to values of $\rho^{(j)}$ that are closer to zero.

We assume

$$\gamma^{(j)} = m(\boldsymbol{n}^{(j)}) + \varepsilon_j$$

where m is a regression function, and the ε_j are centred, homoscedastic independent random variables. In this case we use a ridge regression function for m.

Algorithm 2: An implementation of an Approximate Bayesian Computation algorithm.

```
Set j = 1
while i ≤ N_P do
    Sample T* = (γ*, α*_1, α*_2, β*) from τ(T).
    Simulate a realisation of the process n*
    if ρ(n*, n_obs) ≤ ϵ then
        Set γ_j = γ*
        j = j + 1
    end if
end while
return γ = (γ_1, γ_2, ··· , γ_{N_P})
```

Once the regression is performed on the posterior sample, a weighted posterior sample is obtained by performing the following correction to the $\gamma^{(j)}$ via the following equation

$$\hat{\gamma}^{(j)} = \hat{m}\left(\boldsymbol{n}^{(j)}\right) + \frac{\hat{\sigma}\left(\boldsymbol{n}_{obs}\right)}{\hat{\sigma}\left(\boldsymbol{n}^{(j)}\right)}\hat{\varepsilon}_j, \tag{4.6}$$

where $\hat{\sigma}(\cdot)$ is the estimated conditional standard deviation [6].

From this corrected posterior sample we may calculate the empirical median, and upper and lower bounds for a 95% posterior probability region. Note though that a result of the correction method is that it is now possible for the corrected posterior distribution to contain values of γ that are less than zero. Hence, the lower bound of the $(1-\chi)\%$ posterior probability distribution may be less than zero.

4.5 Parameter Distribution Estimation via Numerical Integration

Instead of the simulation approach described in Section 4.4, we investigate a numerical approximation to the posterior distribution of γ, given prior beliefs about the admixture graph parameters.

Let $\boldsymbol{\alpha}$ have prior distribution, $\boldsymbol{\alpha} \sim \text{Dirichlet}(\boldsymbol{a})$, where $\boldsymbol{a} = (a_1, a_2, a_3)$, which yields

$$P(\boldsymbol{\alpha}|\boldsymbol{a}) = \frac{1}{B(\boldsymbol{a})} \alpha_1^{a_1-1} \left(1 - \alpha_1 - \alpha_3\right)^{a_2-1} \alpha_3^{a_3-1},$$

where

$$B(\boldsymbol{a}) = \frac{\Gamma(a_1)\Gamma(a_2)\Gamma(a_3)}{\Gamma(a_1 + a_2 + a_3)},$$

and

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

Let β have prior distribution $\beta \sim U\left[\beta_\ell, \beta_u\right]$, which yields that

$$P(\beta) = \frac{1}{\beta_u - \beta_\ell}, \;\; \beta \in \left[\beta_\ell, \beta_u\right],$$

and let γ have prior distribution $\gamma \sim U\left[0, 1/2\right]$, which yields that

$$P(\gamma) = 2, \;\; \gamma \in \left[0, 1/2\right].$$

We restrict γ to the support $[0, 1/2]$ to better utilise computational effort due to the fact that in most cases it is trivial to identify the species that contributes more to the hybrid offspring.

From Bayes' formula we have that

$$P\left(\gamma, \beta, \boldsymbol{\alpha}|\boldsymbol{n}\right) = \frac{P(\boldsymbol{n}|\gamma, \beta, \boldsymbol{\alpha}) P(\gamma, \beta, \boldsymbol{\alpha})}{P(\boldsymbol{n})}.$$

Assuming that parameters γ,β and $\boldsymbol{\alpha}$ are independent, then

$$\begin{aligned} P\left(\gamma,\beta,\boldsymbol{\alpha}\middle|\boldsymbol{n}\right) &= \frac{P(\boldsymbol{n}\middle|\gamma,\beta,\boldsymbol{\alpha})P(\gamma)P(\beta)P(\boldsymbol{\alpha})}{P(\boldsymbol{n})} \\ &= K\left(\frac{\alpha_1+(1-\gamma)\beta}{1+\beta}\right)^{n_1}\left(\frac{1-\alpha_1-\alpha_2}{1+\beta}\right)^{n_2}\left(\frac{\alpha_3+\gamma\beta}{1+\beta}\right)^{n_3} \\ &\quad\times\alpha_1^{a_1-1}\left(1-\alpha_1-\alpha_3\right)^{a_2-1}\alpha_3^{a_3-1}, \end{aligned}$$

where

$$K=\frac{2B(\boldsymbol{a})(n_1+n_2+n_3)!}{P(\boldsymbol{n})n_1!n_2!n_3!(\beta_u-\beta_\ell)}.$$

Consider the behaviour of $P\left(\gamma,\beta,\boldsymbol{\alpha}\middle|\boldsymbol{n}\right)$ as the $n_i\to\infty$.

Since

$$\left(\frac{\alpha_1+(1-\gamma)\beta}{1+\beta}\right)<1,\left(\frac{1-\alpha_1-\alpha_3}{1+\beta}\right)<1,\left(\frac{\alpha_3+\gamma\beta}{1+\beta}\right)<1,$$

then for a large number of pattern counts,

$$\left(\frac{\alpha_1+(1-\gamma)\beta}{1+\beta}\right)^{n_1}\left(\frac{1-\alpha_1-\alpha_3}{1+\beta}\right)^{n_2}\left(\frac{\alpha_3+\gamma\beta}{1+\beta}\right)^{n_3}\to 0.$$

This leads to underflow issues when performing numerical integration. To avoid this issue, we introduce $\frac{1}{r^N}$, a constant normalisation parameter, where

$$r=\left[\frac{n_1}{N}\right]^{\frac{n_1}{N}}\left[\frac{n_2}{N}\right]^{\frac{n_2}{N}}\left[\frac{n_3}{N}\right]^{\frac{n_3}{N}}.$$

If the function

$$f\left(\boldsymbol{\pi}\middle|\boldsymbol{n}\right)=\frac{(\sum_{i=1}^3 n_i)!}{\prod_{i=1}^3 n_i!}\prod_{i=1}^3\pi_i^{n_i}$$

is uniquely maximised by $\hat{\boldsymbol{\pi}}=\left(\frac{n_1}{N},\frac{n_2}{N},\frac{n_3}{N}\right)$, then so must a function proportional to $f\left(\boldsymbol{\pi}\middle|\boldsymbol{n}\right)$, specifically

$$f^*\left(\boldsymbol{\pi}\middle|\boldsymbol{n}\right)=\frac{1}{r^N}\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}.$$

The maximum value this function takes is,

$$
\begin{aligned}
f\left(\hat{\boldsymbol{\pi}}\middle|\boldsymbol{n}\right) &= \frac{\left[\frac{n_1}{N}\right]^{n_1}\left[\frac{n_2}{N}\right]^{n_2}\left[\frac{n_3}{N}\right]^{n_3}}{\left(\left[\frac{n_1}{N}\right]^{\frac{n_1}{N}}\left[\frac{n_2}{N}\right]^{\frac{n_2}{N}}\left[\frac{n_3}{N}\right]^{\frac{n_3}{N}}\right)^N} \\
&= \frac{\left[\frac{n_1}{N}\right]^{n_1}\left[\frac{n_2}{N}\right]^{n_2}\left[\frac{n_3}{N}\right]^{n_3}}{\left[\frac{n_1}{N}\right]^{n_1}\left[\frac{n_2}{N}\right]^{n_2}\left[\frac{n_3}{N}\right]^{n_3}} \\
&= 1.
\end{aligned}
$$

Hence, in regions of appreciably non-zero probability density, the rescaled density function $f^*(\boldsymbol{\pi}\big|\boldsymbol{n})$ is less likely to suffer from underflow.

We then have that the marginal distribution of γ is

$$
\begin{aligned}
P(\gamma|\boldsymbol{n}) &= \int_{\beta_\ell}^{\beta_u}\int_0^1\int_0^{1-\alpha_3} P\left(\gamma,\beta,\boldsymbol{\alpha}\middle|\boldsymbol{n}\right) d\alpha_1 d\alpha_3 d\beta \\
&= \int_{\beta_\ell}^{\beta_u}\int_0^1\int_0^{1-\alpha_3} \frac{K}{r^N}\left(\frac{\alpha_1+(1-\gamma)\beta}{1+\beta}\right)^{n_1}\left(\frac{1-\alpha_1-\alpha_3}{1+\beta}\right)^{n_2}\left(\frac{\alpha_3+\gamma\beta}{1+\beta}\right)^{n_3}\times \\
&\qquad \alpha_1^{a_1-1}(1-\alpha_1-\alpha_3)^{a_2-1}\alpha_3^{a_3-1} d\alpha_1 d\alpha_3 d\beta \\
&\propto \int_{\beta_\ell}^{\beta_u}\int_0^1\int_0^{1-\alpha_3} \frac{1}{r^N}\left(\frac{\alpha_1+(1-\gamma)\beta}{1+\beta}\right)^{n_1}\left(\frac{1-\alpha_1-\alpha_3}{1+\beta}\right)^{n_2}\left(\frac{\alpha_3+\gamma\beta}{1+\beta}\right)^{n_3}\times \\
&\qquad \alpha_1^{a_1-1}(1-\alpha_1-\alpha_3)^{a_2-1}\alpha_3^{a_3-1} d\alpha_1 d\alpha_3 d\beta \qquad (4.7)
\end{aligned}
$$

Note that we omit the normalising constant K to reduce computational complexity, and to reduce the probability of underflow occurring. It is trivial to renormalise estimates of the posterior density by simply ensuring that the total probability mass sums to one.

There is no simple elementary function that is the solution of Equation (4.7), and so we use numerical integration to find an approximation for the marginal posterior density of γ.

Consider an interval on which we wish to consider γ, denoted $\gamma \in [0, \gamma_{max}]$. Assuming a uniform grid spacing for $\gamma \in [0, \gamma_{max}]$, such that $\Delta_\gamma = \gamma_{max}/m$, we aim to produce

estimates of the marginal posterior density of γ, evaluated at

$$\boldsymbol{\gamma} = (0, \Delta_\gamma, 2\Delta_\gamma, \cdots, i\Delta_\gamma, \cdots, \gamma_{max}).$$

For the function

$$\begin{aligned}
&\phi(\gamma_i, \beta_j, \alpha_{1,k}, \alpha_{3,\ell}, \boldsymbol{n}, \boldsymbol{a}) \\
=&\frac{1}{r^N}\left(\frac{\alpha_{1,\ell} + (1-\gamma_i)\beta_j}{1+\beta_j}\right)^{n_1}\left(\frac{1-\alpha_{1,\ell}-\alpha_{3,\ell}}{1+\beta_j}\right)^{n_2}\left(\frac{\alpha_{3,k}+\gamma_i\beta_j}{1+\beta_j}\right)^{n_3} \\
&\times \alpha_{1,k}^{a_1-1}(1-\alpha_{1,k}-\alpha_{3,k})^{a_2-1}\alpha_{3,\ell}^{a_3-1},
\end{aligned}$$

we estimate the marginal posterior density for γ_i by numerically integrating over all values of β, α_1 and α_3 for the function. We do this by also considering a uniform grid spacing for the parameters

$$\left\{(\beta, \alpha_1, \alpha_3)\,\middle|\,\beta_\ell \le \beta \le \beta_u, 0 \le \alpha_1 \le 1, 0 \le \alpha_3 \le 1-\alpha_1\right\}$$

of the form

$$\Delta_\beta = \frac{\beta_u - \beta_\ell}{m}, \Delta\alpha_1 = \frac{1}{m}, \Delta\alpha_3 = \frac{1}{m}, \beta_j = j\Delta_\beta, \alpha_{1,k} = \frac{k}{m}, \alpha_{3,\ell} = \frac{\ell}{m},$$

and

$$\alpha_k^* = \lfloor 1 - \frac{\alpha_{1,k}}{m} \rfloor.$$

Using a 3-dimensional form of the trapezoidal rule we get

$$P(\gamma_i|\boldsymbol{n}) \approx \tilde{I}_{\gamma_i} = \sum_{j=0}^{m}\sum_{k=0}^{m}\sum_{\ell=0}^{\alpha_k^*}\frac{1}{2^h}\phi(\gamma_i, \beta_j, \alpha_{1,k}, \alpha_{3,\ell}, \boldsymbol{n}, \boldsymbol{a})\Delta_\beta\Delta_{\alpha_1}\Delta_{\alpha_3}$$

where

$$h = \mathbb{1}_{\{\beta_j=\beta_\ell\}} + \mathbb{1}_{\{\beta_j=\beta_u\}} + \mathbb{1}_{\{\alpha_{1,k}=0\}} + \mathbb{1}_{\{\alpha_{1,k}=1\}} + \mathbb{1}_{\{\alpha_{3,\ell}=0\}} + \mathbb{1}_{\{\alpha_{3,\ell}=1-\alpha_{1,k}\}},$$

are simply the boundary cases.

Finally, we normalise the posterior estimate to have total probability mass one, *i.e.*

$$\hat{p}(\gamma_i) = \tilde{I}_{\gamma_i} / \sum_{\gamma_j \in \boldsymbol{\gamma}} \tilde{I}_{\gamma_j}.$$

From these discrete estimates of the posterior density of γ we may calculate an estimate of the median of the posterior distribution

$$\hat{M}_\gamma = \gamma_i,$$

such that

$$i = \arg\min_j \sum_{j=0}^{m} \hat{p}(\gamma_j) \geq \frac{1}{2}.$$

Further, we obtain *conservative* $(1-\chi)\%$ probability intervals by selecting values of γ_i, denoted $(\ell, u) = (\gamma_k, \gamma_m)$ such that

$$k = \arg\max_j \sum_{j=0}^{m} \hat{p}(\gamma_j) \leq \frac{\chi}{2}, \tag{4.8}$$

and

$$m = \arg\min_j \sum_{j=0}^{m} \hat{p}(\gamma_j) \geq 1 - \frac{\chi}{2}. \tag{4.9}$$

Due to the discrete grid of values of γ_i at which we evaluate the posterior density, it is unlikely that these intervals contain exactly $(1-\chi)\%$ of the posterior density, and as such will give values of ℓ and u that are outside the true interval.

4.6 Analysis of Simulated Data

4.6.1 Experimental Design

We begin simulating data under two scenarios, Scenario A and Scenario B, which differ only by the values β may take. We use calendar time as a scaled proxy for evolutionary time, and use estimates of ancient sampling dates from published research.

Scenario A describes biologically reasonable conditions under which admixture may occur, and we base the simulation parameters on the human(*Homo sapiens*) and Neanderthal (*Homo neanderthalensis*) admixture event dating back to between 40

to 50 kya [40, 18]. We let Species A represent a Neanderthal individual, Species C represent a Moroccan individual with no European ancestry, and Species B represent a post-hybridisation Western Eurasian human hybrid offspring of Africans and Neanderthals. The time until the MRCA (T_{MRCA}) of humans and Neanderthals is thought to have been somewhere between 550 kya and 700 kya, yielding an approximate value of $t_m \approx (1 \times 10^6, 1.4 \times 10^6)$ [40]. Since Neanderthal individuals have been sampled before and after the hybridisation event, we let $t_c \approx 1$ ky [45]. A Western Eurasian individual dating to between 37 and 39 kya has been found and successfully sequenced, and so we use this as a proxy to let $t_b \in [1, 13]$ ky [38]. Finally a Moroccan individual with no European ancestry has been found dating to between 14 to 15 kya, and so we let $t_a \in [25, 36]$ ky [50] .

These tip branch lengths, when rescaled by the total length of the branches since the hybridisation event, yield an approximate interval of

$$\beta \in \left(\frac{1 \times 10^6}{51 \times 10^3}, \frac{1.4 \times 10^6}{36 \times 10^3} \right) \approx (19.61, 38.89) \,.$$

We allow greater flexibility in the tip lengths by simply assuming that

$$\boldsymbol{\alpha} \sim \mathit{Dirichlet}\,(5, 5, 5)\,,$$

that is, the α_i, $i = 1, 2, 3$, are simply equally likely to take any value between zero and one, and must sum to one. This allows for greater relative branch lengths for all three species, and hence emulates more potential population histories.

In Scenario B we allow less time since the speciation event separating Species A and C, and more time since the hybridisation event creating Species B. Recall that this is a relative measure, meaning that a small value of β is caused by a small value of t_m relative to $t_a + t_b + t_c$. There are two ways in which this could happen.

First, it could be that t_m is very small. This would indicate that a very small amount of time has passed since the speciation event that produced Species A and C. One may consider particularly small values of t_m unreasonable for Species A and C to have sufficiently diverged, and so we discount this interpretation.

However, it may also be the case that while t_m is sufficiently large for speciation to occur, so much time has passed since the hybridisation event and the sampling of one, or all, of the species, that $t_a + t_b + t_c$ is very large, relative to t_m. In this case, most of the site patterns that we observe are the result of point mutations on the tips of the tree, yielding proportionally less information about the hybridisation event in the site pattern counts.

We base these simulations on a maximum likelihood tree obtained from the mtDNA of a Neanderthal individual (*Homo neanderthalensis*), a human Yoruban individual and the revised Cambridge reference sequence [3, 22, 19]. This yields estimates of $t_a = t_b = t_c = 2.3110 \times 10^{-3}$, $t_m = 2.319 \times 10^{-2}$, and $\beta = 3.3448$. We allow additional variability by simulating values of β such that

$$\beta \in (3, 4)\,.$$

Except for the first five percent of simulations where we impose $\gamma = 0$, we uniformly sample values of $\gamma \in (0, 0.25]$, since it is almost always possible to tell to which of Species A or C that Species B is most closely related. We omit the region $(0.25, 0.5]$ as values of γ this high are unlikely.

We use conservative total site pattern counts of 1×10^5 to allow for poor coverage in ancient sampling. We simulate 5×10^6 simulations for the ABC analysis, and choose N_M, the number of simulations for the ABC analyses such that the posterior sample size was 5000,and hence we retain the 1% 'closest' simulations. We then selected a value $m = 125$, the grid size for the numerical integration method such that the confidence intervals were of approximately equal width, resulting in a total of 2.5162×10^8 individual calculations.

We took $M = 1000$ independent samples of the admixture graph parameters, denoted γ_i, $\boldsymbol{\alpha}_i$ and β_i from the prior distributions given in Table 4.1. For each sampled parameter set, we calculated the site pattern probabilities, denoted $\boldsymbol{\pi}_i$. From the site pattern probabilities we took a sample of site pattern counts from the multinomial

Parameter	Description	Scenario A	Scenario B
γ	Species A ancestry proportion	$\frac{95}{100} \times U[0, 0.25] + \frac{5}{100} \times \mathbb{1}_{\{\gamma=0\}}$	$\frac{95}{100} \times U[0, 0.25] + \frac{5}{100} \times \mathbb{1}_{\{\gamma=0\}}$
β	Scaled ancestral branch length	$U[35, 40]$	$U[3, 4]$
$\boldsymbol{\alpha}$	Scaled external branch lengths	$Dirichlet(5, 5, 5)$	$Dirichlet(5, 5, 5)$
N	Number of loci	1×10^5	1×10^5
m	Grid size for integration	125	125
ξ	Proportion of kept simulations for ABC	0.001	0.001
N_M	Number of simulations for ABC	5×10^6	5×10^6

Table 4.1: Table of tuning parameters and prior distributions for parameter values for Scenario A and Scenario B of the simulated data.

distribution

$$\boldsymbol{n}_i \sim MN(1 \times 10^5, \boldsymbol{\pi}_i),$$

yielding 1000 independent simulated site patterns counts, with known mixing parameter.

For each simulated site pattern count we used both the ABC and the numerical integration approaches to estimate the marginal posterior distribution of γ_i. We then calculate an estimate of the posterior median, denoted $\widehat{M}_i^{ABC}$ for the ABC approach, and $\widehat{M}_i^{NI}$ for the numerical integration approach (see Figure 4.5). Note that when discussing the estimated posterior median in general for both methods, we simply use the notation $\widehat{M}_i$.

For each estimated posterior median, $\widehat{M}_i^{ABC}$ and $\widehat{M}_i^{NI}$, we define the residual of the estimators

$$r_i^{ABC} = \widehat{M}_i^{ABC} - \gamma_i$$

and

$$r_i^{NI} = \widehat{M}_i^{NI} - \gamma_i$$

respectively.

Finally, summary statistics for the residuals may be calculated, such as, $\bar{r}$, the mean of the observed residual of the posterior median,

$$\bar{r}^{ABC} = \frac{1}{M}\sum_{i=1}^{M} r_i^{ABC} \quad \text{and} \quad \bar{r}^{NI} = \frac{1}{M}\sum_{i=1}^{M} r_i^{NI},$$

respectively. The sample standard deviations, denoted s_r^{ABC} and s_r^{NI}, may also be calculated.

4.6.2 Results for Scenario A

From Figure 4.6 we can see that both the ABC and numerical integration methods both appear to estimate the true value of γ well, and both methods yield a Spearman

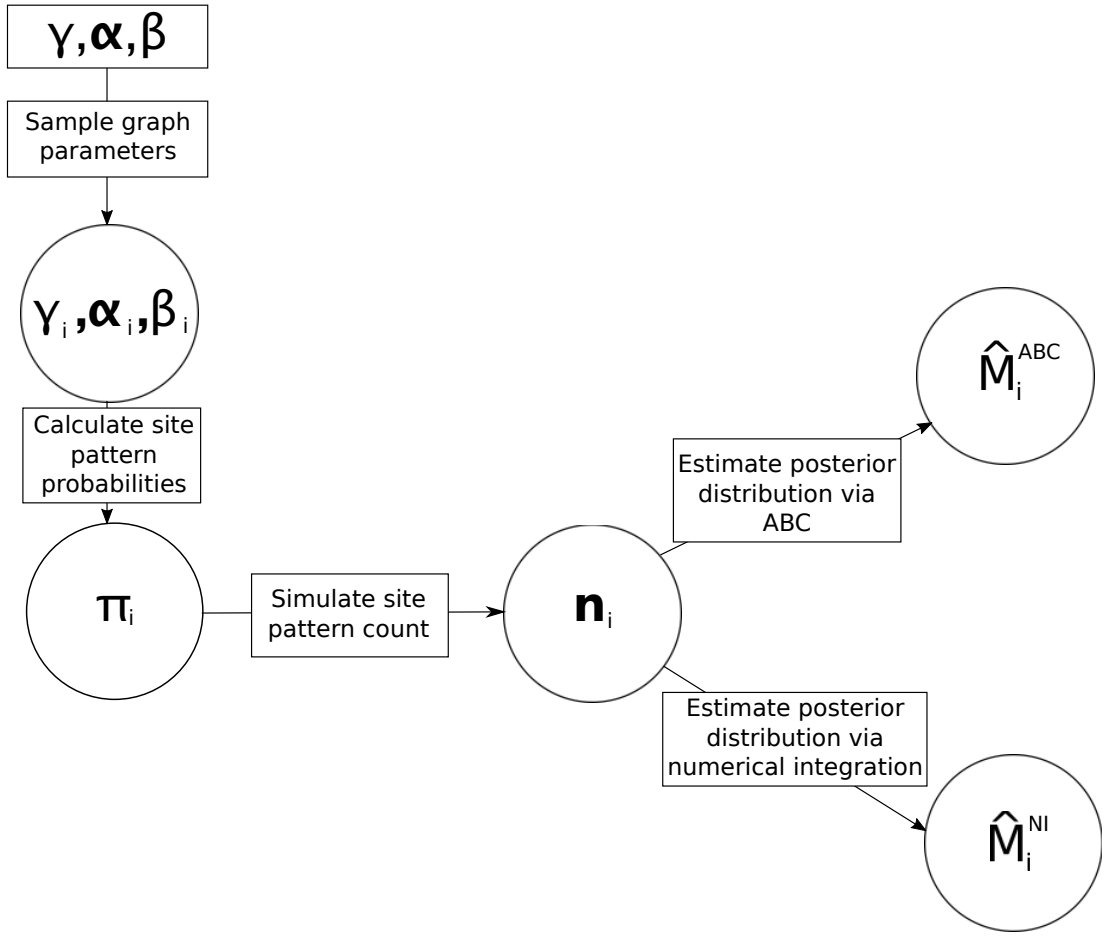


Figure 4.5: A flow diagram for the simulation study. See Table 4.1 for specific simulation parameter values.

sample correlation of $\rho_S = 0.9989$ between the true value of γ and the estimated posterior median $\widehat{M}$ for all 1000 simulations.

There is clear positive bias for $\widehat{M}$ for values of γ close to zero, as we never estimate $\widehat{M} < 0$ (see Figure 4.7). This is expected for both the ABC and numerical integration estimates since the prior distribution for the mixing parameter

$$\gamma \in [0, 0.25]$$

gives zero density in the posterior distribution of γ for values of $\gamma < 0$. However, one must also consider the posterior probability intervals in these cases. For each of

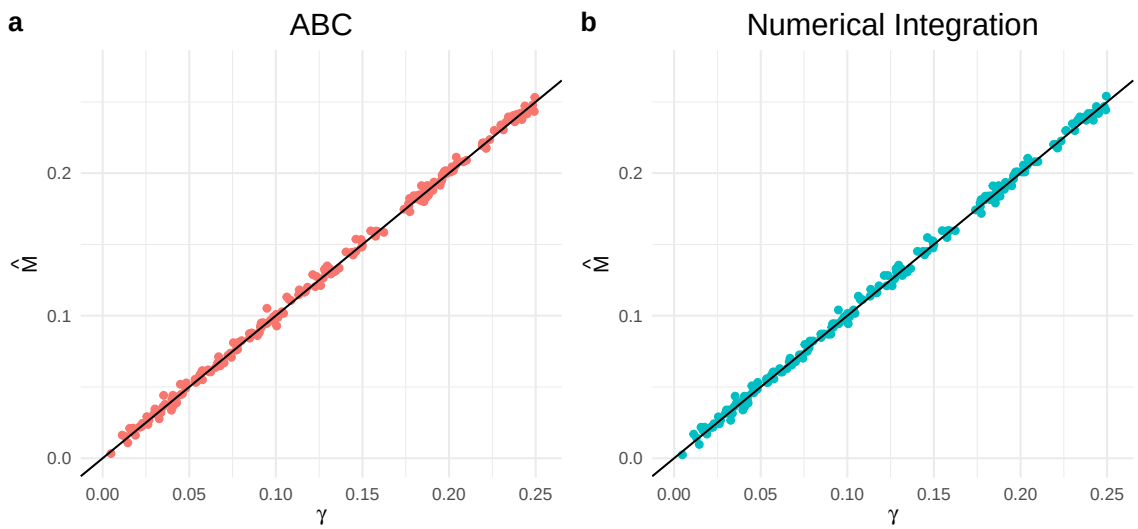


Figure 4.6: For Scenario A: Scatter-plots of $\hat{M}_i$, the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method. Note that we remove the cases where $\gamma = 0$, and take a random sample of 200 posterior estimates for ease of visualisation.

the fifty simulations where $\gamma = 0$, the conservative 95% posterior probability interval obtained via numerical integration included the boundary case $\gamma = 0$. In contrast, only eighteen of the the intervals obtained via ABC contained zero.

From Figure 4.8 we see that the apparent upward bias for $\hat{M}$ decreases quickly for values of $\gamma > 0$. In fact, for values of $\gamma \geq 2.967 \times 10^{-2}$, both linear models of the form

$$\hat{r}^{ABC} = \beta_\ell^{ABC} + \beta_u^{ABC} \times \gamma + \epsilon_i^{ABC}, \tag{4.10}$$

and

$$\hat{r}^{NI} = \beta_\ell^{NI} + \beta_u^{NI} \times \gamma + \epsilon_i^{NI}, \tag{4.11}$$

where $\epsilon_i^{ABC}, \epsilon_i^{NI} \sim N(0, \sigma^2)$ are independent, produce estimates of the coefficients that are not significantly different from $\beta_j^{ABC} = 0$ and $\beta_j^{NI} = 0$, where $j \in \{\ell, u\}$.

From Table 4.2 we also observe that, for both methods, the conservative 95% posterior probability intervals of the posterior median contain the true value of γ for more than 95% of simulations.

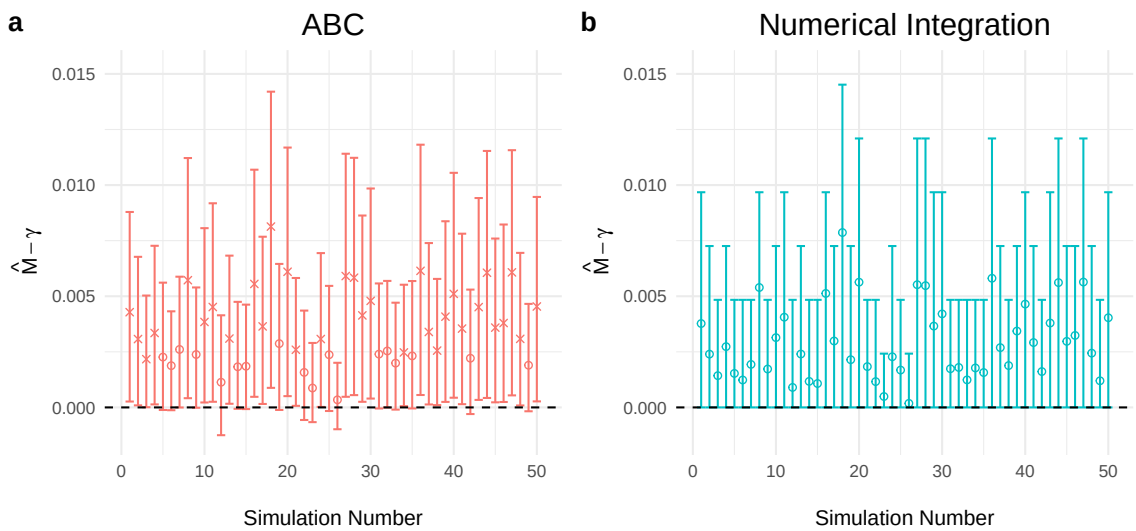


Figure 4.7: For Scenario A: Scatter-plots of $\hat{M}_i - \gamma$, the residual of the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method for the fifty simulations where $\gamma = 0$. Error bars indicating the conservative 95% posterior probability region, and plotting characters indicate that the interval contains zero ($\circ$) or did not ($\times$).

	is γ in CI
ABC	0.964
Integration	0.992

Table 4.2: For Scenario A, the proportion of simulations for which the 95% probability interval contained the true value of γ.

From Table 4.3 we observe that the residual of the posterior median was on average very close to zero, with corresponding sample standard deviation approximately 4.36 and 4.74 times greater than the sample means of the residuals for the ABC and numerical integration approaches respectively. Hence, we observe that our methods reliably predict the true value of γ.

The two methods of estimation produce extremely similar results. A correlation coefficient between the values of M^{ABC} and M^{NI} of 0.99995 is observed, and a

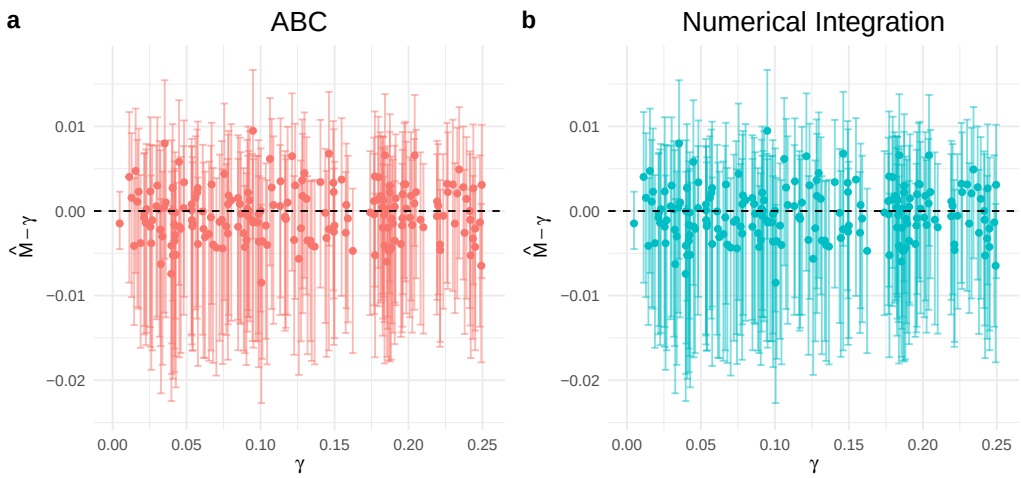


Figure 4.8: For Scenario A: Scatter-plots of $\hat{M}_i - \gamma$, the residual of the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method. Error bars indicate the conservative 95% posterior probability region. Note that we remove the cases where $\gamma = 0$, and take a random sample of 200 posterior estimates for ease of visualisation.

	$\bar{r}_{\hat{M}_\gamma}$	$s_{\hat{M}_\gamma}$
ABC	0.0007202246	0.003142405
Integration	0.0006703309	0.003179442

Table 4.3: Table of the sample median and sample standard deviation for the residuals of the approximate posterior median of γ, obtained via ABC and numerical integration, for Scenario A.

matched-pairs t-test yields no significant difference between the true mean values of M^{ABC} and M^{NI} ($p = 0.6616$).

Of the fifty simulations where $\gamma = 0$, and hence no hybridisation has occurred, the posterior probability interval obtained via ABC did not contain zero in 32 (64%) of the simulations, whereas the numerical integration posterior probability interval contained zero every time (see Figure 4.7). Clearly then, for the boundary case of $\gamma = 0$, the numerical integration approach performs better. However, from Figure 4.7

we can see that although the 95% posterior probability interval for the ABC method contained zero in only eighteen out of the fifty simulations, the lower bound of the posterior probability interval was very close to zero. In fact, the mean lower bound for intervals that did not contain zero was 2.87×10^{-4}.

We compare the difference in the performance of the methods for simulations where $\gamma > 0$. From Figure 4.9 we observe that M_i^{ABC} and M_i^{NI} are almost always within 1.5×10^{-3} of one another when $\gamma > 0$. However the 95% confidence interval about a generalised additive model (GAM), which we use to account for potential non-linearity, includes zero for all values of $\gamma > 0$, although the trend line is upwardly biased as $\gamma \to 0$. One could argue that the numerical integration approach, which is more accurate for small values of γ should be preferred then.

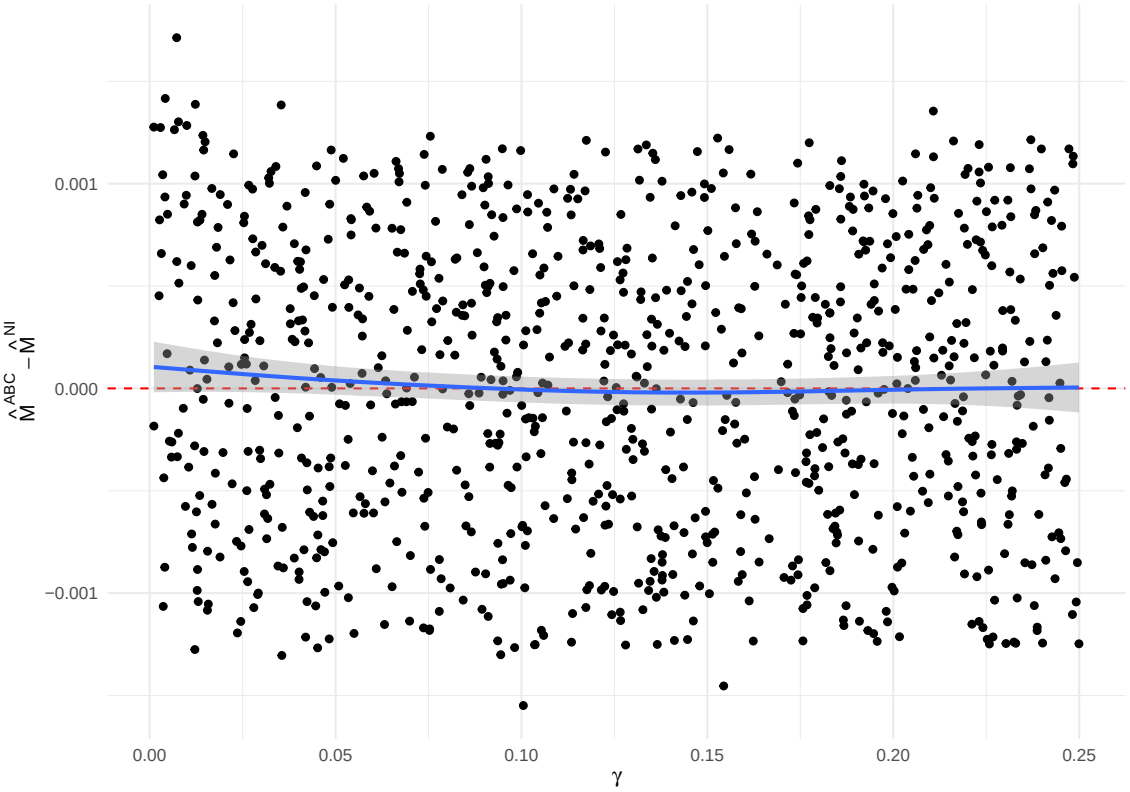


Figure 4.9: For Scenario A: a scatter-plot of the difference in estimates of γ, $\hat{M}_i^{ABC} - \hat{M}_i^{NI}$ with a trend line calculated using a generalised additive model, for $\gamma > 0$.

The methods had significantly different computation run times according to a matched pairs t-test, with a p-value $< 2.2 \times 10^{-16}$. The mean runtime for the numerical integration analyses was 15.2 seconds, compared to 44.79 seconds for the ABC method.

We investigated the average computational runtime for both methods for a range of precision parameters, namely the grid size for the numerical integration method, and the number of simulations for the ABC method (see Figure 4.10). For the numerical integration method we analysed data sets with grid sizes of $m = 75, 100, 125$ and 150, and for the ABC method we analysed data sets with the total number of simulations $N_M = 1 \times 10^5, 5 \times 10^5, 1 \times 10^6$ and 5×10^6. For each parameter value, we analysed fifty independent data sets. Computation runtimes were measured using the R-package `microbenchmark` [30]. Both methods grow exponentially in computational runtime, however it should be noted that the methods may not be directly comparable for these values of m and N_m, and so we cannot claim that either method is significantly computationally more efficient than the other for some specified level of accuracy.

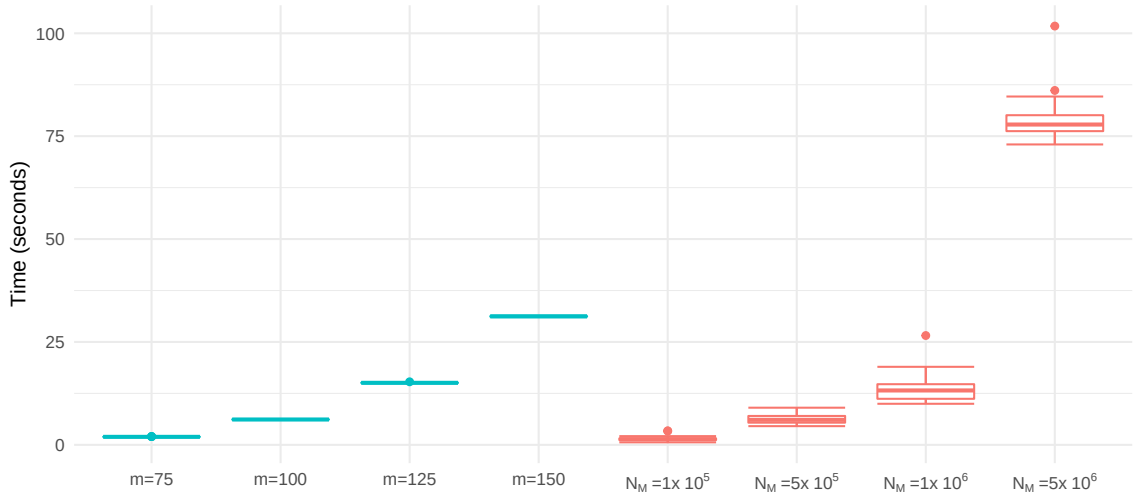


Figure 4.10: Boxplots comparing computational runtime for the numerical integration method (blue) and the ABC method (red) for varying values of the grid size m and number of simulations N_M.

We tested the sensitivity of the width of the 95% posterior probability interval for the

ABC method to changes in N_M to identify when the ABC method was sufficiently "accurate". Using the same sampled branch lengths as in the simulation study (for Scenario A), we simulated 1000 site patterns, however each data set had a value of $\gamma = 0.15$. We then randomly selected $N_m \in (1 \times 10^4, 1 \times 10^6)$, and chose ξ such that $\lfloor \xi N_M \rfloor = 500$ so that every corrected approximate posterior distribution was made up of the same number of observations. Surprisingly, we found that the number of simulations N_M was not significantly correlated with the interval width ($\rho_S = -0.019$, $p = 0.5476$) or the residual mean ($\rho_S = 0.007$, $p = 0.8223$), according to a Spearman correlation test, where ρ_S is the Spearman correlation coefficient. That is, even for relatively small numbers of simulations, the ABC method has seemingly converged to a relatively consistent estimated posterior distribution.

From this simulation study we have shown that under reasonable biological conditions, where a large amount of evolutionary time separates Species A and C (*i.e.* when β is relatively large compared to the extant branch tips), our methods perform well. It should be noted that while both the ABC and numerical integration methods performed similarly, both showed a clear positive bias for values of γ close to zero, and this bias was more pronounced for ABC.

4.6.3 Results for Scenario B

Next we simulate site pattern counts for Scenario B which differs from Scenario A only in that β is much smaller. Recall that for Scenario A we had that $\beta \in (35, 40)$, whereas for Scenario B we have that $\beta \in (3, 4)$.

From Figure 4.11 it can be seen that both methods again appear to approximate the true value of γ, yielding a Spearman sample correlation of $\rho_S = 0.9959$. However, the sample standard deviation of the $\widehat{M}$ has increased from approximately 3.1×10^{-3} to 2.2×10^{-2}, and the positive bias of $\widehat{M_i}$ for values of γ that are *relatively* close to zero has increased. Linear models of the form given in Equations (4.10) and (4.11) indicate significantly non-zero positive bias for values of $\gamma \geq 0.12$.

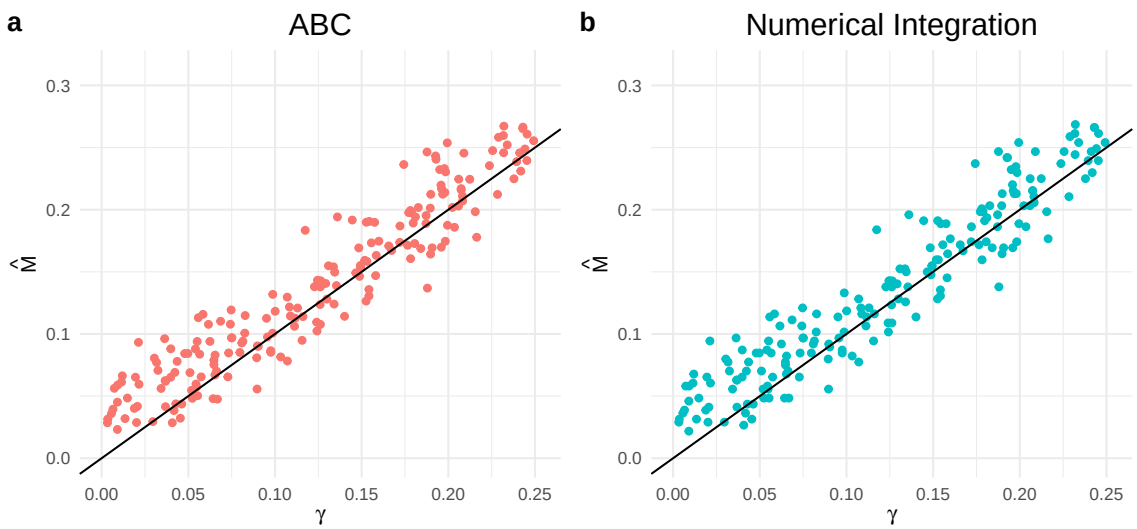


Figure 4.11: For Scenario B: Scatter-plots of $\hat{M}_i$, the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method. Note that we remove the cases where $\gamma = 0$, and take a random sample of 200 posterior estimates for ease of visualisation.

The positive bias of $\widehat{M}$ can be observed clearly in Figure 4.11, and this has now significantly affected the proportion of conservative 95% posterior probability intervals that contain the true value of γ. From Table 4.4 we observe that the intervals obtained via the numerical integration method appear to perform well as they contain the true value of γ for 98.2% of the simulated values, and the intervals obtained via the ABC approach contain the true value of γ for 94.4% of the simulations.

	γ in CI
ABC	0.944
Integration	0.982

Table 4.4: For Scenario B, the number of simulations for which the 95% probability interval contained the true value of γ.

However, for the fifty simulations where $\gamma = 0$, the intervals obtained via the numerical integration method contain the true value of γ for 72% of the simulated values,

whereas the intervals obtained via the ABC approach never contained the true value of γ (see Figure 4.12). This indicates an increase in Type I error for smaller relative values of β. It should be noted when the posterior probability intervals did not contain zero, that the mean of the lower bounds for the posterior probability intervals was 2.24×10^{-3} and 1.55×10^{-3} for the ABC and numerical integration methods, respectively.

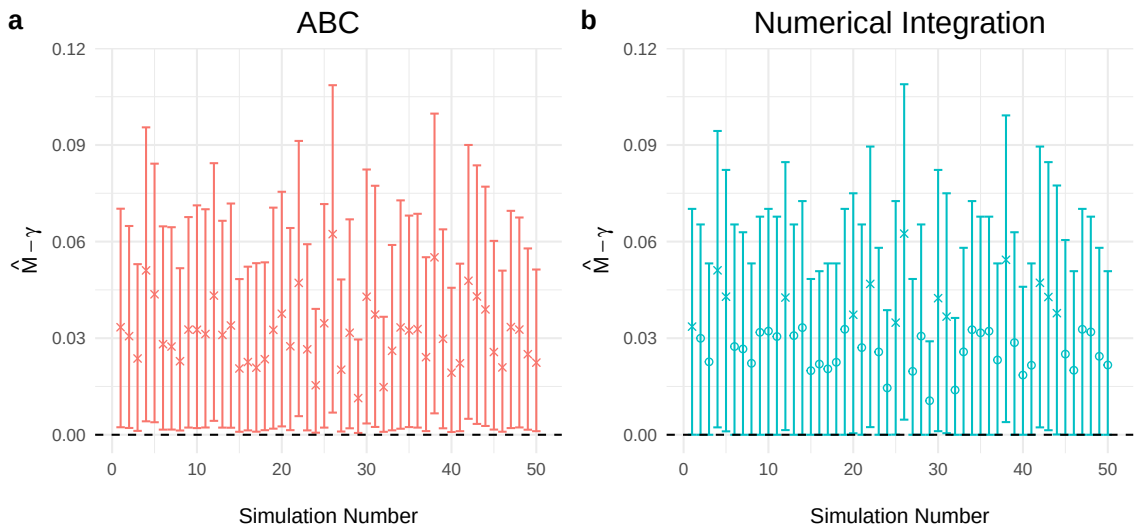


Figure 4.12: For Scenario B: Scatter-plots of $\hat{M}_i - \gamma$, the residual of the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method for the fifty simulations where $\gamma = 0$. Error bars indicating the conservative 95% posterior probability region, and plotting characters indicate that the interval contains zero ($\circ$) or does not contain zero ($\times$).

To quantify the effect of small values of γ, we fit logistic regression models of the form

$$P(Y_i^{ABC} = 1 \big| \gamma = \gamma_i) = \frac{1}{1 + e^{-(\beta_\ell^{ABC} + \beta_u^{ABC}\gamma_i)}}$$

and

$$P(Y_i^{NI} = 1 \big| \gamma = \gamma_i) = \frac{1}{1 + e^{-(\beta_\ell^{NI} + \beta_u^{NI}\gamma_i)}},$$

where Y_i^{ABC} and Y_i^{NI} equal one if the ith posterior probability interval contains γ_i, and zero if it does not, for the ABC and numerical integration methods respectively.

These models indicate that γ is a significant predictor of whether or not the posterior probability interval contains the true value of γ_i for both methods (p-values of 4.24×10^{-10} and 1.11×10^{-3}, respectively). We also find that we can expect to have approximately 95% probability of the posterior probability interval containing γ when $\gamma \geq 0.0146$ and 0.00714 for the ABC and numerical integration methods respectively. Hence, while the point estimate of γ_i, $\widehat{M_i}$, is certainly upwardly biased, the approximate posterior probability interval appears to perform well, even for very small, non-zero values of γ. However, it must be conceded that our ABC method cannot be trusted to identify cases where $\gamma = 0$ for relatively small values of β.

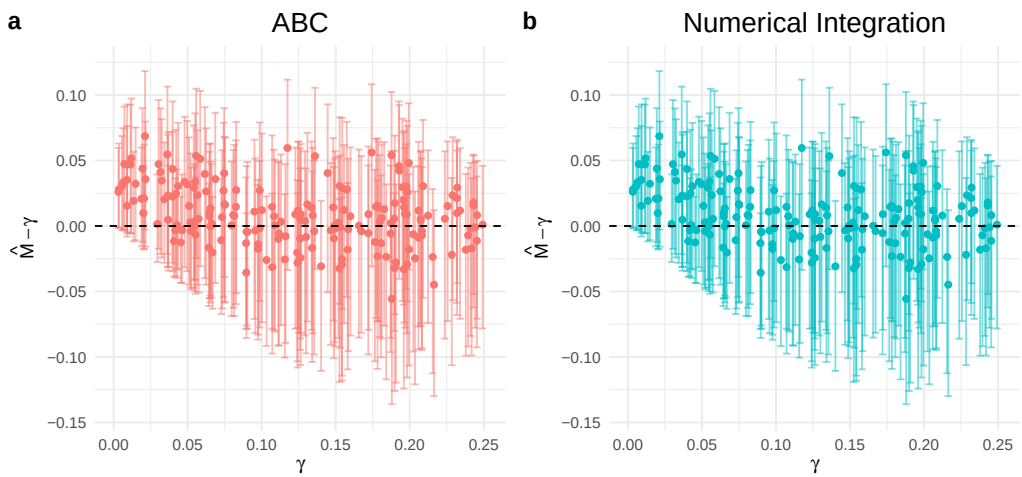


Figure 4.13: For Scenario B: Scatter-plots of $\hat{M}_i - \gamma$, the residual of the estimate of the posterior median of γ, and the true value of γ using (a) the ABC method, and (b) the numerical integration method. Error bars indicate the conservative 95% posterior probability region. Note that we remove the cases where $\gamma = 0$, and take a random sample of 200 posterior estimates for ease of visualisation.

4.7 Application to Empirical Data

To test the performance of our method on real data, we aim to infer the proportion of Neanderthal ancestry in nine anatomically modern humans in Europe after the so-called "out of Africa expansion", but prior to the last glacial maximum (LGM) approximately 26.5 kya [11, 48]. We compare our results to previously published results from a study by Fu *et al.* [18]. Specifically, we select samples obtained from the Ostuni Cave in Italy, the Dolní Věstonice and Pavlov1 archaeological sites in the Czech Republic, the Peştera Muierii cave system in Romania, and a single sample (Kostenki12) from the Kostenki archaeological site in the Pokrovsky Valley of Russia. We also selected only samples such that they were obtained using the 3.7M SNP Panel [31]. Note that since our model cannot incorporate multiple admixture events, we ignore ancient European samples from after the LGM, and the resulting population turnover which brought with it a far more complicated ancestry for modern humans [35].

To estimate the proportion of Neanderthal ancestry in each of the ancient European samples dating from between 27.6 and 31.282 kya, we compare the genomes nine genomes to the genome of a modern individual from Yoruba, and the genome of a Neanderthal individual from the Altai Mountains in Russia (dating from approximately 50,300 years before present) [12, 18, 38]. The allocations for the different species (A, B and C) on the admixture graph are given in Figure 4.14. Note that for each of the nine analyses, we use the same Neanderthal and Yoruba samples to estimate the specific Neanderthal ancestry for each ancient European individual.

It should be noted that in the study by Fu *et al.*, non-admixed humans are represented by a pool of nine genomes from West and Central African modern samples. These samples come from the Mbuti, Yoruba peoples (both West African) and the Mende people (Central Africa). In contrast, we use a single sample obtained from a Yoruba individual from the 1000 Genomes Project. Similarly, Fu *et al.* use a pooled sample of an Altai Neanderthal and a Siberian Devonian individual, where we use only the

Altai Neanderthal genome for our archaic human sample. Due to these differences, we expect subtly different, although relatively consistent, estimates of Neanderthal ancestry for our analysis.

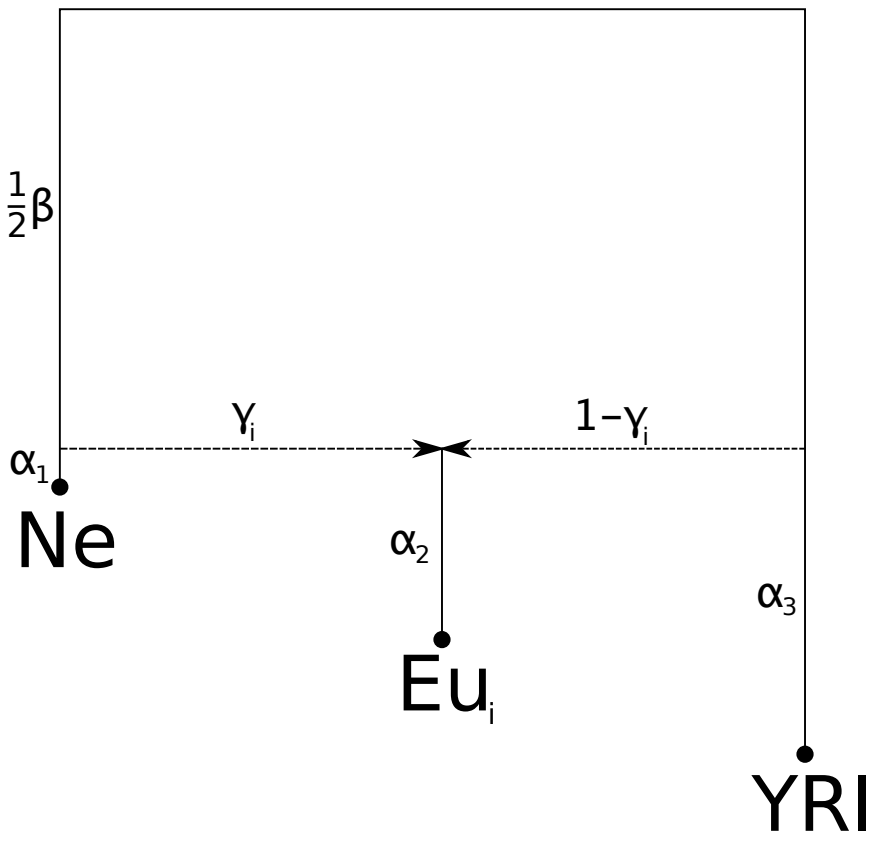


Figure 4.14: The parameterisation of the admixture graph used to estimate the proportion of Neanderthal (Ne) and Yoruba (YRI) ancestry in an ancient European (EU$_i$).

Ancient European sequences were obtained from the European Nucleotide Archive (accession number PRJEB13123) [18]. We use sample NA18488 from the 1000 Genomes Project for the Yoruba sample, and we use the Altai Neanderthal sample first published by Prüfer *et al.* (accession number ERP002097) [12, 38]

For the prior distributions for the parameters of the admixture graph, we use a prior for $\beta^* \in U\,[5, 50]$, for each analysis, as significant evolutionary time has passed

since the most recent common ancestor of Neanderthals and African anatomically modern humans [15]. For the prior distribution of $\boldsymbol{\alpha}$, we note that the relative branch lengths are unlikely to be equal, and so a Dirichlet distribution with equal expected values makes little sense here. The modern Yoruba individual is likely to have the longest relative branch length, due to the longest amount of calendar time between the admixture event and the sampling time, and a relatively large population size, and so we set $a_3^* = 10$. The sampling time of the Altai Neanderthal (50.3 ± 2.2 kya) is relatively close to the estimated admixture time, and so will have a very short relative branch length, and so we set $a_1^* = \frac{1}{4}$ for every analysis. Finally, the pre-ice age European will have branch lengths dependent on their sampling times.

For the least ancient European sample (27.6 kya) we set $a_{2,i}^* = 1$, and for the most ancient European sample (31.282 kya) we set $a_{2,i}^* = 2$. For the remaining samples that fell in between these sampling dates, we used a simple linear interpolation to choose $a_{2,i}^*$, *i.e.*, for the ith sampling time $27.6 \times 10^3 \leq t \leq 31.282 \times 10^3$, we set

$$a_{2,i}^* = 1 + \frac{t - 27.6 \times 10^3}{31.282 \times 10^3 - 27.6 \times 10^3}.$$

Finally, we normalise the vector $\boldsymbol{a}_i^* = \left\{\frac{1}{4}, a_{2,i}^*, 10\right\}$, to control the total variance of the prior distribution for $\boldsymbol{\alpha}$, by setting hyper parameters

$$\boldsymbol{a}_i = (a_1, a_2, a_3)_i = \frac{\boldsymbol{a}^*}{1/4 + a_{2,i}^* + 10},$$

that is

$$\boldsymbol{\alpha}_i \sim \text{Dirichlet}\left(\boldsymbol{a}_i\right).$$

A sensitivity analysis showed no significant change in results for the arbitrarily chosen endpoints of one and two for the $a_{2,i}^*$, as long as the upper bound for the $a_{2,i}^*$ was no greater than half of a_3^*. For a complete table of details for the samples used in the analyses, see Table 4.5.

In Sections 4.6.2 and 4.6.3 we showed that estimates of the ancestry proportion γ can be strongly upwardly biased for small values of γ, especially for the ABC method.

Proportions of Neanderthal ancestry in admixed anatomically modern humans have been shown to be less than 2% in modern populations, and less than 10% for ancient samples [44, 18]. Hence, we employ only the numerical integration approach here. For the following analysis we use a grid size defined by $m = 125$.

Our results appear to be consistent with the results obtained via the ratio of f_4 statistics by Fu *et al.* [18]. For every sample, except the Vestonice43 sample, the posterior probability interval we calculated contained the point estimate obtained by Fu *et al.* (see Figure 4.15). In the case of Vestonice43, we estimate 5.5% Neanderthal ancestry, with an upper bound of 6.29% ancestry. Fu *et al.* report a point estimate of 6.9% ancestry, with a lower bound of 5.2%. So, while our posterior probability interval does not contain the estimate obtained by Fu *et al.*, our point estimate is contained in their confidence interval (see Table 4.6).

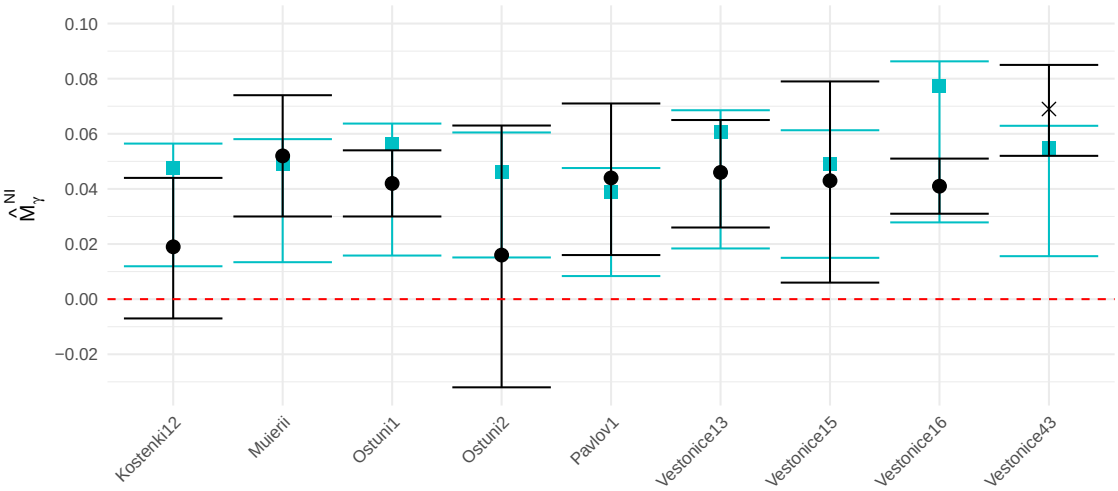


Figure 4.15: Scatter-plots of $\hat{M}_i$, the proportion or Neanderthal ancestry in pre-ice age European humans obtained via the numerical integration method. Error bars indicate the conservative 95% posterior probability region.

In two samples Fu *et al.* report negative lower bounds for the confidence interval of the proportion of Neanderthal ancestry, -0.7% and -3.2% for Konstenki12 and Ostuni2 respectively. Hence, one would reject any significant evidence for Neanderthal

ancestry in these samples. We report lower bounds of 1.19% and 1.51% Neanderthal ancestry for Konstenki12 and Ostuni2 respectively, which are well above the lower bounds suggested by Fu *et al.*. Further, we do not believe that there was a systematic inflation in the point estimates and margin of error reported in our analyses, compared to that of Fu *et al.*. Our analysis reported point estimates of Neanderthal ancestry greater than reported by Fu *et al.* in only six out of nine of the analyses. Similarly, our analyses yield a margin of error that was greater than the margin of error reported by Fu *et al.* in only five out of nine of the analyses.

4.8 Conclusion

In this work we developed a method for estimating γ, the proportion of ancestry on a three taxon tree, when Species B is known to be a hybrid offspring of Species A and Species C. We showed through simulation that our method was able to accurately estimate γ for a range of biologically reasonable scenarios. We then showed that our method was able to produce estimates of γ for pre-ice age European humans, consistent with those obtained via the popular ratio of f_4 statistics [34].

However, we noted that as the amount of evolutionary time since the MRCA of Species A and C and the admixture event becomes small relative to the total amount of evolutionary time for all species since the admixture event, our method can be upwardly biased for small values of γ. This was particularly so for the ABC method.

We claim that our method is relatively computationally fast, although we omitted the computation time associated with the necessary pre-processing to calculate the site pattern counts. In reality, this is where the vast majority of the computation time is spent for our method. Preprocessing also plays a large role in the computation time for the ratio of the f_4 statistics, although the bootstrap method for estimating the standard deviation of the ratio of the f_4 statistics is also computationally expensive.

Our reanalysis of the pre-ice age European samples yielded consistent point estimates

of ancestry proportions, that were at times greater than, less than, or roughly equal to the results from Fu *et al.* Similarly, we also recovered margins of error with no systematic relationship with those obtained via the ratio of f_4 statistics. Hence, we have a novel method for estimating the proportion of ancestry proportions that could be used to strengthen results obtained via the ratio of f_4 statistics. In the following chapter we use both our method and the ratio of f_4 statistics to estimate, and provide further evidence for the estimated proportion of ancestry in some ancient bison samples.

EU	Age (kya)	n_1	n_2	n_3	a_1	a_2	a_3	Longitude	Latitude
Kostenki12	3.242×10^4	265	252	5071	0.1525	0.02067	0.8268	39.3	51.23
Muierii	3.33×10^4	358	297	6569	0.1633	0.02041	0.8163	23.46	45.11
Ostuni1	2.762×10^4	786	671	1.271×10^4	0.08889	0.02222	0.8889	17.57	40.73
Ostuni2	2.898×10^4	57	26	1138	0.1078	0.02176	0.8704	17.57	40.73
Pavlov1	3.026×10^4	148	163	3472	0.125	0.02134	0.8536	16.39	48.53
Vestonice13	3.087×10^4	583	444	8769	0.133	0.02115	0.8459	16.39	48.53
Vestonice15	3.087×10^4	130	82	2391	0.133	0.02115	0.8459	16.39	48.53
Vestonice16	3.087×10^4	1448	1149	1.685×10^4	0.133	0.02115	0.8459	16.39	48.53
Vestonice43	3.087×10^4	508	402	8466	0.133	0.02115	0.8459	16.39	48.53

Table 4.5: Metadata for the nine ancient European samples used in the analyses.

EU	$\hat{M}_{\gamma}^{NI}$	NI CI	Fu Estimate	Fu CI
Kostenki12	0.048	(0.01193,0.05645)	0.019	(-0.007,0.044)
Muierii	0.049	(0.01341,0.05806)	0.052	(0.03,0.074)
Ostuni1	0.056	(0.01582,0.06371)	0.042	(0.03,0.054)
Ostuni2	0.046	(0.01512,0.06048)	0.016	(-0.032,0.063)
Pavlov1	0.039	(0.00838,0.04758)	0.044	(0.016,0.071)
Vestonice13	0.06	(0.01839,0.06855)	0.046	(0.026,0.065)
Vestonice15	0.049	(0.01499,0.06129)	0.043	(0.006,0.079)
Vestonice16	0.077	(0.02785,0.08629)	0.041	(0.031,0.051)
Vestonice43	0.055	(0.01559,0.0629)	0.069	(0.052,0.085)

Table 4.6: Neanderthal ancestry proportions ($\hat{M}_{\gamma}^{NI}$) for ancient European samples estimated via numerical integration and from Fu *et al.*

Chapter 5

An Application of Modelling Admixture via Site Pattern Distributions

5.1 Introduction

In this chapter we present the paper titled "Early cave art and ancient DNA record the origin of European bison". This paper was published in *Nature Communication* on the 18th of March, 2017 and is an application of the work presented in Chapter 4.

During the Late Pleistocene, between 11.7 and 126 thousand years before present (kya), a close relative of the American bison, the Steppe bison (*Bison priscus*) and the ancestor of modern cattle, the aurochs (*Bos primigenius*) were the two forms of recognised bovids in Europe, and were extremely well represented in the fossil record. At around 11.7 kya, the wisent (*Bison bonasus*) suddenly appears in the early Holocene fossil record shortly after the disappearance of the Steppe bison during the megafaunal extinctions of the Late Pleistocene.

In an effort to understand the replacement of Steppe bison by wisent, 38 new samples,

ranging from 14 to (greater than) 50 kya were sequenced. A phylogenetic analysis of mtDNA revealed the presence of a previously undetected clade of bison, tentatively titled Clade-X. Clade-X was found to be most closely related to cattle, wisent and aurochs, and relatively distantly related to Steppe bison, and American bison. Of interest was that modern and ancient wisent samples were found to form a single, separate clade from Clade-X.

A phylogenetic analysis of 10,000 genome-wide nuclear sites yielded, for the nuclear genome, that Steppe bison, ancient wisent and Clade-X form a clade closer to American bison than modern wisent samples, but close to two pre-bottleneck wisent samples. This incongruence in mitochondrial and nuclear genomes suggested a hybridisation event may have occurred at some point in the history of the wisent. We found strong evidence to suggest that ancient wisent are comprised of approximately 10% Aurochs ancestry and 90% Steppe bison ancestry, and are the result of a female Aurochs and male Steppe bison mating.

Wisent living outside of the region inhabited by the hybrid species were reintroduced to Clade-X (at least 20 kya), likely due to the arrival of the last glacial maximum. These species, wisent and Clade-X, would have had differing morphologies due to the strong Steppe contribution to the Clade-X nuclear genome. However, due to the maternal inheritance of mtDNA, Clade-X would have appeared to be wisent-like from mtDNA, but Steppe-like in morphology. It seems then that our results also agreed with the cave art of the last 30 thousand years in Europe which had actually recorded the change in bison morphology.

Our contribution to this research was in the writing of the manuscript, all bioinformatic analyses of the sequence data, and in the interpretation of these results. One specific contribution was the new method presented in Chapter 4, to strengthen the results of the ratio of f_4 statistics used to estimate the proportion of hybridisation in the ancient wisent samples.

Here we include the main publication, but we also include the supplementary in-

formation for completeness. We direct the reader to our contribution on Page 17 of the supplementary information, although this methodology is fully described in Chapter 4.

5.2 Statement of Authorship

Statement of Authorship

Title of Paper	Early cave art and ancient DNA record the origin of European bison
Publication Status	Published
Publication Details	Soubrier, J., et al. 2016. "Early Cave Art and Ancient DNA Record the Origin of European Bison." *Nature Communications* 7 (October): 13158. doi:10.1038/ncomms13158.

Principal Authors

Julien Soubrier			
Contribution to the Paper	Designed experiments. Performed bioinformatics analyses: processed and analysed NGS data, phylogenetics. Analysed and interpreted results. Wrote the paper with help from all co-authors.		
Overall percentage (%)	40		
Signature		Date	09/10/18

Adam Rohrlach(Candidate)			
Contribution to the Paper	Designed new method for inference of admixture parameter via mixture multinomial modelling site pattern counts. Performed post-simulation ABC analysis of sequence data to infer shared history of Bison X		
Overall percentage (%)	20		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am one of two primary authors of this paper.		
Signature		Date	24/10/2018

Co-Author Contributions

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

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

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

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

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Designed experiments, provided samples, interpreted results. Wrote the paper with help from all co-authors.		
Signature		Date	12/10/18

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Designed experiments, laboratory work, analyses, interpreted results.		
Signature		Date	18/10/18

Name of Co-Author	Graham Gower		
Contribution to the Paper	Designed experiments. Performed bioinformatics analyses: processed and analysed nuclear data (Paleomix, Principal Component Analysis, D and f statistics, Hypergeometric test, sensitivity analysis, co-contributor of ABC analysis). Analysed and interpreted results. Wrote the paper with help from all co-authors.		
Signature		Date	12.10.18

Name of Co-Author	Kieren Mitchell		
Contribution to the Paper	Designed and supervised laboratory experiments, advised on analyses, interpreted results.		
Signature		Date	12.10.18

Name of Co-Author	Wolfgang Haak		
Contribution to the Paper	Collected Bison samples from Russia, participated in early study design.		
Signature		Date	21/11/2016

Name of Co-Author	Johannes Krause		
Contribution to the Paper	Together with Frauke Langbein and Alexander Immel processed and provided sequence data from Bison from the Ukraine (provided to him by Marie-Anne Julien).		
Signature		Date	15.11.16

Name of Co-Author	Frauke Langbein		
Contribution to the Paper	Together with her supervisor Johannes Krause and Alexander Immel processed and provided sequence data from Bison from the Ukraine (provided by Marie-Anne Julien to Johannes Krause). Signed by Johannes Krause on behalf of her, since she has not been reachable.		
Signature		Date	15.11.16

Name of Co-Author	Alexander Immel		
Contribution to the Paper	Together with his supervisor Johannes Krause and Frauke Langbein processed and provided sequence data from Bison from the Ukraine (provided by Marie-Anne Julien to Johannes Krause).		
Signature		Date	15.11.16

Name of Co-Author	Amelie Scheu		
Contribution to the Paper	Provided samples, background information and data.		
Signature		Date	18.11.2016

Name of Co-Author	Beth Shapiro		
Contribution to the Paper	Laboratory work, interpretation of results, comments on manuscript.		
Signature		Date	16 Nov 2016

Name of Co-Author	Colin Groves		
Contribution to the Paper	Provided morphological and taxonomic background; suggested the link with cave art. Wrote the paper with help from all co-authors.		
Signature		Date	14/11/16

Name of Co-Author	David Chivall		
Contribution to the Paper	Radiocarbon dating of bison samples		
Signature		Date	18th November 2016

Name of Co-Author	Emilia Hofman-Kamińska		
Contribution to the Paper	Provided samples, interpretations of results and comments on the study.		
Signature		Date	18.11.2016

Name of Co-Author	Federica Fontana		
Contribution to the Paper	Sample collecting. Data for sample contextualisation (Riparo Tagliente, IT)		
Signature		Date	19 November 2016

Name of Co-Author	Gennady Baryshnikov		
Contribution to the Paper	Bone material from field excavations.		
Signature		Date	14.11.2016

Name of Co-Author	Jared Decker		
Contribution to the Paper	Provided feedback on interpretation of the results. Along with Jeremy Taylor and Bob Schnabel, provide modern bison data.		
Signature		Date	14 November 2016

Name of Co-Author	Greger Larson		
Contribution to the Paper	Designed and carried out experiments. Obtained samples. Wrote the paper with the help from all co-authors.		
Signature		Date	22/11/2016

Name of Co-Author	Jerry Taylor		
Contribution to the Paper	Provided samples/data. Edited manuscript.		
Signature		Date	November 14, 2016.

Name of Co-Author	Johannes van der Plicht		
Contribution to the Paper	provided radiocarbon dates		
Signature		Date	14 november 2016

Name of Co-Author	Ayla van Loenen		
Contribution to the Paper	Performed laboratory genetic analyses of mitochondrial and nuclear data that contributed towards the body of genetic data analysed in this paper, initial data processing steps of aforementioned genetic data, edited manuscript.		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis.		
Signature		Date	12/8/18

Name of Co-Author	Vladimir Doronichev		
Contribution to the Paper	Contributed samples and provided comments on this study		
Signature		Date	14.11.2016

Name of Co-Author	Liubov Golovanova		
Contribution to the Paper	Contributed samples and provided comments on this study		
Signature		Date	14.11.2016

Name of Co-Author	Ludovic Orlando		
Contribution to the Paper	Provided feedback in data analyses and interpretation.		
Signature		Date	2016.12.01

Name of Co-Author	Małgorzata Tokarska		
Contribution to the Paper	Supplying samples, co-editing of the manuscript		
Signature		Date	14.11. 2016

Name of Co-Author	Michael Lee		
Contribution to the Paper	Assisted with phylogenetic analyses and data interpretation		
Signature		Date	14.11.16

Name of Co-Author	Pavel Kosintsev		
Contribution to the Paper	Provided samples, interpreted results		
Signature		Date	22.11.16

Name of Co-Author	Pere Bover		
Contribution to the Paper	Performed laboratory work, submission of sequences to GenBank and general comments to the manuscript		
Signature		Date	20/12/2017

Name of Co-Author	Rafal Kowalczyk		
Contribution to the Paper	Provided samples, analysed and interpreted the results.		
Signature		Date	18 11 2016

Name of Co-Author	Ruth Bollongino		
Contribution to the Paper	Provided some mt-sequences and samples, discussed results and manuscript		
Signature		Date	16/11/2016

Name of Co-Author	Simon Ho		
Contribution to the Paper	Provided advice on phylogenetic analysis. Edited the draft manuscript.		
Signature		Date	14-Nov-16

Name of Co-Author	Stephen M. Richards		
Contribution to the Paper	Designed experiments, laboratory work, analyses, interpreted results.		
Signature		Date	14/11/16

Name of Co-Author	Tom Higham		
Contribution to the Paper	AMS radiocarbon dating of bone collagen extracts		
Signature		Date 14/11/2016	

To whom it may concern,

As the Director of the Australian Centre for Ancient DNA, the lab at which the significant proportion of the work for the publication "*Early cave art and ancient DNA record the origin of European bison*" was performed, I certify that candidate Adam Benjamin Rohrlach completed the work as indicated in the Statement of Authorship.

Unfortunately Adam was unable to obtain the signatures of Joachim Burger, Kefei Chen, Evelyne Crégut-Bonnoure, Katerina Douka, Damien Fordham, Carole Fritz, Jan Glimmerveen, Antonio Guerreschi, Marie-Anne Julien, Oleksandra Krovota, Robert Schnabel, Gilles Tosello, Jean-Denis Vigne and Oliver Wooley. However, I can confirm that Adam made significant efforts to try and obtain statements from all authors.

Sincerely,

Professor Alan Cooper

nature COMMUNICATIONS

ARTICLE

Received 22 Apr 2016 | Accepted 9 Sep 2016 | Published 18 Oct 2016

DOI: 10.1038/ncomms13158 OPEN

Early cave art and ancient DNA record the origin of European bison

Julien Soubrier[1,*], Graham Gower[1,*], Kefei Chen[1], Stephen M. Richards[1], Bastien Llamas[1], Kieren J. Mitchell[1], Simon Y.W. Ho[2], Pavel Kosintsev[3], Michael S.Y. Lee[4,5], Gennady Baryshnikov[6], Ruth Bollongino[7], Pere Bover[1,8], Joachim Burger[7], David Chivall[9], Evelyne Crégut-Bonnoure[10,11], Jared E. Decker[12], Vladimir B. Doronichev[13], Katerina Douka[9], Damien A. Fordham[14], Federica Fontana[15], Carole Fritz[16], Jan Glimmerveen[17], Liubov V. Golovanova[13], Colin Groves[18], Antonio Guerreschi[15], Wolfgang Haak[1,19], Tom Higham[9], Emilia Hofman-Kamińska[20], Alexander Immel[19], Marie-Anne Julien[21,22], Johannes Krause[19], Oleksandra Krotova[23], Frauke Langbein[24], Greger Larson[25], Adam Rohrlach[26], Amelie Scheu[7], Robert D. Schnabel[12], Jeremy F. Taylor[12], Małgorzata Tokarska[20], Gilles Tosello[27], Johannes van der Plicht[28], Ayla van Loenen[1], Jean-Denis Vigne[29], Oliver Wooley[1], Ludovic Orlando[30,31], Rafał Kowalczyk[20], Beth Shapiro[32,33] & Alan Cooper[1]

The two living species of bison (European and American) are among the few terrestrial megafauna to have survived the late Pleistocene extinctions. Despite the extensive bovid fossil record in Eurasia, the evolutionary history of the European bison (or wisent, *Bison bonasus*) before the Holocene (<11.7 thousand years ago (kya)) remains a mystery. We use complete ancient mitochondrial genomes and genome-wide nuclear DNA surveys to reveal that the wisent is the product of hybridization between the extinct steppe bison (*Bison priscus*) and ancestors of modern cattle (aurochs, *Bos primigenius*) before 120 kya, and contains up to 10% aurochs genomic ancestry. Although undetected within the fossil record, ancestors of the wisent have alternated ecological dominance with steppe bison in association with major environmental shifts since at least 55 kya. Early cave artists recorded distinct morphological forms consistent with these replacement events, around the Last Glacial Maximum (LGM, ~21–18 kya).

[1] Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia. [2] School of Biological Sciences, University of Sydney, Sydney, New South Wales 2006, Australia. [3] Institute of Plant and Animal Ecology, Russian Academy of Sciences, 202 8 Marta Street, 620144 Ekaterinburg, Russia. [4] School of Biological Sciences, Flinders University, South Australia 5001, Australia. [5] Earth Sciences Section, South Australian Museum, North Terrace, Adelaide, South Australia 5000, Australia. [6] Zoological Institute RAS, Universitetskaya Naberezhnaya 1, 199034 St Petersburg, Russia. [7] Palaeogenetics Group, Institute of Anthropology, University of Mainz D-55128, Mainz, Germany. [8] Department of Biodiversity and Conservation, Institut Mediterrani d'Estudis Avançats (CSIC-UIB), Cr. Miquel Marquès 21, 07190 Esporles, Illes Balears. [9] Oxford Radiocarbon Accelerator Unit, Research Laboratory for Archaeology and the History of Art, University of Oxford, Oxford OX1 3QY, UK. [10] Museum Requien, 67 rue Joseph Vernet, 84000 Avignon, France. [11] Laboratoire TRACES UMR5608, Université Toulouse Jean Jaurès - Maison de la Recherche, 5 allée Antonio Machado, 31058 Toulouse, France. [12] Division of Animal Sciences, University of Missouri, Columbia, Missouri 65211, USA. [13] ANO Laboratory of Prehistory, 14 Linia 3e 11, 199034 St Petersburg, Russia. [14] Environment Institute and School of Biological Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia. [15] Dipartimento di Studi Umanistici, Università degli Studi di Ferrara, 12 Via Paradiso, 44121 Ferrara, Italy. [16] CNRS, TRACES, UMR 5608 et CREAP, MSHS Toulouse, USR 3414, Maison de la Recherche, 5 allées Antonio Machado, 31058 Toulouse, France. [17] CERPOLEX/Mammuthus, Anna Paulownastraat 25A, NL-2518 BA Den Haag, The Netherlands. [18] School of Archaeology and Anthropology, Australian National University, Building 14, Canberra, Australian National University 0200, Australia. [19] Max Planck Institute for the Science of Human History, 07745 Jena, Germany. [20] Mammal Research Institute, Polish Academy of Sciences, Waszkiewicza 1c, 17-230 Białowieża, Poland. [21] Department of Archaeology, Centre for the Archaeology of Human Origins, University of Southampton, Avenue Campus, Southampton SO17 1BF, UK. [22] Unité Histoire naturelle de l'Homme préhistorique (UMR 7194), Sorbonne Universités, Muséum national d'Histoire narurelle, CNRS, 1 rue René Panhard, 75013 Paris, France. [23] Department of Stone Age, Institute of Archaeology, National Ukrainian Academy of Science, 04210 Kiev, Ukraine. [24] Institute for Archaeological Sciences, Archaeo and Palaeogenetics, University of Tübingen, 72070 Tübingen, Germany. [25] Palaeogenomics and Bio-Archaeology Research Network, Research Laboratory for Archaeology, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK. [26] School of Mathematical Sciences, The University of Adelaide, Adelaide, South Australia 5005, Australia. [27] Chercheur associé, CREAP, MSHS Toulouse, URS 3414, Maison de la Recherche, 5 allées Antonio Machado, 31058 Toulouse, France. [28] Centre for Isotope Research, Radiocarbon Laboratory, University of Groningen, Nijenborg 4, NI-9747 AG Groningen, The Netherlands. [29] Centre National de la Recherche Scientifique, Muséum National d'Histoire Naturelle, Sorbonne Universités, UMR7209, 'Archéozoologie, archéobotanique: sociétés, pratiques et environnements', CP56, 55 rue Buffon, 75005 Paris, France. [30] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, ØsterVoldgade 5-7, Copenhagen 1350K, Denmark. [31] Université de Toulouse, University Paul Sabatier, Laboratoire AMIS, CNRS UMR 5288, 37 Allées Jules Guesde, Toulouse 31000, France. [32] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [33] UCSC Genomics Institute, University of California Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.S. (email: julien.soubrier@adelaide.edu.au) or to A.C. (email: alan.cooper@adelaide.edu.au).

The extensive Late Pleistocene fossil record of bovids in Europe consists of two recognized forms: the aurochs (*Bos primigenius*), ancestor of modern cattle, and the mid/late Pleistocene 'steppe bison' (*Bison priscus*), which also ranged across Beringia as far as western Canada[1,2]. The European bison, or wisent (*Bison bonasus*), has no recognized Pleistocene fossil record and seems to suddenly appear in the early Holocene (<11.7 kya)[3,4], shortly after the disappearance of the steppe bison during the megafaunal extinctions of the Late Pleistocene[5–7]. The Holocene range of wisent included all lowlands of Europe, and several highland areas of eastern Europe (where it was termed the Caucasian form *B. bonasus caucasicus*) but range reduction and hunting by humans brought the species close to extinction, with modern populations descending from just 12 mostly Polish individuals that lived in the 1920s (refs 8,9). Nuclear DNA sequences and the morphology of the wisent show close similarities to American bison (*B. bison*), but wisent mitochondrial DNA (mtDNA) indicates a closer relationship with cattle. This suggests some form of introgression from cattle or a related *Bos* species[10–12], potentially associated with the recent extreme bottleneck event.

Both aurochs and bison feature heavily in Palaeolithic cave art, with 820 depictions displaying bison individuals (~21% of known cave ornamentation[13]). The diversity of bison representations has been explained as putative cultural and individual variations of style through time, since the steppe bison was assumed to be the only bison present in Late Paleolithic Europe[14–16]. However, two distinct morphological forms of bison (Fig. 1, Supplementary Information section) are clearly apparent in cave art: a long-horned form similar to modern American bison (which are thought to be descended from steppe bison), with very robust forequarters and oblique dorsal line, and a second form with thinner double-curved horns, smaller hump and more balanced body proportions, similar to wisent. The former is abundant in art older than the Last Glacial Maximum (LGM, ~22–18 kya), while the latter dominates Magdalenian art (~17–12 kya, see Supplementary Information section). Similarly, two distinct morphological forms of Late Pleistocene bison have been reported from North Sea sediments[17].

To further examine the potential existence of a previously unrecognized fossil bison species within Europe, we sequenced ancient mtDNA and nuclear DNA from bones and teeth of 64 Late Pleistocene/Holocene bison specimens.

We reveal that the wisent lineage originated from hybridization between the aurochs and steppe bison, and this new form alternated ecologically with steppe bison throughout the Late Pleistocene and appears to have been recorded by early cave artists.

Results

New group of ancient European bison. The mtDNA sequences of 38 specimens, dated from >50 to 14 kya and ranging from the Caucasus, Urals, North Sea, France and Italy, formed a previously unrecognized genetic clade, hereafter referred to as CladeX, related to modern and historical wisent (including the Caucasian form; Fig. 2a,b). By using the radiocarbon-dated specimens to calibrate our phylogenetic estimate of the timescale, we inferred that the divergence between CladeX and modern wisent lineages occurred ~120 (92–152) kya, likely during the last (Eemian) interglacial. Both these mitochondrial clades are more closely related to cattle than to bison, suggesting that they are descended from an ancient hybridization event that took place >120 kya (presumably between steppe bison and an ancestral form of aurochs, from which the mitochondrial lineage was acquired).

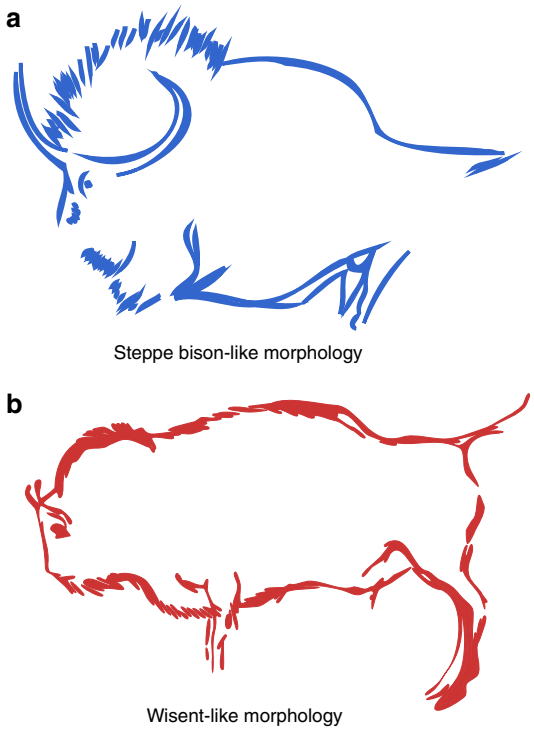


Figure 1 | Cave painting example of steppe bison-like and wisent-like morphs. (**a**) Reproduction from Lascaux cave (France), from the Solutrean or early Magdalenian period (~20,000 kya—picture adapted from ref. 53). (**b**) Reproduction from the Pergouset cave (France), from the Magdalenian period (<17,000 kya—picture adapted from ref. 54).

Hybrid origin of wisent and ancient European bison. To investigate the potential hybrid origins of wisent and CladeX, we used target enrichment and high-throughput methods to sequence ~10,000 genome-wide bovine single-nucleotide polymorphisms (SNPs) from nine members of CladeX, an ancient (>55 kyr) and a historical (1911 AD) wisent specimen and two steppe bison (30 and >50 kyr). Principal Component Analysis (PCA) and phylogenetic analysis (Fig. 3 and Supplementary Fig. 10) of the nuclear data demonstrate that members of CladeX are closely related to the steppe bison. D-statistic[18] analyses confirm a closer affinity of both CladeX and the ancient wisent to steppe bison than to modern wisent (Fig. 3b), which is explicable because of rapid genetic drift during the severe bottleneck leading to modern wisent. Concordantly, our historical wisent sample (Caucasian, from 1911) displays a signal intermediate between modern wisent and both CladeX and steppe bison (Fig. 3b(3–5),c).

The nuclear and mitochondrial analyses together suggest that the common ancestor of the wisent and CladeX mitochondrial lineages originated from asymmetrical hybridization (or sustained introgression) between male steppe bison and female aurochs (see Supplementary Fig. 20). This scenario is consistent with the heavily polygynous mating system of most large bovids[19], and the observation that hybridization between either extant bison species and cattle usually results in F1 male infertility, consistent with Haldane's Rule of heterogametic crosses[20–22]. However, it is unclear whether hybridization took place only once or multiple times, and how and at what point after the initial hybridization event(s) the wisent–CladeX forms became distinct from the steppe bison.

To examine the extent of genetic isolation maintained through time by the hybrid forms (wisent and CladeX) from steppe bison, we characterized the genomic signals originating from either steppe bison or aurochs in the wisent and CladeX lineages.

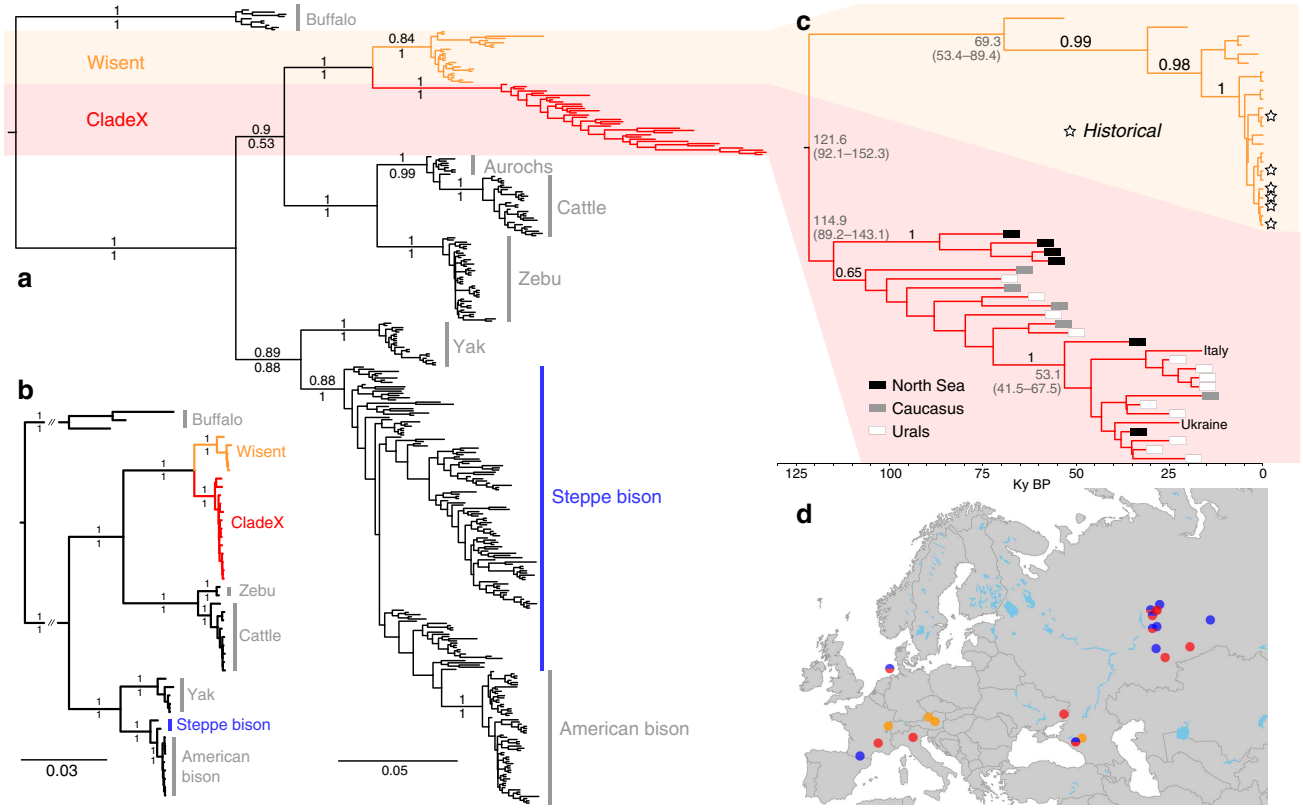


Figure 2 | Identification of CladeX. (**a**) Phylogenetic tree inferred from bovine mitochondrial control region sequences, showing the new clade of bison individuals. The positions of the newly sequenced individuals are marked in red for CladeX. (**b**) Bovine phylogeny estimated from whole-mitochondrial genome sequences, showing strong support for the grouping of wisent and CladeX with cattle (cow) and zebu. For both trees (**a**,**b**) numbers above branches represent the posterior probabilities from Bayesian inference, numbers below branches represent approximate likelihood ratio test support values from maximum-likelihood analysis and scale bars represent nucleotide substitutions per site from the Bayesian analysis. (**c**) Maximum-clade-credibility tree of CladeX and wisent estimated using Bayesian analysis and calibrated with radiocarbon dates associated with the sequenced bones. Dates of samples older than 50 kyr were estimated in the phylogenetic reconstruction. (**d**) Map showing all sampling locations, using the same colour code (red for CladeX, orange for wisent and blue for steppe bison).

Calculations of f_4 ratios[23] show the same high proportion of nuclear signal from steppe bison ($\geq$89.1%) and low proportion from aurochs ($\leq$10.9%) in both wisent and CladeX (Fig. 3d and Supplementary Table 6). Independent calculation of hybridization levels from ABC comparisons with simulated data also shows clear evidence of hybridization, with similar proportions of nuclear signal (97.2% probability that there is at least 1% aurochs ancestry and a 87.6% probability that there is at least 5% aurochs ancestry; see Supplementary Note 2 and Supplementary Tables 10 and 11). The agreement between these two methods is compelling evidence of hybridization. In addition, a greater number of derived alleles are common to both wisent and CladeX lineages (either from the imprint of steppe bison ancestry, aurochs ancestry, or from post-hybridization drift) than expected from multiple hybridization events (see Supplementary Note 2 and Supplementary Tables 8 and 9), implying that CladeX represents part of the Late Pleistocene wisent diversity. The age of the oldest genotyped specimens of CladeX (23 kyr) and wisent (>55 kyr) confirm that the initial hybridization event (or ultimate significant introgression of steppe bison) occurred before 55 kya. Together, the long-term stability of the nuclear and mitochondrial signal in wisent and CladeX indicates that the hybrid bison lineage maintained a marked degree of genetic isolation throughout the Late Pleistocene, consistent with the different morphologies observed in the North Sea specimens[17].

Hybrid and steppe Bison represent different ecological forms. The temporal distribution of genotyped individuals reveals that wisent mitochondrial lineages (including CladeX) are only observed before 50 kya and after 34 kya, when steppe bison appears to be largely absent from the European landscape (Fig. 4). The detailed records of the southern Ural sites allow the timing of the population replacements between steppe bison and wisent to be correlated with major palaeoenvironmental shifts, revealing that the wisent was associated with colder, more tundra-like landscapes and absence of a warm summer (Supplementary Fig. 22). Stable isotope data ($\partial^{13}C/\partial^{15}N$; Supplementary Fig. 23) and environment reconstructions show that wisent were present in a more diverse environment than steppe bison, with a more variable diet, suggesting that these two taxa occupied separate ecological niches.

Discussions

Contrary to previous palaeontological interpretations, the ancestors of modern wisent were present in Europe throughout the Late Pleistocene, and the two different bison morphs depicted in Paleolithic art suggest that early artists recorded the replacement of the steppe bison by the hybrid form (including CladeX) in Western Europe around the LGM. Two bison individuals have been genotyped from European caves during this period: a 19-kyr-old steppe bison from Southern France[24] and a

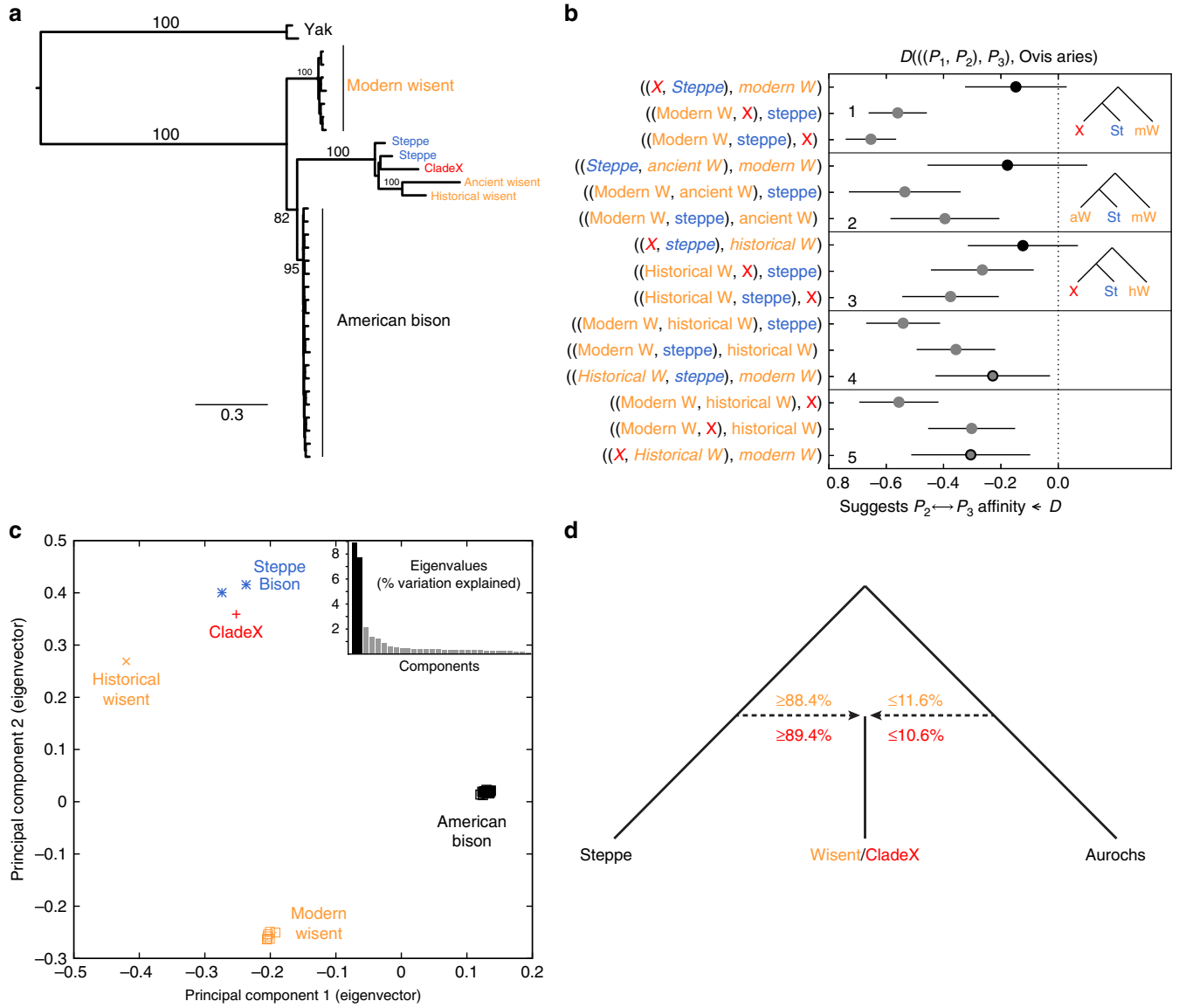


Figure 3 | Genome-wide data comparison of bison. (**a**) Maximum-likelihood phylogeny of modern and ancient bison from ~10,000 genome-wide nuclear sites, showing the close relationship between CladeX and steppe bison. However, a bifurcating phylogeny is not capable of displaying the complex relationships between these taxa (see Supplementary Fig. 8). Numbers above branches represent bootstrap values. (**b**) D-statistics from the same ~10,000 nuclear sites, using sheep as outgroup. For three bison populations, assuming two bifurcations and no hybridizations, three possible phylogenetic topologies can be evaluated using D-statistics, with the value closest to 0, indicating which topology is the most parsimonious. The topology being tested is shown on the vertical axis. Error bars are three s.e.'s (from block jackknife) either side of the data point. Data points that are significantly different from zero are shown in grey. The data point representing the topology in **a**, among a set of three possible topologies, is shown with a black outline. (**c**) Principal Component Analysis of ~10,000 genome-wide nuclear sites (ancient wisent not included due to the sensitivity of PCA to missing data, see Supplementary Fig. 10). (**d**) Proportion of steppe bison and aurochs ancestry in both wisent and CladeX lineages, calculated with f_4 ratios.

16-kyr-old wisent (CladeX) from Northern Italy (present study), corresponding to the timing of the morphological transition from steppe bison-like to wisent-like morphotypes apparent in cave art.

Combined evidence from genomic data, paleoenvironmental reconstructions and cave paintings strongly suggest that the hybridization of steppe bison with an ancient aurochs lineage during the late Pleistocene led to a morphologically and ecologically distinct form, which maintained its integrity and survived environmental changes on the European landscape until modern times. Although further analyses of deeper ancient genome sequencing will be necessary to characterize the phenotypic consequences of such hybridization, this adds to recent evidence of the importance of hybridization as a mechanism for speciation and adaptation of mammals[25–29] as is already accepted for plants. Lastly, the paraphyly of *Bos* with respect to *Bison*, and the evidence of meaningful hybridization between aurochs and bison, support the argument that both groups should be combined under the genus *Bos*[12,19,30].

Methods

Ancient DNA samples description and processing. Samples from a total of 87 putative bison bones were collected from three regions across Europe: Urals, Caucasus and Western Europe (Supplementary Data 1).

Dating of 45 samples that yielded DNA was performed at the Oxford Radiocarbon Accelerator Unit of the University of Oxford (OxA numbers), and the Ångström Laboratory of the University of Uppsala, Sweden, for the Swiss sample (Ua-42583). The calibration of radiocarbon dates was performed using OxCal v4.1 with the IntCal13 curve[31] (Supplementary Data 1).

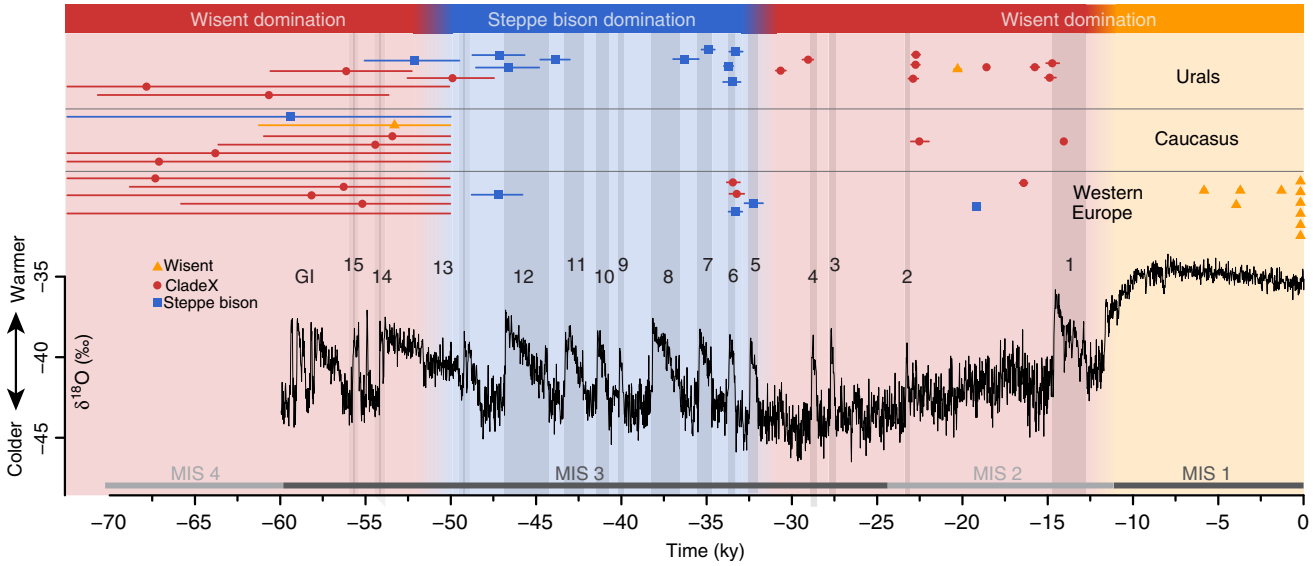


Figure 4 | Temporal and geographical distribution of bison in Europe. Individual calibrated AMS dates from the present study and published data are plotted on top of the NGRIP $\delta^{18}O$ record[55]. Age ranges for infinite AMS dates are from molecular clock estimates (Fig. 2c). Greenland interstadials (GIs) are numbered in black and marine isotope stages (MIS) in grey.

All ancient DNA work was conducted in clean-room facilities at the University of Adelaide's Australian Centre for Ancient DNA, Australia (ACAD), and at the University of Tuebingen, Germany (UT) following the published guidelines[32].

Samples were extracted using either phenol–chloroform[33] or silica-based methods[34,35] (see Supplementary Data 1).

Mitochondrial control region sequences (>400 bp) were successfully amplified from 65 out of 87 analysed samples in one or up to four overlapping fragments, depending on DNA preservation[33]. To provide deeper phylogenetic resolution and further examine the apparent close relationship between *Bos* and wisent mitochondria, whole-mitogenome sequences of 13 CladeX specimens, as well as one ancient wisent, one historical wisent and one steppe bison were generated using hybridization capture with either custom-made[36,37] (see Supplementary Note 1 for details).

In addition, genome-wide nuclear locus capture was attempted on DNA extracts from 13 bison samples (see Supplementary Table 2), using either an ~40,000 or an ~10,000 set of probes (as described in Supplementary Note 1). All targeted loci were part of the BovineSNP50 v2 BeadChip (Illumina) bovine SNP loci used in a previous phylogenetic study[38]. Ultimately, only the 9,908 loci common to both sets were used for comparative analysis.

Genetic data analysis. *Data processing*. Next-generation sequencing data were obtained from enriched libraries using paired-end reactions on Illumina HiSeq or MiSeq machines, and processed using the pipeline Paleomix v1.0.1 (ref. 39). AdapterRemoval v2 (ref. 40) was used to trim adapter sequences, merge the paired reads and eliminate all reads shorter than 25 bp. BWA v0.6.2 was then used to map the processed reads to either the reference mitochondrial genome of the wisent (NC_014044), American bison (NC_012346—only for the steppe bison A3133) or the *Bos taurus* genome reference UMD 3.1 (ref. 41). Minimum mapping quality was set at 25, seeding was disabled and the maximum number of gap opens was set to 2 (see Supplementary Tables 2 and 3).

MapDamage v2 (ref. 42) was used to check that the expected contextual mapping and damage patterns were observed for each library, depending on the enzymatic treatment used during library preparation (see Supplementary Table 3 and Supplementary Figs 1–3 for examples), and to rescale base qualities accordingly.

Phylogenetic analyses. The 60 newly sequenced bovine mitochondrial regions (Supplementary Data 1) were aligned with 302 published sequences (Supplementary Table 4), and a phylogenetic tree was inferred using both maximum-likelihood (PhyML v3 (ref. 43)) and Bayesian (MrBayes v3.2.3 (ref. 44)) methods (Fig. 2a and Supplementary Fig. 4). The same methods were used to obtain the whole-mitogenome phylogeny of 16 newly sequenced bison (Supplementary Data 1) aligned with 31 published sequences (Fig. 2b and Supplementary Fig. 5). To estimate the evolutionary timescale, we used the programme BEAST v1.8.1 (ref. 45) to conduct a Bayesian phylogenetic analysis of all radiocarbon-dated samples from CladeX and wisent (Fig. 1c), using the mean calibrated radiocarbon dates as calibration points. All parameters showed sufficient sampling after 5,000,000 steps, and a date-randomization test supported that the temporal signal from the radiocarbon dates associated with the ancient sequences was sufficient to calibrate the analysis[46] (Supplementary Fig. 6).

Finally, phylogenetic trees were inferred from nuclear loci data using RAxML v8.1.21 (ref. 47), first from published data of modern bovine representatives[38] (using sheep as an outgroup; Supplementary Fig. 7) and then including five ancient samples (two ancient steppe bison, an ancient wisent, a historical wisent and a CladeX bison; Fig. 2a), which had the highest number of nuclear loci successfully called among the ~10 k nuclear bovine SNPs targeted with hybridization capture (see Supplementary Fig. 8).

Principal Component Analysis. PCA (Fig. 3a and Supplementary Fig. 10) was performed using EIGENSOFT version 6.0.1 (ref. 48). In Fig. 3a, CladeX sample A006 was used as the representative of CladeX, as this sample contained the most complete set of nuclear loci called at the bovine SNP loci (see Supplementary Table 2). Other CladeX individuals, as well as ancient wisent, cluster towards coordinates 0.0, 0.0 (see Supplementary Fig. 10), because of missing data.

D and f statistics. Support for the bifurcating nuclear tree (Fig. 2a) was further tested using D-statistics calculated using ADMIXTOOLS version 3.0, git~3065acc5 (ref. 23). Sensitivity to factors like sampling bias, depth of coverage, choice of outgroup, heterozygosity (by haploidization) and missing data did not have notable influences on the outcome (Supplementary Figs 12–15).

The proportion of the wisent's ancestry differentially attributable to the steppe bison, and the aurochs was estimated with AdmixTools using an f_4 ratio[23] with sheep (*Ovis aries*) as the outgroup (Supplementary Figs S16, S17 and 3D). Again, the test was shown to be robust to haploidization.

Finally, to test whether the wisent lineages (including CladeX) have a common hybrid ancestry, or whether multiple independent hybridization events gave rise to distinct wisent lineages (Supplementary Fig. 18), we identify nuclear loci that have an ancestral state in the aurochs lineage, but a derived state in the steppe bison lineage (see Supplementary Note 2 section 'Identification of Derived Alleles'). Hypergeometric tests (Supplementary Tables 8 and 9) showed strong support for an ancestral hybridization event occurring before the divergence of the wisent lineages.

Testing admixture using ABC and simulated data. Admixture proportions were also independently tested using simulated data and an ABC approach. Nuclear genetic count data were simulated for two species trees (as described in Supplementary Fig. 19 and Supplementary Note 2 section) by drawing samples from two Multinomial distributions, where for tree topology X_1, $n^{X_1} \sim \mathrm{Mult}(N, p^{T,X_1})$, and for tree topology X_2, $n^{X_2} \sim \mathrm{Mult}(N, p^{T,X_2})$. The linear combination of these counts was then considered.

ABC was performed using the R package 'abc', with a ridge regression correction for comparison of the simulated and observed data using the 'abc' function[49]. The distance between the observed and simulated data sets is calculated as the Euclidean distance in a three-dimensional space, corrected for the within dimension variability. A tolerance $\epsilon = 0.005$ was chosen so that the closest $\ell \times \epsilon$ simulated data sets are retained. For each analysis we had $\ell = 100,000$, resulting in 500 posterior samples.

We performed leave-one-out cross-validation using the function 'cv4abc' on $\ell' = 250$ randomly selected simulations, and report the prediction error, calculated as

$$E_{\mathrm{pred}} = \frac{\sum_{i=1}^{\ell'} (\hat{\gamma}_i - \gamma_i)^2}{\mathrm{Var}(\gamma_i)}$$

for each analysis. At most, the prediction error was 0.5111 s.d.'s away from zero, and so we observe that the analysis has performed well (see Supplementary Table 10).

Palaeoenvironment reconstruction and stable isotope analyses. The Urals material has the most complete sampling through time (Fig. 4 and Supplementary Fig. 22), allowing us to contrast reconstructed paleoenvironmental proxies for the region (see Supplementary Note 3). Paleovegetation types were inferred for a convex hull of the Ural study region based on geo-referenced site locations for all genotyped ancient samples (Supplementary Fig. 21). Global maps of BIOME4 plant functional types[50] were accessed for 2,000-year time steps throughout the period from 70,000 years ago to the present day, with a $1° \times 1°$ latitude/longitude grid cell resolution. We also generated estimates of the annual mean daily temperature and Köppen–Geiger climate classification[51] using the Hadley Centre Climate model (HadCM3)[52]. Finally, stable isotope values (δ13C and δ15N) obtained for all the genotyped bison individuals from the Ural region were compared between steppe bison and wisent (Supplementary Fig. 23).

Cave paintings. Two consistent morphological types can be distinguished within the diversity of bison representations (see Fig. 1 and Supplementary Figs 24–27). The first type, abundant before the LGM, is characterized by long horns (with one curve), a very oblique dorsal line and a very robust front part of the body (solid shoulders versus hindquarters), all traits similar to the modern American bison. The second type, dominating the more recent paintings between 18 and 15 kya, displays thinner sinuous horns (often with a double curve), a smaller hump and more balanced dimensions between the front and rear of the body, similar to modern wisent and to some extent aurochsen (see also Supplementary Note 4). The coincident morphological and genetic replacement indicate that variation in bison representations in Paleolithic art does not simply represent stylistic evolution, but actually reflects the different forms of bison genotyped in this study (that is, pre and post-hybridization) through time.

Data Availability. All newly sequenced mitochondrial control regions are deposited at the European Nucleotide Archive under the following accession numbers (LT599586–645) and all complete mitochondrial genomes at GenBank (KX592174–89). The BEAST input file (XML) is available as Supplementary Data set 2, the MrBayes input file (Nexus), including all whole-mitochondrial genomes, as Supplementary Data set 3 and the nuclear SNPs as Supplementary Data set 4 (VCF format). All other data are included in the Supplementary Material or available upon request to the corresponding authors.

References

1. Kurtén, B. *Pleistocene Mammals of Europe* (1968).
2. Geist, V. The relation of social evolution and dispersal in ungulates during the Pleistocene, with emphasis on the old world deer and the genus Bison. *Quat. Res.* **1,** 285–315 (1971).
3. Benecke, N. The holocene distribution of European bison: the archaeozoological record. *Munibe Antropol. Arkeol.* **57,** 421–428 (2005).
4. Bocherens, H., Hofman-Kamińska, E., Drucker, D. G., Schmölcke, U. & Kowalczyk, R. European Bison as a refugee species? Evidence from isotopic data on early holocene bison and other large herbivores in Northern Europe. *PLoS ONE* **10,** e0115090 (2015).
5. Stuart, A. J. Mammalian extinctions in the late Pleistocene of Northern Eurasia and North America. *Biol. Rev.* **66,** 453–562 (1991).
6. Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature* **479,** 359–364 (2011).
7. Cooper, A. *et al.* Abrupt warming events drove Late Pleistocene Holarctic megafaunal turnover. *Science* **349,** 602–606 (2015).
8. Slatis, H. M. An analysis of inbreeding in the European Bison. *Genetics* **45,** 275–287 (1960).
9. Tokarska, M., Pertoldi, C., Kowalczyk, R. & Perzanowski, K. Genetic status of the European bison Bison bonasus after extinction in the wild and subsequent recovery. *Mammal Rev.* **41,** 151–162 (2011).
10. Verkaar, E. L. C., Nijman, I. J., Beeke, M., Hanekamp, E. & Lenstra, J. A. Maternal and paternal lineages in cross-breeding bovine species. Has wisent a hybrid origin? *Mol. Biol. Evol.* **21,** 1165–1170 (2004).
11. Hassanin, A. *et al.* Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.* **335,** 32–50 (2012).
12. Bibi, F. A multi-calibrated mitochondrial phylogeny of extant Bovidae (Artiodactyla, Ruminantia) and the importance of the fossil record to systematics. *BMC Evol. Biol.* **13,** 166 (2013).
13. Sauvet, G. & Wlodarczyk, L'art Pariétal, miroir des sociétés paléolithiques. *Zephyrus Rev. Prehist. Arqueol.* **53,** 217–240 (2000).
14. Breuil, H. *Quatre Cents Siècles d'art Pariétal; Les Cavernes Ornées de l'âge du Renne* (Centre d'études et de documentation préhistoriques, 1952).
15. Leroi-Gourhan, A. *Préhistoire de l'art Occidental* (1965).
16. Petrognani, S. *De Chauvet à Lascaux: l'art des Cavernes, Reflet de sociétés Préhistoriques en Mutation* (Editions Errance, 2013).
17. Drees, M. & Post, K. Bison bonasus from the North Sea, the Netherlands. *Cranium* **24,** 48–52 (2007).
18. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28,** 2239–2252 (2011).
19. Groves, C. Current taxonomy and diversity of crown ruminants above the species level. *Zitteliana B* **32,** 5–14 (2014).
20. Haldane, J. B. S. Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12,** 101–109 (1922).
21. Hedrick, P. W. Conservation genetics and North American bison (*Bison bison*). *J. Hered.* **100,** 411–420 (2009).
22. Derr, J. N. *et al.* Phenotypic effects of cattle mitochondrial DNA in American bison. *Conserv. Biol.* **26,** 1130–1136 (2012).
23. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
24. Marsolier-Kergoat, M.-C. *et al.* Hunting the extinct steppe bison (*Bison priscus*) mitochondrial genome in the Trois-Frères Paleolithic Painted Cave. *PLoS ONE* **10,** e0128267 (2015).
25. Ropiquet, A. & Hassanin, A. Hybrid origin of the Pliocene ancestor of wild goats. *Mol. Phylogenet. Evol.* **41,** 395–404 (2006).
26. Larsen, P. A., Marchán-Rivadeneira, M. R. & Baker, R. J. Natural hybridization generates mammalian lineage with species characteristics. *Proc. Natl Acad. Sci. USA.* **107,** 11447–11452 (2010).
27. Song, Y. *et al.* Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. *Curr. Biol.* **21,** 1296–1301 (2011).
28. Amaral, A. R., Lovewell, G., Coelho, M. M., Amato, G. & Rosenbaum, H. C. Hybrid speciation in a marine mammal: the clymene dolphin (*Stenella clymene*). *PLoS ONE* **9,** e83645 (2014).
29. Lister, A. M. & Sher, A. V. Evolution and dispersal of mammoths across the Northern Hemisphere. *Science* **350,** 805–809 (2015).
30. Groves, C. & Grubb, P. *Ungulate Taxonomy* (Johns Hopkins University Press, 2011).
31. Reimer, P. J. *et al.* IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55,** 1869–1887 (2013).
32. Willerslev, E. & Cooper, A. Ancient DNA. *Proc. R Soc. B Biol. Sci.* **272,** 3–16 (2005).
33. Shapiro, B. *et al.* Rise and fall of the Beringian steppe bison. *Science* **306,** 1561–1565 (2004).
34. Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. *Nat. Commun.* **4,** 1764 (2013).
35. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2,** 1756–1762 (2007).
36. Llamas, B. *et al.* Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2,** e1501385 (2016).
37. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products. *PLoS ONE* **5,** e14004 (2010).
38. Decker, J. E. *et al.* Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. *Proc. Natl Acad. Sci. USA* **106,** 18644–18649 (2009).
39. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9,** 1056–1082 (2014).
40. Lindgreen, S. AdapterRemoval easy cleaning of next generation sequencing reads. *BMC Res. Notes* **5,** 337 (2012).
41. Zimin, A. V. *et al.* A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* **10,** R42 (2009).
42. 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).
43. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59,** 307–321 (2010).
44. Ronquist, F. *et al.* MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61,** 539–542 (2012).
45. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).
46. Ho, S. Y. W. *et al.* Bayesian estimation of substitution rates from ancient DNA sequences with low information content. *Syst. Biol.* **60,** 366–375 (2011).
47. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
48. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
49. Csilléry, K., François, O. & Blum, M. G.B. abc: an R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3,** 475–479 (2012).
50. Kaplan, J. O. *Geophysical Applications of Vegetation Modeling* (Lund University, 2001).
51. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **11,** 1633–1644 (2007).

52. Singarayer, J. S. & Valdes, P. J. High-latitude climate sensitivity to ice-sheet forcing over the last 120 kyr. *Quat. Sci. Rev.* **29,** 43–55 (2010).
53. Leroi-Gourhan, A. & Allain, J. *Lascaux Inconnu* (CNRS, 1979).
54. Lorblanchet, M. *La Grotte Ornée de Pergouset (Saint-Géry, Lot). Un Sanctuaire Secret Paléolithique* (Maison des Sciences de l'Homme, 2001).
55. Wolff, E. W., Chappellaz, J., Blunier, T., Rasmussen, S. O. & Svensson, A. Millennial-scale variability during the last glacial: the ice core record. *Quat. Sci. Rev.* **29,** 2828–2838 (2010).

Acknowledgements

We are grateful to A.A. Krotova (Institute of Archeology Ukrainian Academy of Sciences), K. Wysocka (Vinnytsia Regional Local History Museum), M. Blant (Swiss Institute for Speleology and Karst Studies), G. Zazula and E. Hall (Yukon Palaeontology Program), C. Lefèvre (Muséum National d'Histoire Naturelle), M. Leonardi (Natural History Museum of Denmark), J.P. Brugal (Laboratoire méditerranéen de préhistoire Europe Afrique and Musée d'Orgnac), the Natural History Museum of Vienna and the Paleontological Institute of Moscow for providing access to samples.We thank A. Lister, K. Helgen and J. Tuke for their comments on the study, as well as A. Vorobiev, Y. Clément and M.E.H. Jones for their help in the project. This research was supported by the Australian Research Council, the European Commission (PIRSES-GA-2009-247652—BIOGEAST), the Polish National Science Centre (N N304 301940 and 2013/11/B/NZ8/00914), the Danish National Research Foundation (DNRF94), the Marie Curie International Outgoing Fellowship (7th European Community Framework Program—MEDITADNA, PIOF-GA-2011-300854, FP7-PEOPLE) and the Russian Foundation for Basic Research (N 15-04-03882).

Author contributions

J.S., G.G., K.C., S.M.R., B.L., K.J.M., S.Y.W.H., M.S.Y.L., B.S., A.R. and A.C. designed experiments; P.K., G.B., R.B., J.B., E.C.-B., V.B.D., F.F., J.G., L.V.G., A.G., W.H., M.-A.J., E.H.-K., O.K., F.L., G.L., A.S., M.T., J.v.d.P., J.-D.V., L.O. and R.K. provided samples, interpretations of results and comments on the study; K.C., S.M.R., B.L., P.B., W.H., J.K., A.I., A.v.L. and B.S. performed laboratory genetic analyses; D.C., K.D., T.H. and J.v.d.P. performed radiocarbon-dating analyses; J.S., G.G., S.Y.W.H., M.S.Y.L., J.E.D., R.D.S., A.R. and O.W. performed bioinformatic analyses; P.K. and D.A.F. performed palaeoenvironmental analyses; C.F. and G.T. provided data and interpretation of cave art; J.S., G.G., B.L., K.J.M., M.S.Y.L., J.E.D., C.G., W.H., J.F.T., L.O., R.K. and A.C. analysed the results; and A.C. and J.S. wrote the paper with help from all co-authors.

Additional information

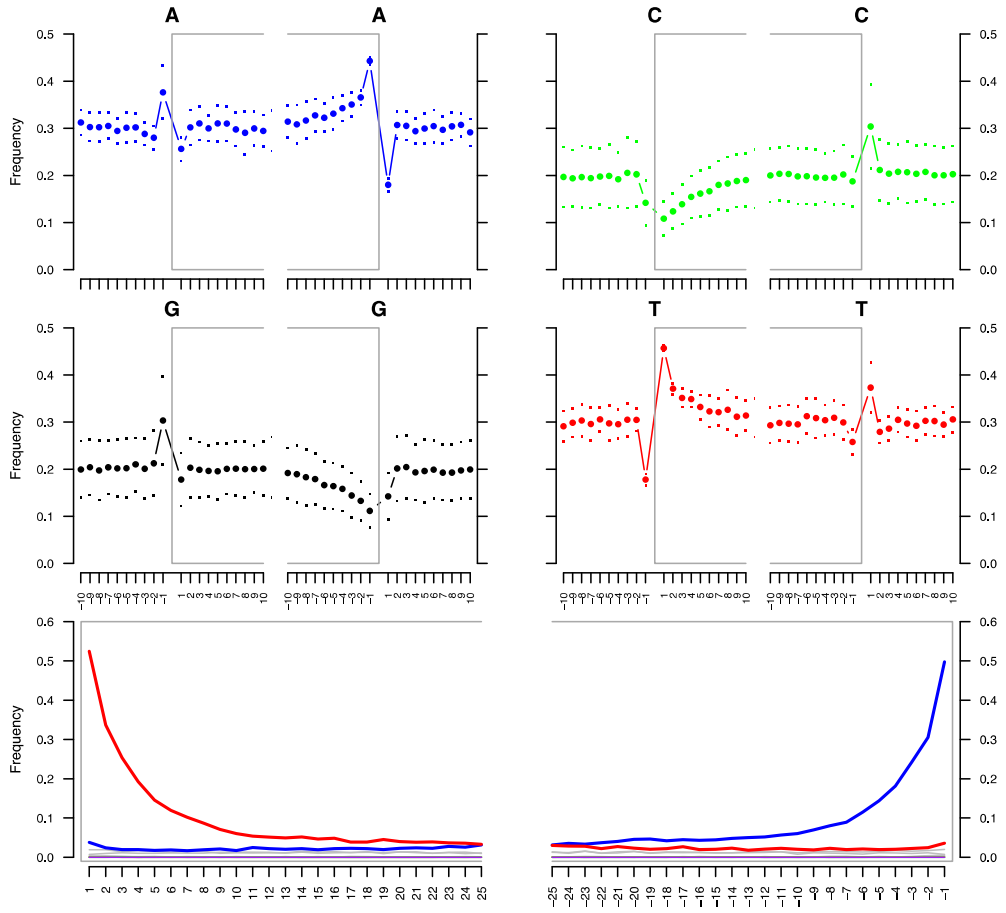
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

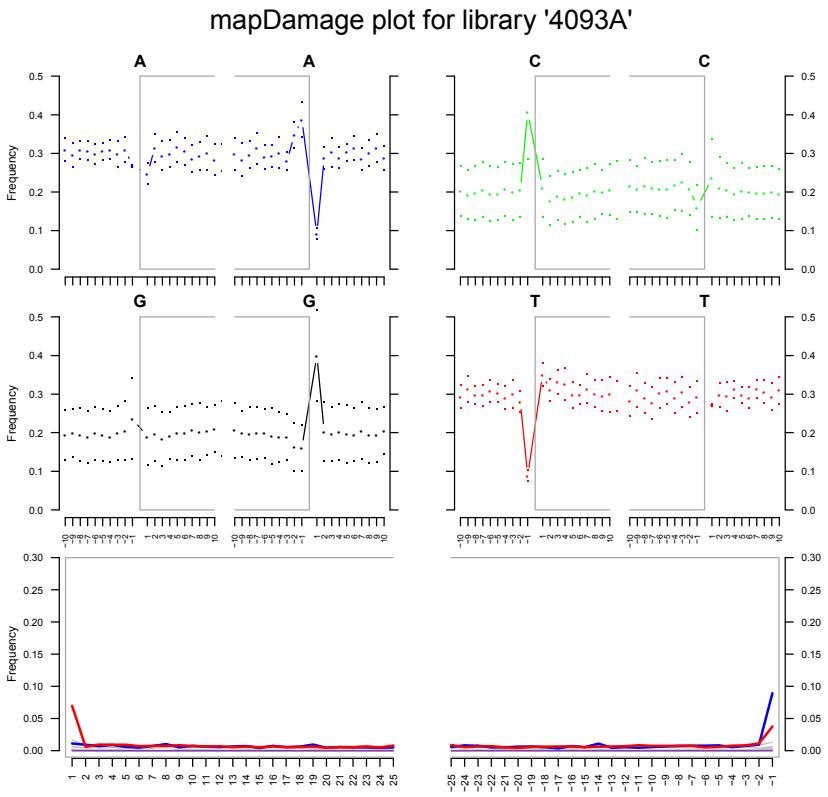
Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Soubrier, J. *et al.* Early cave art and ancient DNA record the origin of European bison. *Nat. Commun.* **7**, 13158 doi: 10.1038/ncomms13158 (2016).

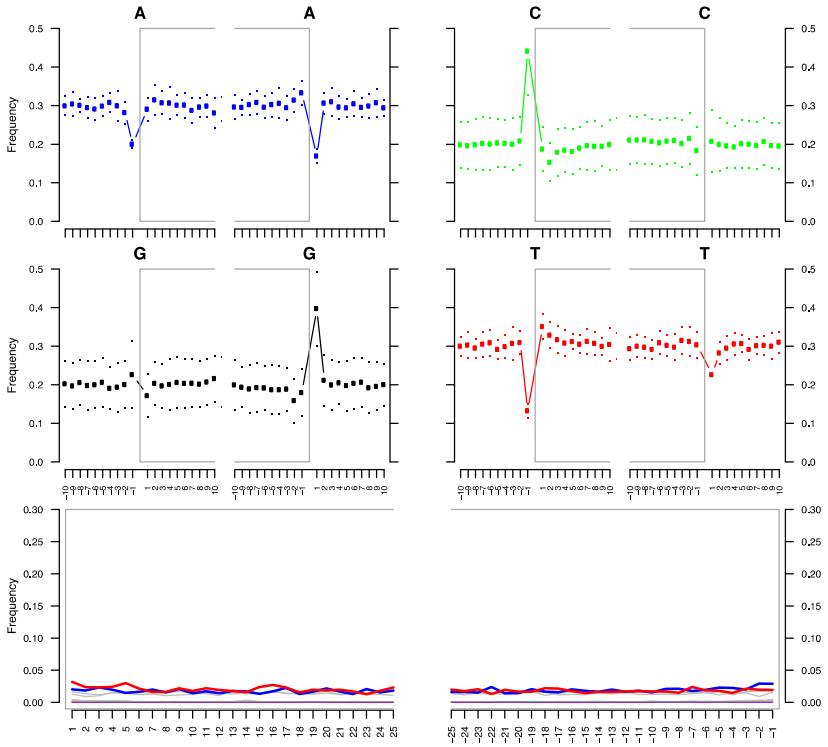
Supplementary Figures



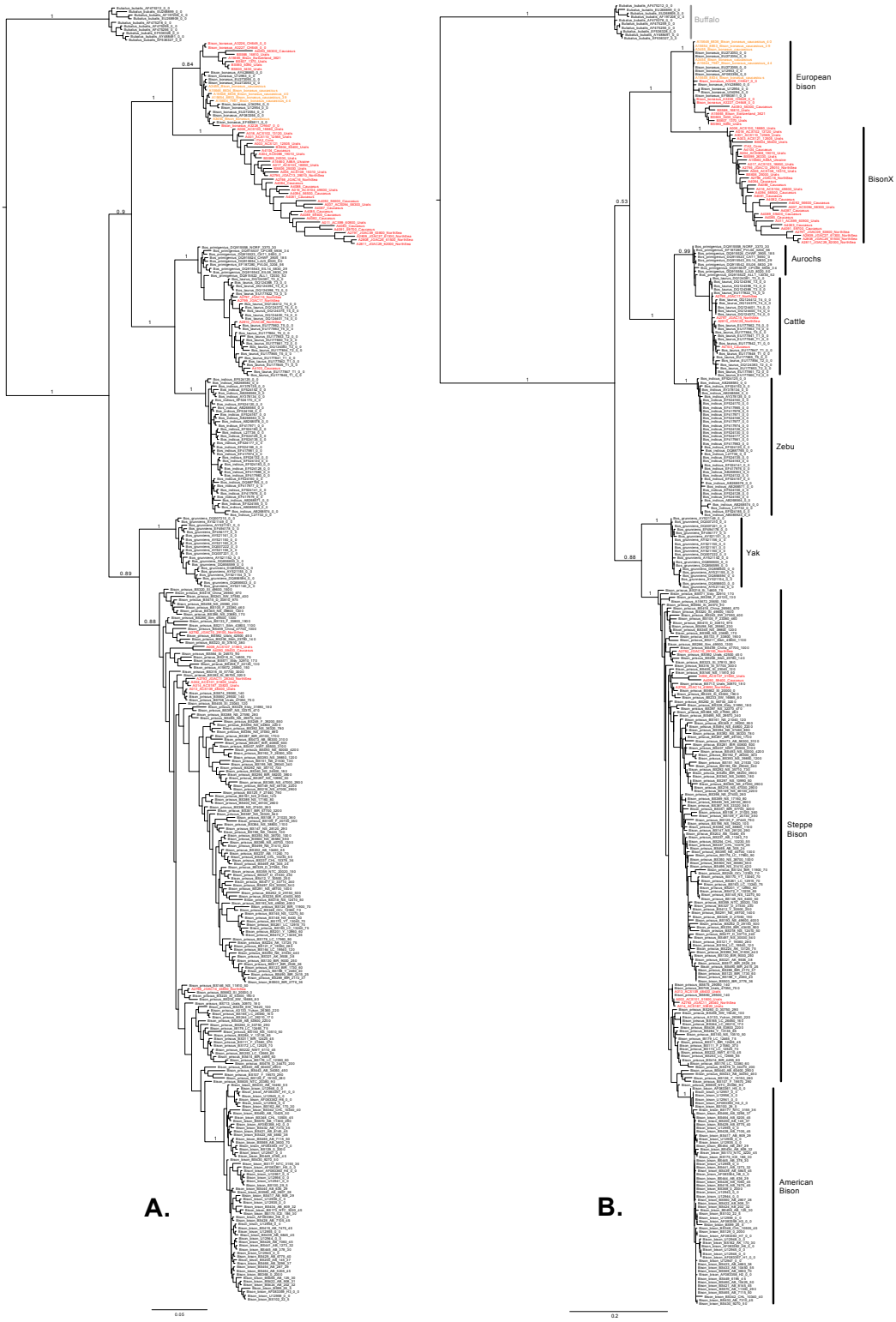
Supplementary Fig 1. Example of damage profile (sample LE257) obtained after sequencing of the whole mitochondrial genome using no treatment for the library preparation. As expected, there is an excess of purines found at the genomic position preceding the mapped reads, and an excess of C>T transitions at the first few positions of the reads.



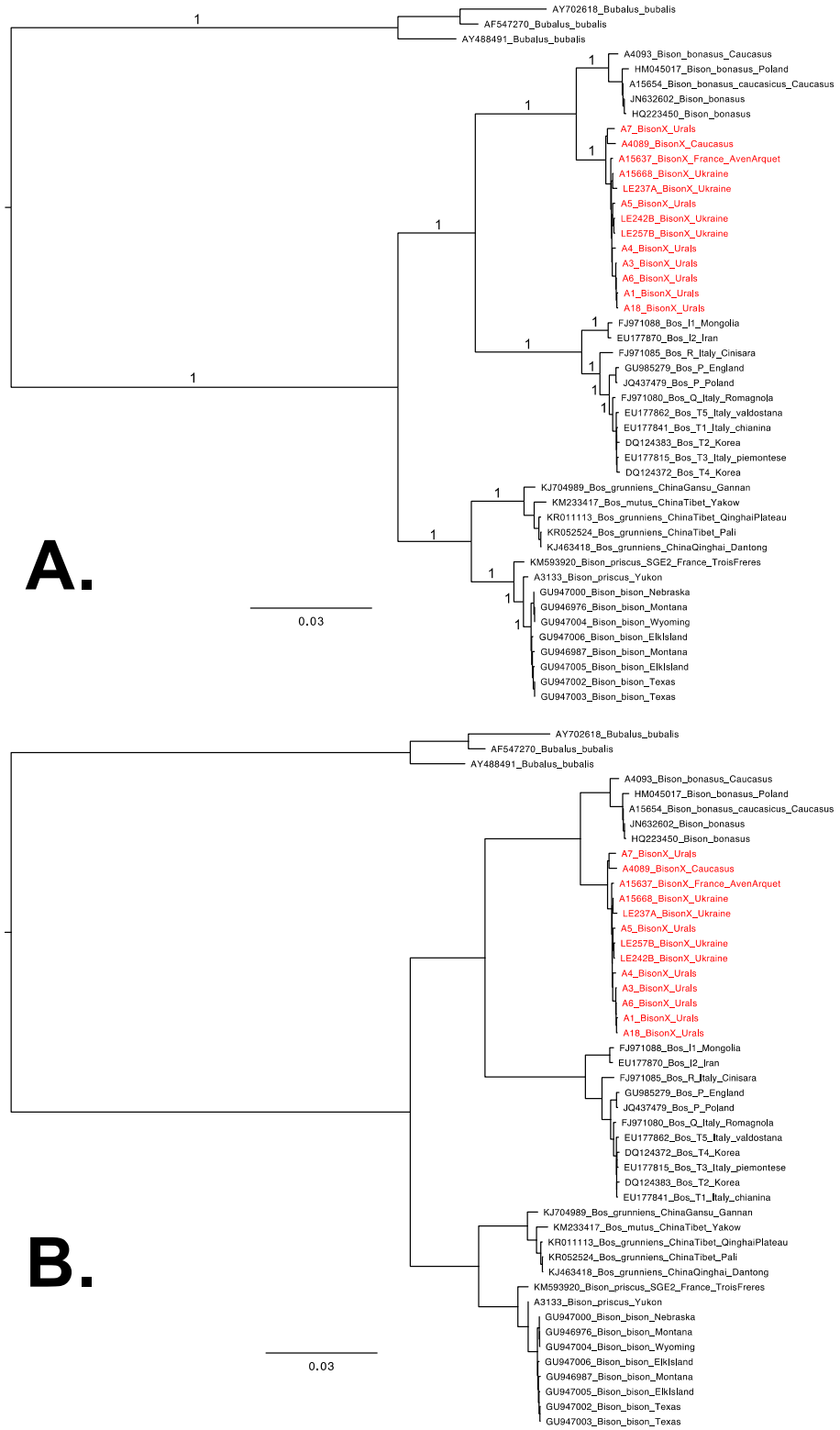
Supplementary Fig 2. Example of damage profile (sample A4093) obtained after sequencing of the whole mitochondrial genome using UDG-half treatment for the library preparation. As expected, there is an excess of cytosine found at the genomic position preceding the mapped reads, and an excess of C>T (and complementary G>A) transitions at the first (last) position of the reads.



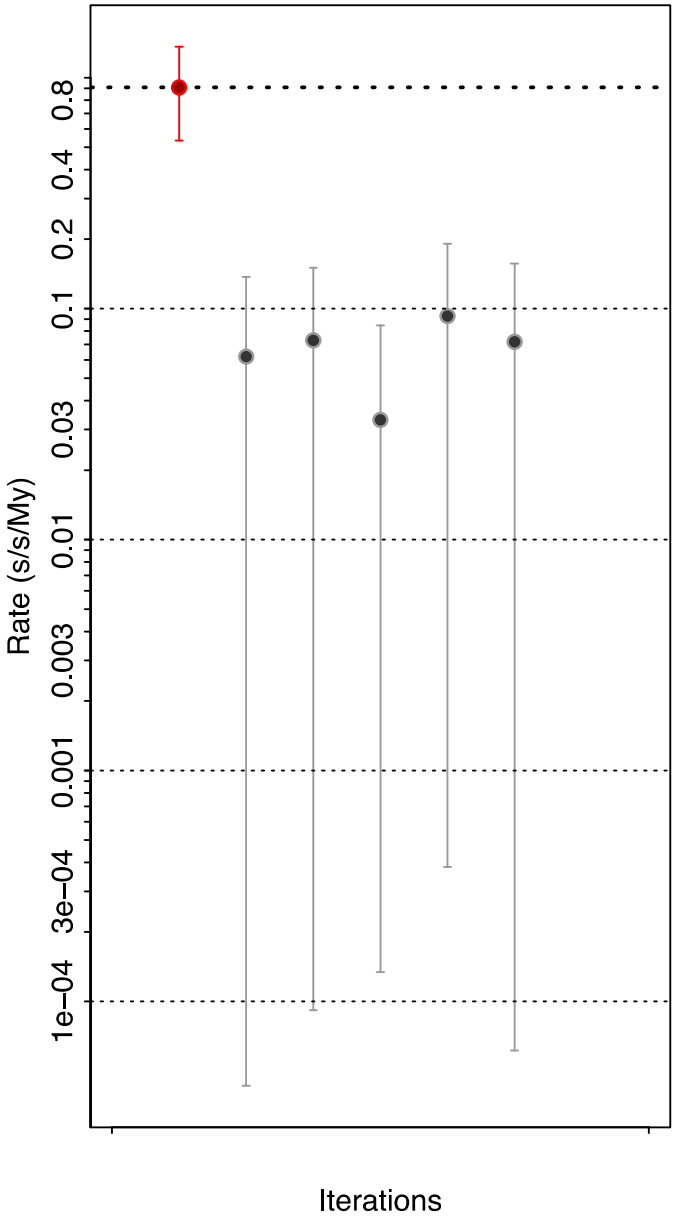
Supplementary Fig 3. Example of damage profile (sample A18) obtained after sequencing of the whole mitochondrial genome using full USER treatment for the library preparation. As expected, there is an excess of cytosine found at the genomic position preceding the mapped reads, and no excess of C>T transitions at the start of the reads.



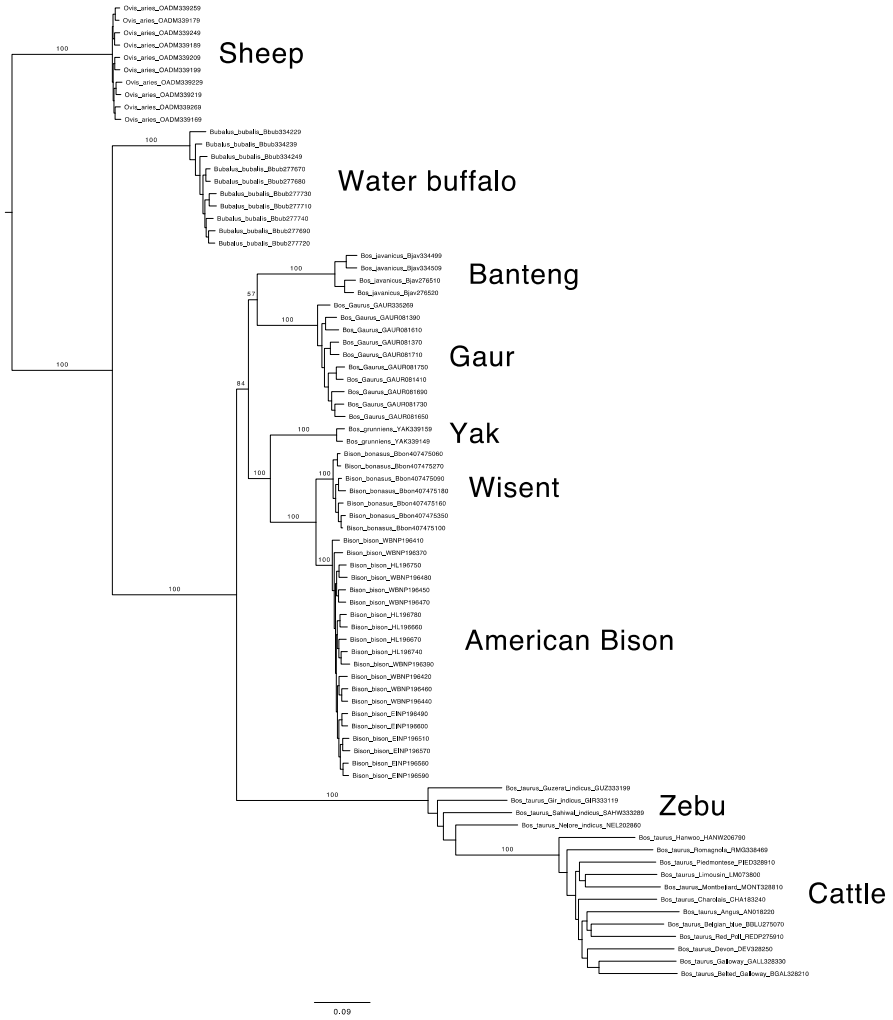
Supplementary Fig 4. Phylogenetic trees of mitochondrial control region sequences from 362 bovid samples. **A.** Majority-rule consensus tree from MrBayes. **B.** Maximum-likelihood tree from PhyML. The 60 newly sequenced individuals are in red font, with the Caucasian bison (*B. bonasus caucasicus*) in orange. Scale bars are given in substitutions per site.



Supplementary Fig 5. Phylogenetic trees inferred from whole mitochondrial genomes. **A.** Majority-rule consensus tree from MrBayes. **B.** Maximum-likelihood tree from PhyML. CladeX bison individuals are colored in red. Scale bars are given in substitutions per site.

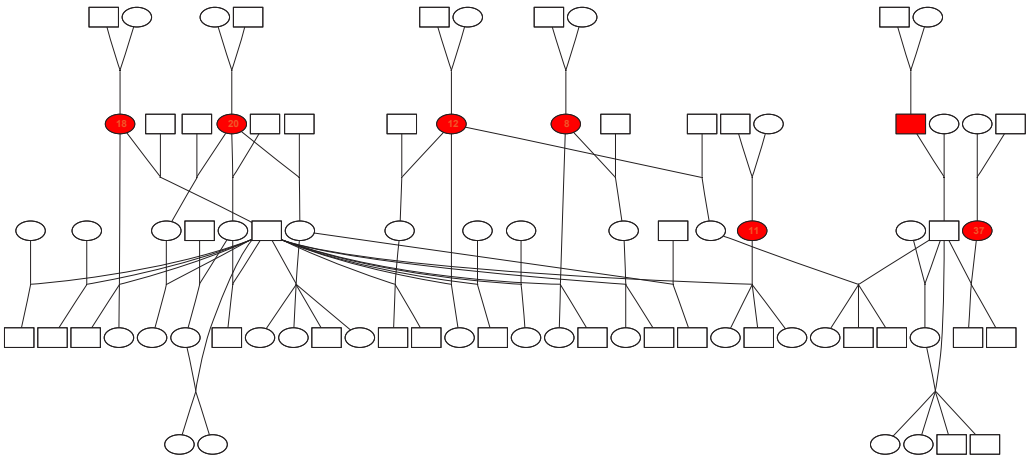


Supplementary Fig 6. Date-randomization test. The red circle and dotted line represent the mean estimate of the molecular rate obtained in the phylogenetic analysis of wisent and CladeX, calibrated using the radiocarbon dates associated with the ancient sequences. The grey lines represent the 95% HPD intervals of rates estimated with randomized dates. None of these margins overlap with the mean rate estimate from the original data set, demonstrating that the radiocarbon dates used for this study contain sufficient temporal information for calibrating the molecular clock.

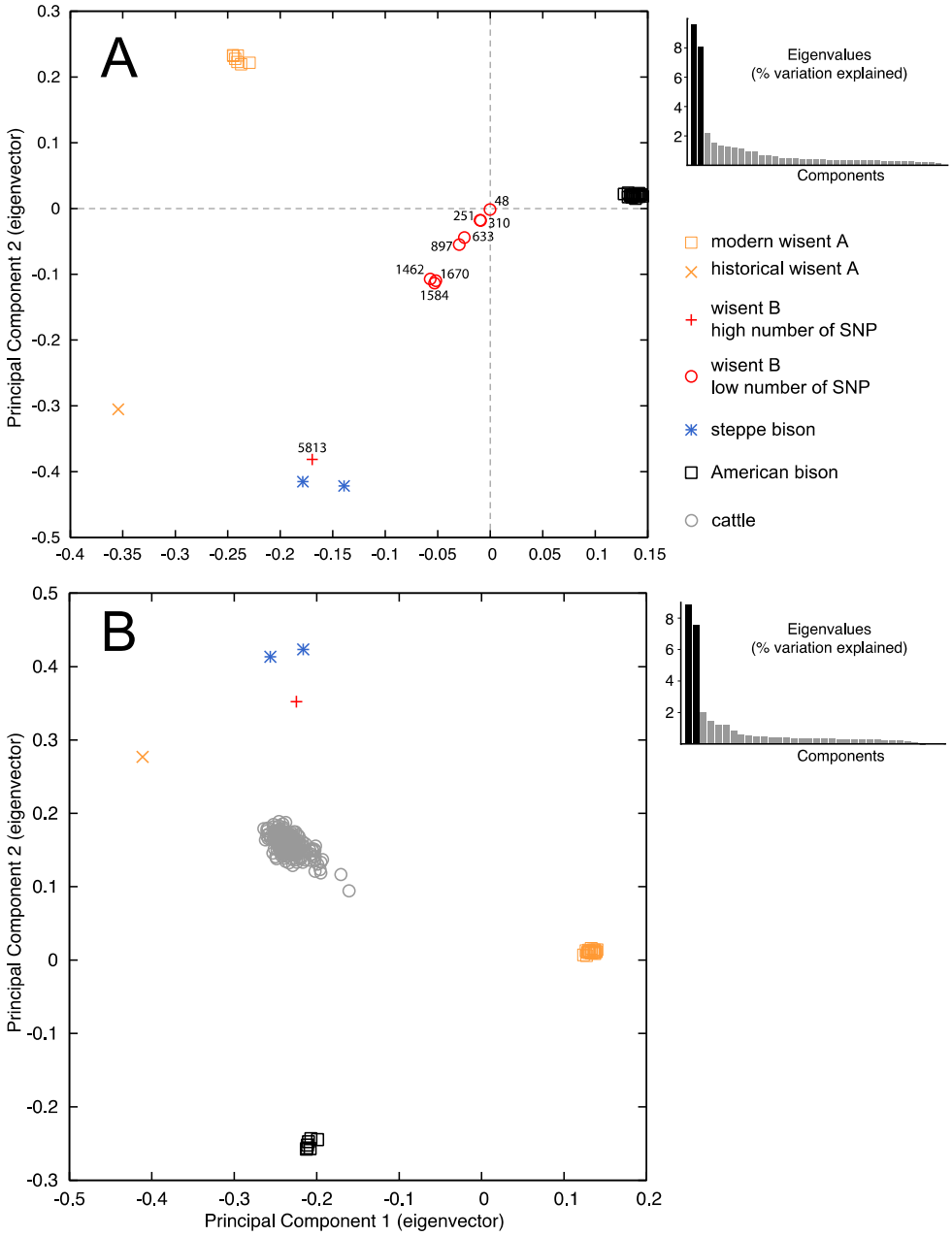


Supplementary Fig 7. Maximum-likelihood phylogeny of modern bovid species (and sheep as outgroup) from ~40k nuclear loci.

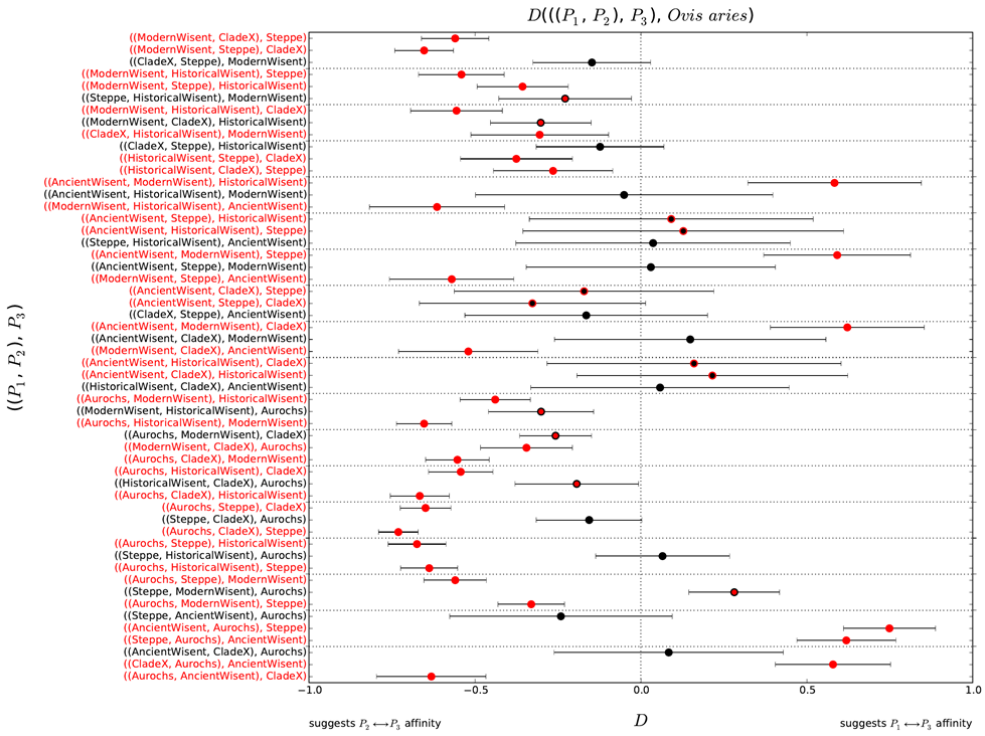
Supplementary Fig 8. Maximum-likelihood phylogenies of modern and ancient bison (and yak as outgroup), from ~10k nuclear loci. **A.** Phylogeny including the two ancient steppe bison. **B.** Phylogeny including the three pre-modern wisent. **C.** Phylogeny including the two steppe bison and three pre-modern wisent (ancient, historical and CladX). **D.** Replicate of **C.** but only using transversions for the non-modern samples.



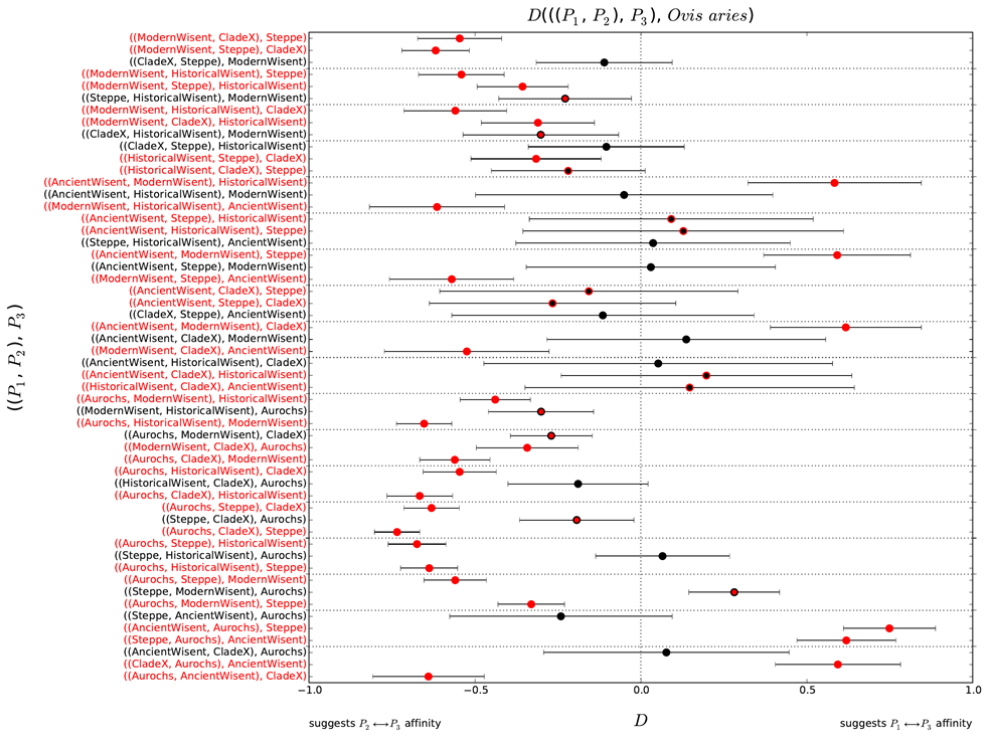
Supplementary Fig 9. Pedigree of wisent from the Białowieża Forest (Poland), from which seven genotyped individuals (in red) were included in the present study.



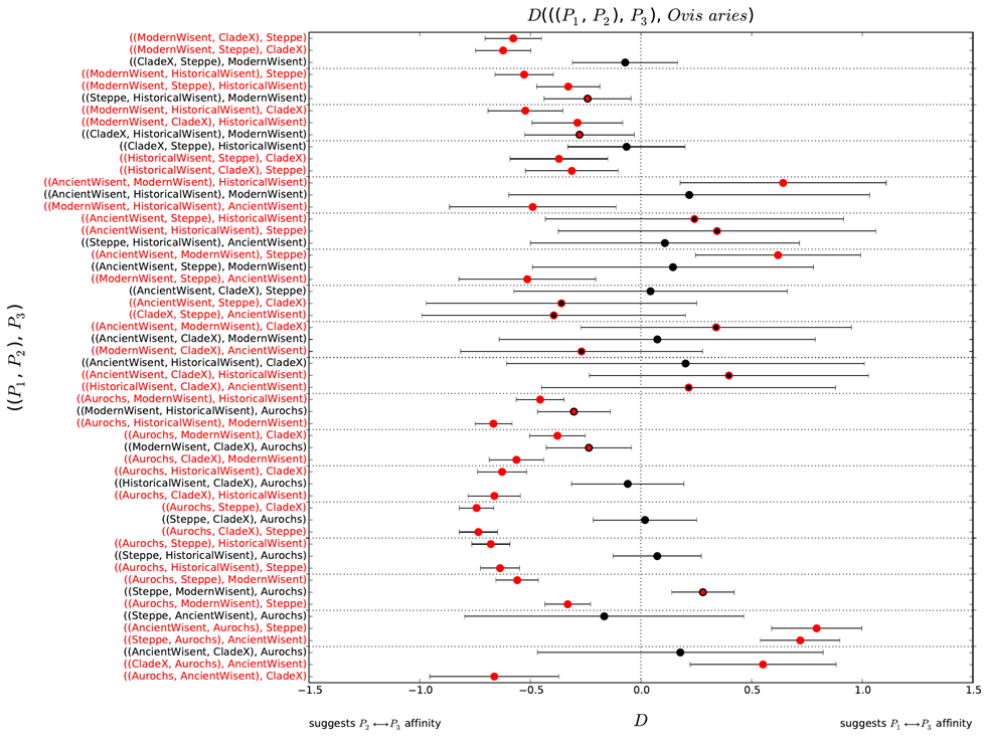
Supplementary Fig 10: A) Principal Component Analysis for nine CladeX individuals (including sample A006), one historical wisent, one ancient wisent, two steppe bison, seven modern wisent and 20 American bison. The numbers on the plot report the number of loci called for the individuals clustering towards zero coordinates (from Supplementary Table 2). Eigenvector 1 explains 9.58% of the variation, while eigenvector 2 explains 7.96% of the variation. B) Same Principal Component Analysis as Figure 3C with cattle individuals from Decker et al. (2009) projected onto original components.



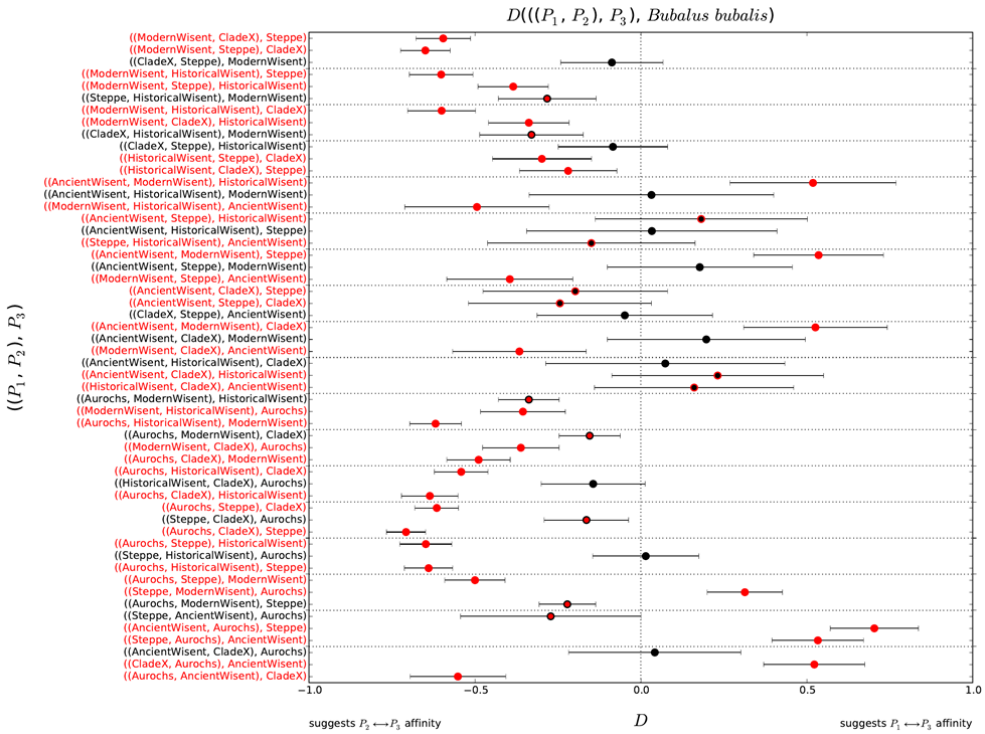
Supplementary Fig 11: Topology testing using D statistics, with sheep as outgroup. The topology being tested is shown on the vertical axis, with the most parsimonious of three possible topologies written in black. Data points that are significantly different (more than three standard errors) from zero are shown in red. The data point representing the topology closest to zero, amongst a set of three possible topologies, is shown with a black outline. Error bars are three standard errors either side of the data point, where the standard error was calculated using a block jackknife.



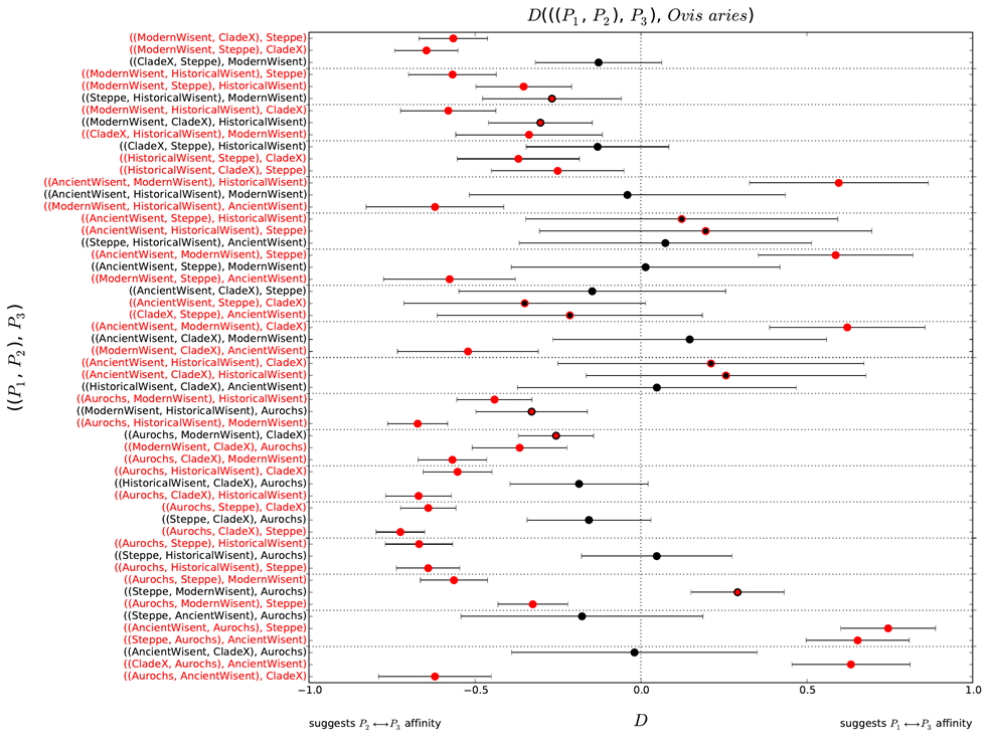
Supplementary Fig 12: Topology testing using D statistics, with sheep as outgroup. As in Supplementary Figure 11, except that sample A006 has been omitted from the CladeX group.



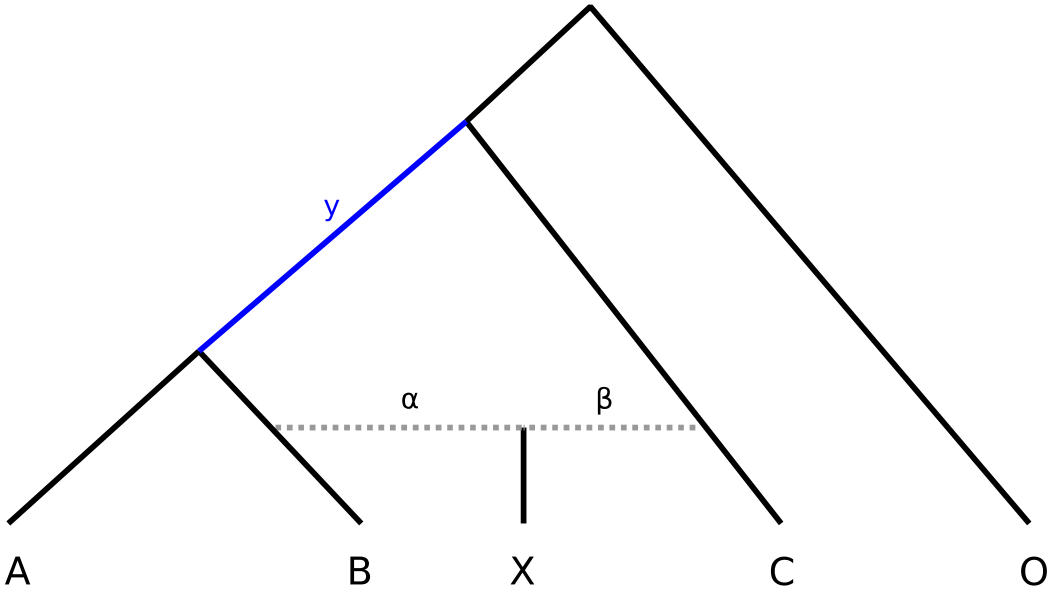
Supplementary Fig 13: Topology testing using D statistics, with sheep as outgroup. As in Supplementary Figure 11, except that genotypes called from read depths <2 have been omitted for extinct individuals.



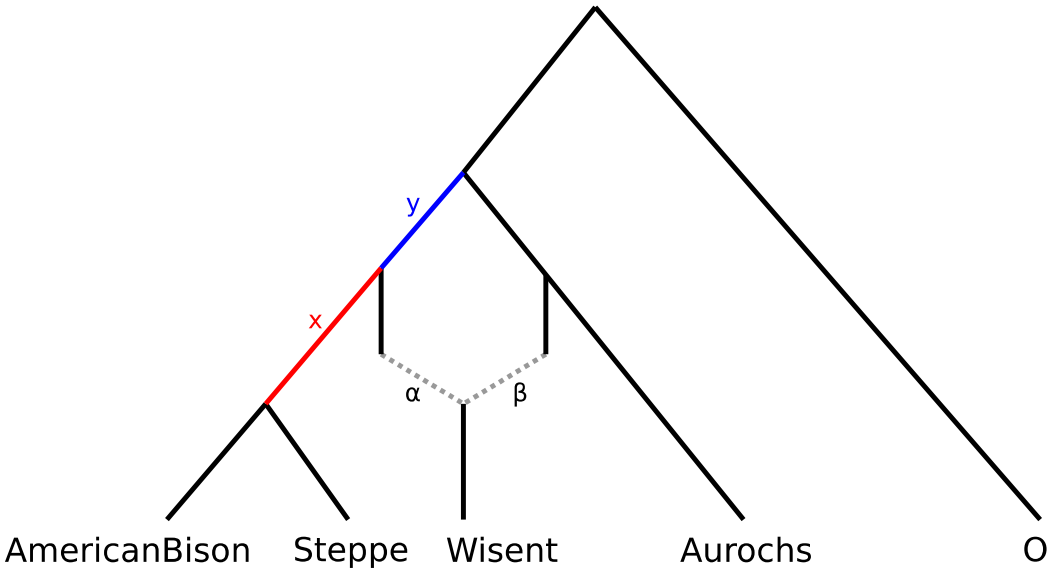
Supplementary Fig 14: Topology testing using D statistics, with Asian water buffalo as outgroup. As in Supplementary Figure 11, except the outgroup has been changed.



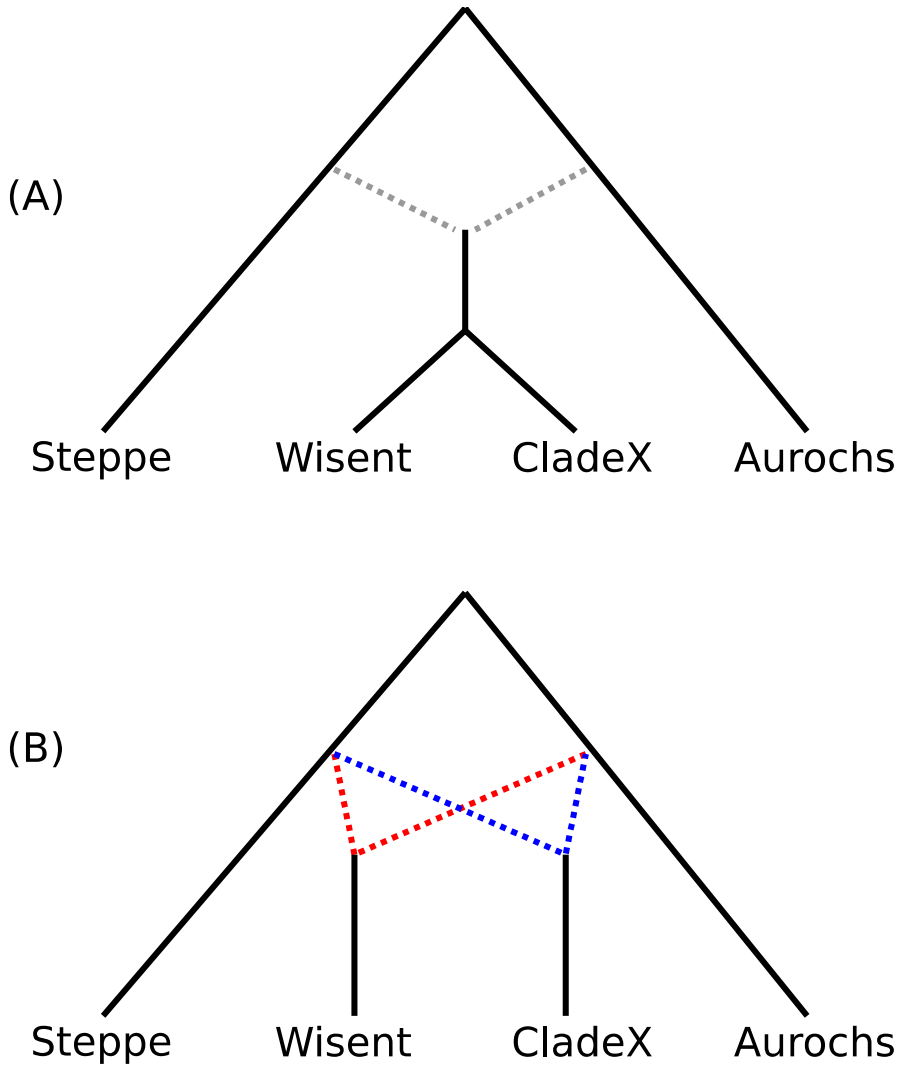
Supplementary Fig 15: Topology testing using D statistics, with sheep as outgroup. As in Supplementary Figure 11, except in extinct individuals, alleles have been randomly sampled from sites called as heterozygotes to simulate haploid sampling.



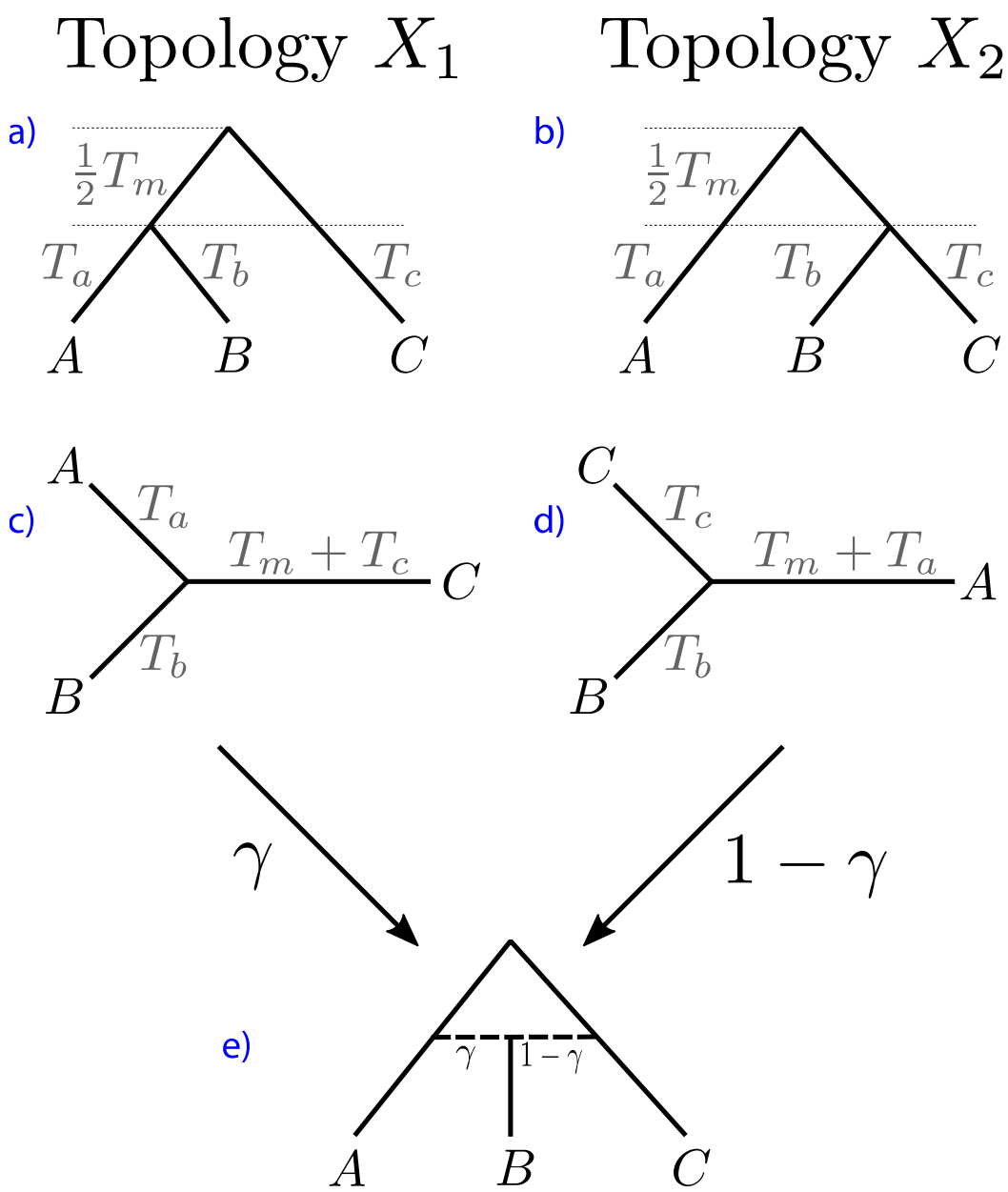
Supplementary Fig 16: An admixture graph showing the ancestry of X, where α is the proportion of ancestry from B and β=1-α is the proportion of ancestry from C.



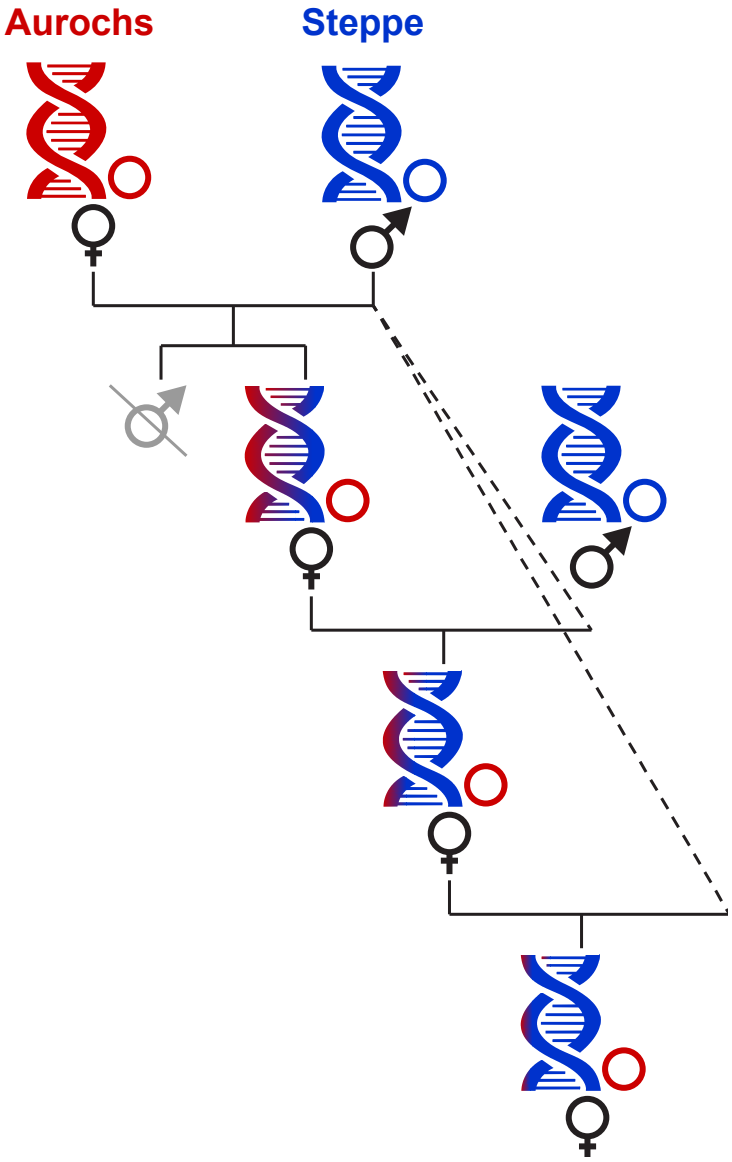
Supplementary Fig 17: An admixture graph showing the ancestry of the wisent, where α is the proportion of ancestry from steppe and β=1-α is the proportion of ancestry from aurochs.



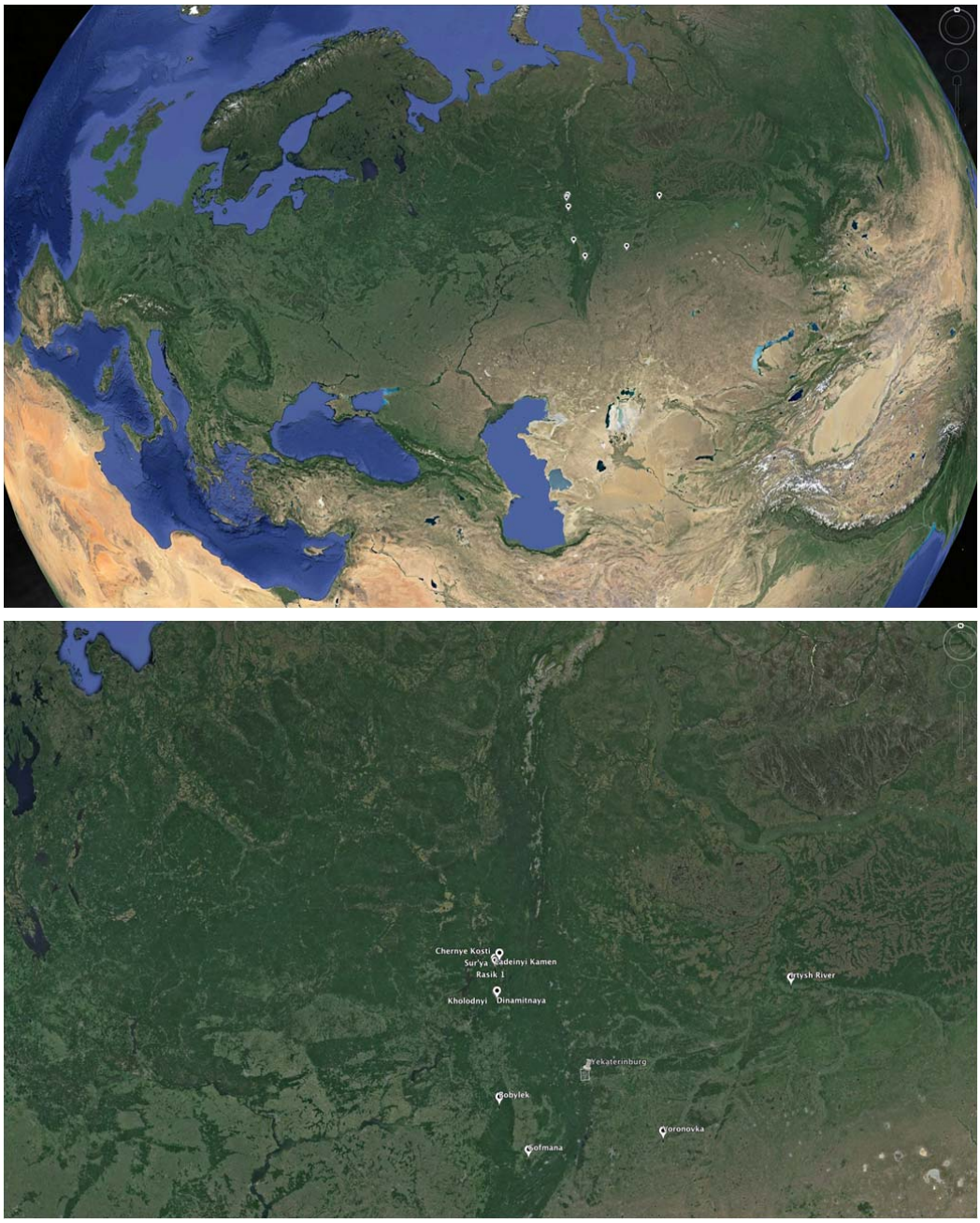
Supplementary Fig 18: Admixture graphs representing (A) a single hybridisation event prior to the divergence of the wisent, and (B) two independent hybridisation events leading to a wisent clade and a CladeX.



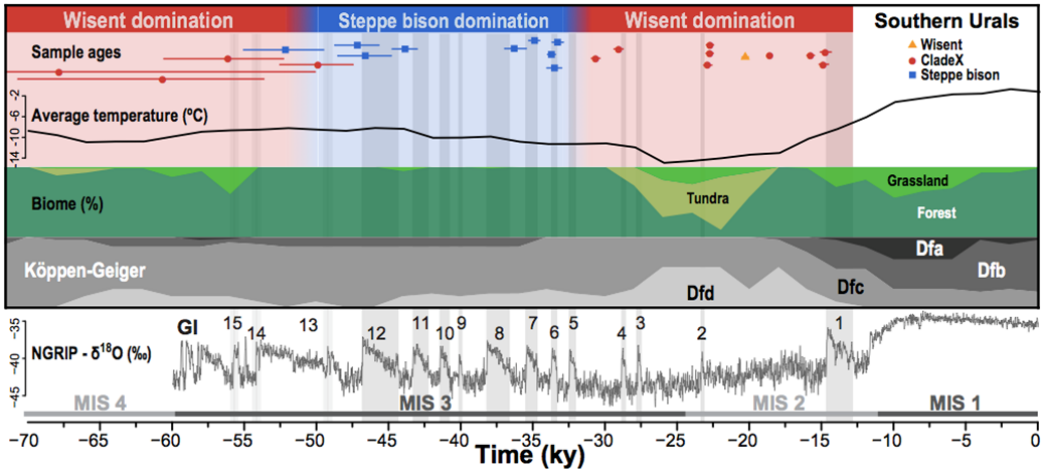
Supplementary Fig 19: A hybrid species tree (e), where individual B is a hybrid of A and C lineages, has two contributing species trees, (a) topology X_1, and (b) topology X_2, with proportion γ from topology X_1 and proportion $1-\gamma$ from topology X_2. The unrooted gene trees are shown for (c) topology X_1, and (d) topology X_2. Branch lengths T_a, T_b, T_c and T_m have units $2N_e\mu$ generations.



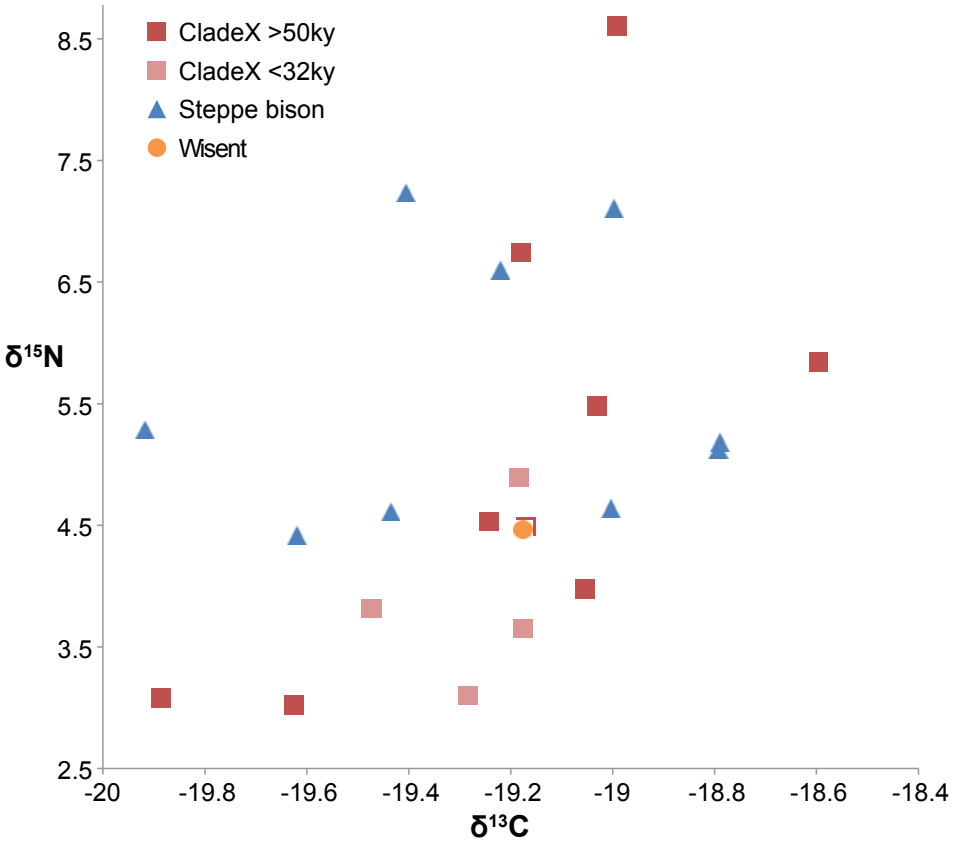
Supplementary Fig 20. Schematic representation of asymmetrical hybridisation between female aurochs and male steppe bison, and its genetic imprint on both nuclear and mitochondrial genomes after a few generations. The coloured double helix represents the nuclear genome, while the circles represent the strictly maternally inherited mitochondrial genome.



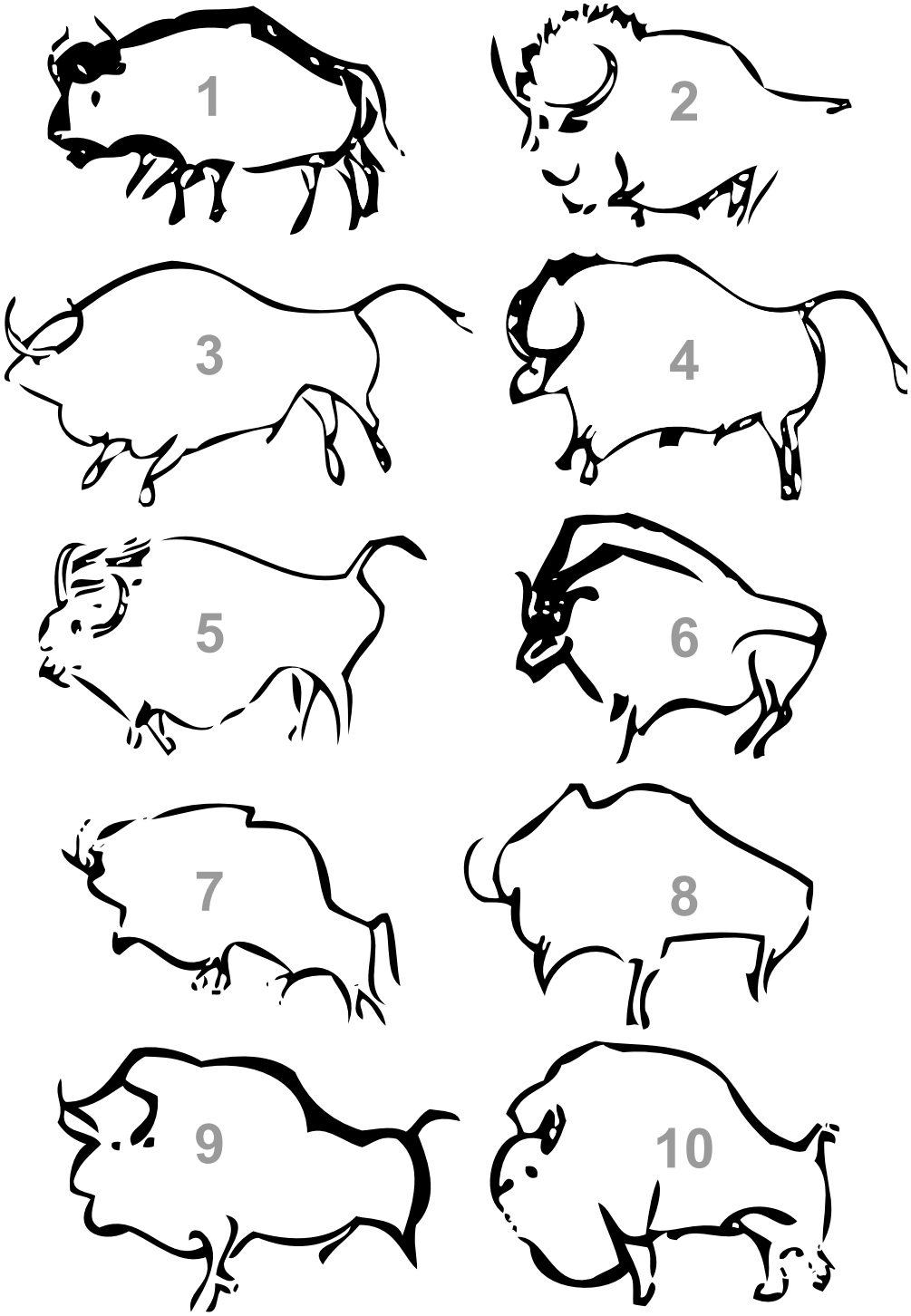
Supplementary Fig 21. Location of all cave sites from which bison samples have been genotyped in the Ural region.



Supplementary Fig 22. Chronology of the Urals samples showing a series of replacement patterns that correlate with climate events. Individual calibrated AMS dates are plotted on top of the NGRIP δO^{18} record [1]. Greenland Interstadials (GI) are numbered in black, and Marine Isotope Stages (MIS) in grey. Inferred average temperature, biome reconstruction and proportion of the area for different Koppen climate classes are shown for the exact region where bison were sampled in southern Urals (Koppen classes: D for 'snow', f for 'fully humid', then a=hot summer; b=warm summer; c=cool summer; d=extremely continental). The most recent population replacement between wisent and steppe bison occurs around 32-33 ky, when major environmental transitions are also observed: 1) Globally, as shown on the NGRIP record with the last major interglacial event (GI 5) before a long period of cold climate; but also 2) Locally, as shown on both the average temperature and biome reconstructions. In this situation, wisent are associated with a cooler climate and the presence of tundra-like vegetation. Although dating resolution is degrading for deeper time, a similar shift is apparent around 50-52 kya. Steppe bison occupied this environment in MIS 3, but have not been detected after this stage and indeed were in a severe population decline by GI 1[2].



Supplementary Fig 23. Stable δ13C and δ15N isotope values for all genotyped bison sampled from the Ural region.



Supplementary Fig 24. Steppe-like morphologies. In European Palaeolithic art, some bison depictions show morphological traits and anatomical details compatible with the morphology of steppe bison (or American bison ancestry). Dates are given as indication based on archaeological occupation determined for each site, or, in the absence of such dating, based on stylistic comparison with other depictions:

1. Grotte Chauvet-Pont d'Arc (Ardèche, France). Blurred black charcoal drawing. Aurignacian period (~35,100 ± 175 calBP. (from C. Fritz and G. Tosello)

2. Grotte de Lascaux (Dordogne, France). Carving. Solutrean (~22,200 ± 380 calBP) or early Magdalenian period (between ~19,300 ± 561 and ~20,597 ± 375 calBP). (adapted from A. Glory[3])

3. Grotte de Lascaux (Dordogne, France). Carving. Solutrean (~22,200 ± 380 calBP) or early Magdalenian period (between ~19,300 ± 561 and ~20,597 ± 375 calBP). (adapted from A. Glory[3])

4. Grotte de Lascaux (Dordogne, France). Carving. Solutrean (~22,200 ± 380 calBP) or early Magdalenian period (between ~19,300 ± 561 and ~20,597 ± 375 calBP). (adapted from A. Glory[3])

5. Grotte du Gabillou (Dordogne, France). Carving. Early Magdalenian period (~20,597 ± 375 calBP). (adapted from J. Gaussen)

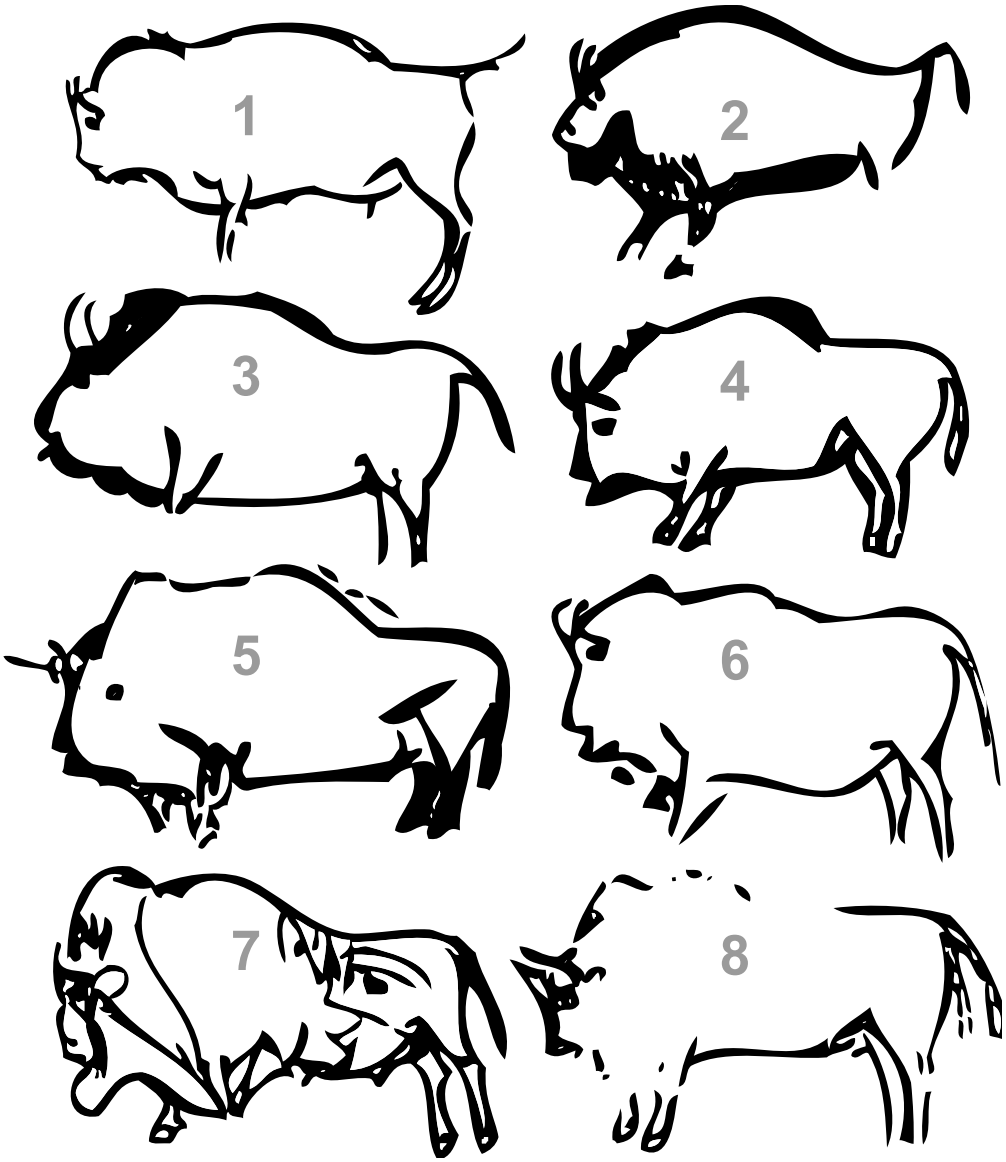
6. Grotte des Trois Frères (Ariège, France). Carving. Gravettian period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

7. Grotte du Pech Merle (Lot, France). Painting (manganese). Gravettian period (~29,447 ± 443 calBP). (adapted from M. Lorblanchet[5])

8. Grotte du Pech Merle (Lot, France). Painting (manganese). Gravettian period (~29,447 ± 443 calBP). (adapted from M. Lorblanchet[5])

9. Grotte de La Pasiega (Cantabria, Spain). Black and red painting. Gravettian or Solutrean period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

10. Abri du Roc de Sers (Charente, France). Carving on limestone. Solutrean period (< 20,442 ± 409 calBP). (adapted from L. Henri-Martin)



Supplementary Fig 25. Wisent-like morphologies. In European Palaeolithic art, some bison depictions show morphological traits and anatomical details compatible with identification of wisent ancestry. Dates are given as indication based on archaeological occupation determined for each site, or, in the absence of such dating, based on stylistic comparison with other depictions:

1. Grotte de Pergouset (Ardèche, France). Carving. Magdalenian period (dating estimated based on stylistic analysis). (adapted from M. Lorblanchet[5])

2. Grotte du Portel (Ariège, France). Painting. Magdalenian period (~14,250 ± 295 calBP). (adapted from H. Breuil[4])

3. Grotte de Niaux (Ariège, France). Painting. Magdalenian period (~17,000 ± 260 calBP). (adapted from H. Breuil[4])

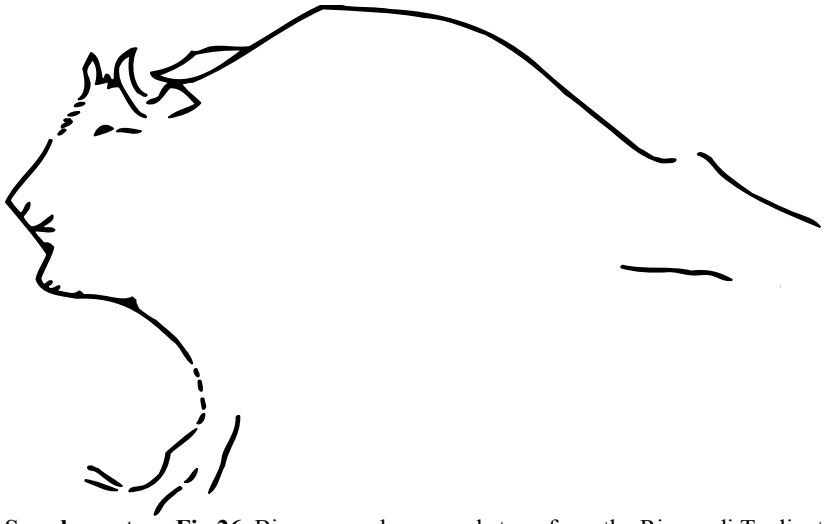
4. Grotte de Niaux (Ariège, France). Painting. Magdalenian period (~17,000 ± 260 calBP). (adapted from H. Breuil[4])

5. Grotte de Fontanet (Ariège, France). Carving. Magdalenian period (between ~14250 ± 295 calBP and ~16,600 ± 1000 calBP). (adapted from A. Glory[3])

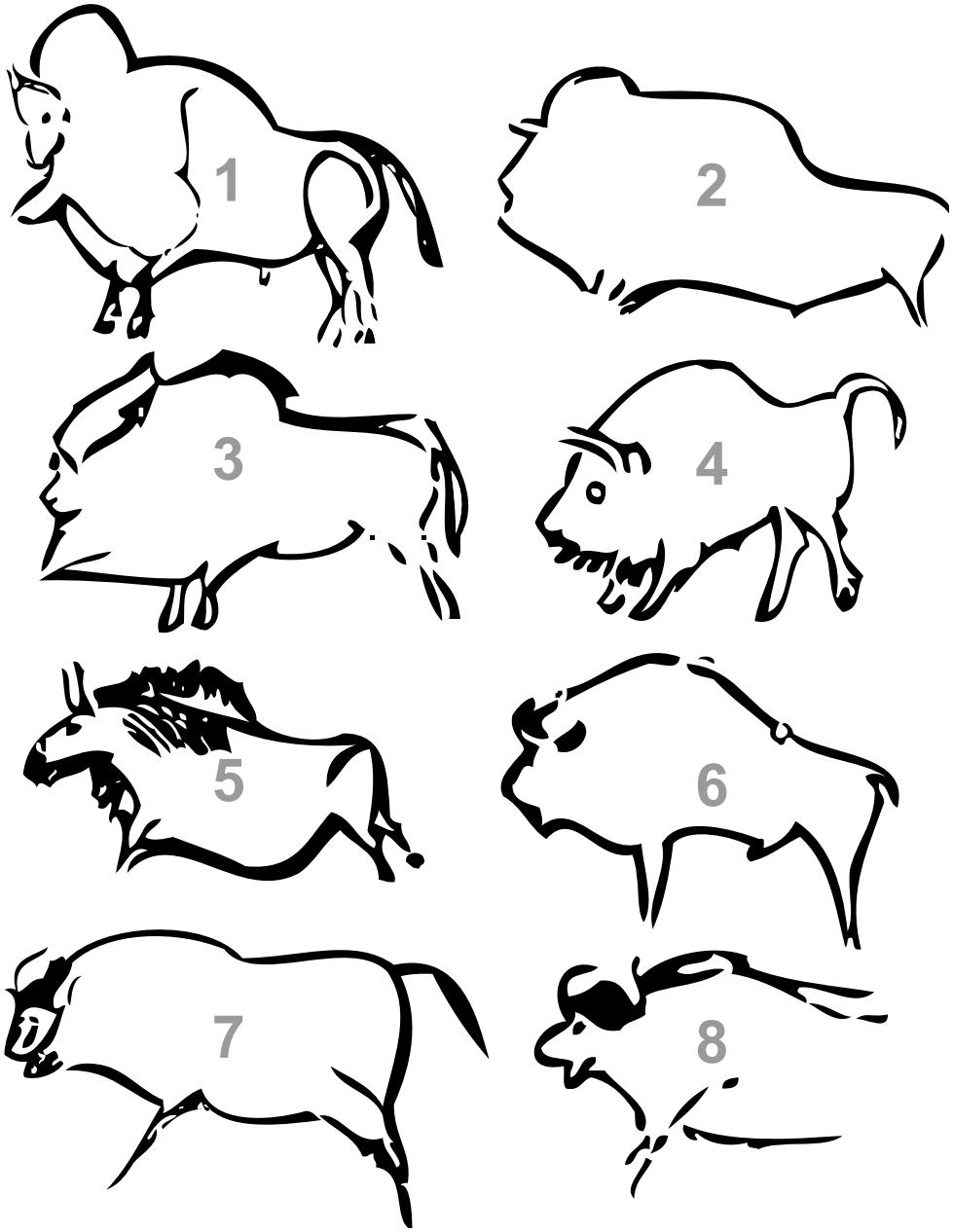
6. Grotte de Rouffignac (Dordogne, France). Painting. Magdalenian period (dating estimated based on stylistic analysis). (adapted from C. Barrière[6])

7. Grotte des Combarelles (Dordogne, France). Carving. Magdalenian period (between ~17,000 and ~14,300 calBP). (adapted from H. Breuil[4])

8. Grotte de Marsoulas (Haute-Garonne, France). Carving. Magdalenian period (dating estimated based on stylistic analysis). (from C. Fritz et G. Tosello)



Supplementary Fig 26. Bison carved on round stone from the Riparo di Tagliente site in Italy



Supplementary Fig 27. Undetermined morphologies. In European Palaeolithic art, some bison depictions show morphological traits and anatomical details that could be compatible with either bison form. These pictures illustrate the limits of cave art analyses for morphological assessment of bison forms, due to varying graphical conventions between cultures. Dates are given as indication based on archaeological occupation determined for each site, or, in the absence of such dating, based on stylistic comparison with other depictions:

1 Grotte de Font-de-Gaume (Dordogne, France). Black and red painting, and carving. Magdalenian period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

2 Grotte de Niaux (Ariège, France). Painting. Magdalenian period (~17,000 ± 260 calBP). (adapted from H. Breuil[4])

3 Grotte des Trois Frères (Ariège, France). Carving. Magdalenian period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

4 Grotte des Trois Frères (Ariège, France). Carving. Magdalenian period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

5 Grotte des Trois Frères (Ariège, France). Carving. Gravettian period (dating estimated based on stylistic analysis). (adapted from H. Breuil[4])

6 Grotte de La Grèze (Dordogne, France). Carving. Gravettian period (dating estimated based on stylistic analysis) (adapted from N. Aujoulat)

7 Grotte Chauvet-Pont d'Arc (Ardèche, France). Blured black charcoal drawing. Aurignacian period (~35100 ± 175 calBP). (from C. Fritz-G. Tosello)

8 Grotte Chauvet-Pont d'Arc (Ardèche, France). Blured black charcoal drawing. Aurignacian period (~35100 ± 175 calBP). (from C. Fritz-G. Tosello)

Supplementary Tables

Supplementary Table 1. Primers and adapters used in this study

	Primer	Primer Sequence (5' - 3')	Length [a]
Set_A1	BovCR-16351F	CAACCCCCAAAGCTGAAG	~96bp
	BovCR-16457R	TGGTTRGGGTACAAAGTCTGTG	
Set_B1	BovCR-16420F	CCATAAATGCAAAGAGCCTCAYCAG	~172bp
	BovCR-16642R	TGCATGGGGCATATAATTTAATGTA	
Set_A2	BovCR-16507F	AATGCATTACCCAAACRGGG	~184bp
	BovCR-16755R	ATTAAGCTCGTGATCTARTGG	
Set_B2	BovCR-16633F[b]	GCCCCATGCATATAAGCAAG	~132bp
	BovCR-16810R[b]	GCCTAGCGGGTTGCTGGTTTCACGC	
Set_A3	BovCR-16765F[b]	GAGCTTAAYTACCATGCCG	~125bp
	BovCR-16998R	CGAGATGTCTTATTTAAGAGGAAAGAATGG	
Set_B3	BovCR-16960F	CATCTGGTTCTTTCTTCAGGGCC	~110bp
	BovCR-80R[b]	CAAGCATCCCCCAAAATAAA	
Frag1	BovCR_16738MF[c,d]	***CACGACGTTGTAAAACGAC***ATYGTACATAGYACATTATGTCAA	~67bp
	BovCR_16810TR[c,d]	***TACGACTCACTATAGGGCGA***GCCTAGCGGGTTGCTGGTTTCACGC	
Frag2	Mamm_12SE[d]	CTATAATCGATAAACCCCGATA	~96bp
	Mamm_12SH[d]	GCTACACCTTGACCTAAC	
	GAII_Indexing_x	CAAGCAGAAGACGGCATACGAGATNNNNNNNGAGTGACTGGAGTTCAGACGTGT	n/a
	IS4_indPCR.P5[e]	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT	n/a
	IS7_short_amp.P5[e]	ACACTCTTTCCCTACACGAC	n/a
	IS8_short_amp.P7[e]	GTGACTGGAGTTCAGACGTGT	n/a
	P5_short_RNAblock	ACACUCUUUCCCUACACGAC	n/a
	P7_short_RNAblock	GUGACUGGAGUUCAGACGUGU	n/a
	Bison_mt1_forward [f]	ACCGCGGTCATACGATTAAC	
	Bison_mt1_reverse [f]	AATTGCGAAGTGGATTTTGG	
	Bison_mt2_forward [f]	ATGAGCCAAAATCCACTTCG	
	Bison_mt2_reverse [f]	TGTATTTGCGTCTGCTCGTC	
	Bison_mt3_forward [f]	CGAATCCACAGCCGAACTAT	
	Bison_mt3_reverse [f]	TATAAAGCACCGCCAAGTCC	

(a): Primers are excluded from the length of PCR amplicon.

(b):[2].

(c): M13 (CAC GAC GTT GTA AAA CGA C) and T7 (TAC GAC TCA CTA TAG GGC GA) sequences were used as tags for primers BovCR_16738F and BovCR_16810R, respectively. This was done to obtain good quality Sanger sequences from short amplicons.

(d): One-step simplex PCRs.

(e): (Meyer and Kircher, "Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing.")

(f): Primer pairs for use to generate DNA baits for mitochondrial DNA capture.

Supplementary Table 2. Summary of nuclear alleles detected at bovine SNP loci: NGS results and locus counts for ancient samples; locus counts for modern samples

		Mapping results for the 9908 SNP positions						Number of SNP called out of the 9908 targeted for each ancient individuals							
								Coverage depth >=1				Coverage depth >=2			
Sample ID	**Method**	**Retained_reads**	**hits_raw**	**hits_unique**	**hits_raw_frac**	**hits_clonality**	**Mean coverage**	**Total**	REF/REF	REF/ALT	ALT/ALT	**Total**	REF/REF	REF/ALT	ALT/ALT
A15526		7045	1821	99	0.26	0.95	0.01	**49**	49	0	0	**1**	1	0	0
A017		1280556	3893	1289	0.00	0.67	0.13	**630**	591	0	39	**88**	49	0	39
A018		967346	3116	538	0.00	0.83	0.05	**253**	241	0	12	**28**	16	0	12
A001		656008	392937	3486	0.60	0.99	0.35	**1484**	1268	2	214	**523**	307	2	214
A003		1706985	12957	3423	0.01	0.74	0.35	**1569**	1363	5	201	**470**	264	5	201
A004	10k capture	240370	132883	645	0.55	1.00	0.07	**315**	287	0	28	**64**	36	0	28
A005		1736500	25788	3519	0.01	0.86	0.35	**1643**	1438	7	198	**464**	259	7	198
A006		10413909	99392	22312	0.01	0.78	2.25	**5690**	3468	104	2118	**4755**	2533	104	2118
A007		3583539	23832	2841	0.01	0.88	0.29	**1307**	1084	1	222	**509**	286	1	222
A15654		1700840	1227601	220913	0.72	0.82	22.28	**8738**	4532	230	3976	**8488**	4282	230	3976
A4093		9400283	62631	4478	0.01	0.93	0.45	**1946**	1480	2	464	**1031**	565	2	464
A3133	Shotgun / 10k	299829433	9812523	465082	0.03	0.95	46.87	**8898**	4579	321	3998	**8680**	4361	321	3998
A875	and 40k capture	3908972	291640	234493	0.07	0.20	23.65	**8433**	4341	342	3750	**8144**	4052	342	3750
CPC98_Aurochs	From published genome							**8882**	4770	1808	2304	**8810**	4698	1808	2304

Supplementary Table 3. Summary statistics for NGS of whole mitochondrial genomes

Sample ID	Retained_reads	hits_raw	hits_unique	hits_raw_frac	hits_clonality	AVG_Depth	STD_Depth	AVE_Length	STD_Length	5pC>T	3pG>A	Library repair
A001	4822143	1618364	86944	0.34	0.95	432.09	224.83	80.82	37.60	0.03	0.02	
A004	5150804	2314449	220697	0.45	0.90	1152.17	541.88	84.88	36.11	0.02	0.02	
A018	3790161	1021750	24699	0.27	0.98	130.53	60.04	85.32	34.05	0.03	0.03	USER
A4089	8618722	5380606	44044	0.62	0.99	237.83	155.46	87.18	33.56	0.02	0.02	
A3133	66864927	1958	1949	0.00	0.00	11.41	6.77	93.92	29.66	0.00	0.01	
A003	985033	371605	64372	0.38	0.83	334.44	112.68	84.31	34.07	0.08	0.07	
A005	521428	262622	39121	0.50	0.85	196.95	65.76	81.59	30.96	0.05	0.09	
A006	456078	120668	44541	0.26	0.63	208.39	93.86	75.86	25.87	0.13	0.17	
A007	431113	175432	43269	0.41	0.75	192.35	85.93	71.74	24.13	0.11	0.08	Partial UDG
A4093	212315	106221	16923	0.50	0.84	73.23	31.26	70.48	24.60	0.07	0.09	
A15637	469884	4401	2621	0.01	0.40	8.85	7.22	50.41	12.17	0.41	0.35	
A15654	294965	29628	28329	0.10	0.04	170.48	89.68	98.23	34.91	0.05	0.02	
A15668	230709	3603	2842	0.02	0.21	11.07	7.80	59.61	15.06	0.07	0.06	
LE237	507023	4271	2677	0.01	0.37	9.84	5.70	58.98	23.99	0.55	0.51	
LE242	6912671	48793	35418	0.01	0.27	120.46	67.86	55.09	18.68	0.61	0.60	None
LE257	4156307	184236	28788	0.04	0.84	94.38	38.34	53.17	20.00	0.52	0.50	

Supplementary Table 4. List of published mitochondrial control region sequences used for phylogenetic analysis. The Urals steppe bison are highlighted in red.

American bison	Bison_priscus_BS146_NS_11810_50	Bison_priscus_BS397_NS_32370_470	Bos_indicus_AY378135_0_0
Bison_bison_AF083357_H1_0_0	Bison_priscus_BS147_NS_28120_290	Bison_priscus_BS398_NS_27400_260	Bos_indicus_DQ887765_0_0
Bison_bison_AF083358_H2_0_0	Bison_priscus_BS148_NS_6400_50	Bison_priscus_BS400_NS_46100_2600	Bos_indicus_EF417971_0_0
Bison_bison_AF083359_H3_0_0	Bison_priscus_BS149_NS_46100_2200	Bison_priscus_BS405_SI_23040_120	Bos_indicus_EF417974_0_0
Bison_bison_AF083360_H4_0_0	Bison_priscus_BS150_NS_10510_50	Bison_priscus_BS407_NWT_55500_3100	Bos_indicus_EF417976_0_0
Bison_bison_AF083361_H5_0_0	Bison_priscus_BS151_NS_21530_130	Bison_priscus_BS412_Y_30500_250	Bos_indicus_EF417977_0_0
Bison_bison_AF083362_H6_0_0	Bison_priscus_BS161_NS_21040_120	Bison_priscus_BS414_BlR_4495_60	Bos_indicus_EF417979_0_0
Bison_bison_AF083363_H7_0_0	Bison_priscus_BS163_LC_13240_75	Bison_priscus_BS415_D_30810_975	Bos_indicus_EF417981_0_0
Bison_bison_AF083364_H8_0_0	Bison_priscus_BS164_LC_19540_120	Bison_priscus_BS418_China_26560_670	Bos_indicus_EF417983_0_0
Bison_bison_BS100_29_5	Bison_priscus_BS165_LC_26460_160	Bison_priscus_BS438_AB_53800_2200	Bos_indicus_EF417985_0_0
Bison_bison_BS102_22_5	Bison_priscus_BS170_YT_13040_70	Bison_priscus_BS440_AB_60400_2900	Bos_indicus_EF524120_0_0
Bison_bison_BS129_0_2000	Bison_priscus_BS172_LC_12525_70	Bison_priscus_BS443_AB_34050_450	Bos_indicus_EF524125_0_0
Bison_bison_BS162_AK_170_30	Bison_priscus_BS176_LC_12380_60	Bison_priscus_BS459_China_47700_1000	Bos_indicus_EF524126_0_0
Bison_bison_BS173_NTC_3220_45	Bison_priscus_BS178_LC_17960_90	Bison_priscus_BS469_AB_305_24	Bos_indicus_EF524128_0_0
Bison_bison_BS175_ICE_186_30	Bison_priscus_BS192_F_26300_300	Bison_priscus_BS472_F_13235_65	Bos_indicus_EF524130_0_0
Bison_bison_BS177_NTC_3155_36	Bison_priscus_BS193_NS_49600_4000	Bison_priscus_BS473_AB_56300_3100	Bos_indicus_EF524132_0_0
Bison_bison_BS200_AB_145_37	Bison_priscus_BS195_NS_29040_340	Bison_priscus_BS477_D_33710_240	Bos_indicus_EF524135_0_0
Bison_bison_BS342_CHL_10340_40	Bison_priscus_BS196_NS_19420_100	Bison_priscus_BS478_D_34470_200	Bos_indicus_EF524141_0_0
Bison_bison_BS348_CHL_10505_45	Bison_priscus_BS198_Y_2460_40	Bison_priscus_BS490_BlR_2415_25	Bos_indicus_EF524152_0_0
Bison_bison_BS368_0_2000	Bison_priscus_BS201_Y_12960_60	Bison_priscus_BS493_NS_50000_4200	Bos_indicus_EF524156_0_0
Bison_bison_BS417_AB_909_29	Bison_priscus_BS202_AB_10460_65	Bison_priscus_BS494_NS_44800_2200	Bos_indicus_EF524160_0_0
Bison_bison_BS419_AB_7475_45	Bison_priscus_BS206_Sibh_23780_140	Bison_priscus_BS495_NS_29570_340	Bos_indicus_EF524166_0_0
Bison_bison_BS421_AB_8145_45	Bison_priscus_BS211_Sibh_43800_1100	Bison_priscus_BS497_NS_30000_540	Bos_indicus_EF524167_0_0
Bison_bison_BS422_AB_908_31	Bison_priscus_BS216_NS_47000_2900	Bison_priscus_BS498_NS_25980_230	Bos_indicus_EF524170_0_0
Bison_bison_BS423_AB_4660_38	Bison_priscus_BS218_Si_14605_75	Bison_priscus_BS499_NS_31410_420	Bos_indicus_EF524177_0_0
Bison_bison_BS424_AB_202_32	Bison_priscus_BS222_NWT_6110_45	Bison_priscus_BS500_NS_35580_550	Bos_indicus_EF524180_0_0
Bison_bison_BS426_AB_7060_45	Bison_priscus_BS223_Si_53300_1900	Bison_priscus_BS517_BlR_2526_26	Bos_indicus_EF524183_0_0
Bison_bison_BS428_AB_7105_45	Bison_priscus_BS224_AK_13125_75	Bison_priscus_BS564_Si_24570_90	Bos_indicus_EF524185_0_0
Bison_bison_BS429_AB_6775_40	Bison_priscus_BS233_SW_16685_80	Bison_priscus_BS571_SIdy_32910_170	Bos_indicus_L27732_0_0
Bison_bison_BS430_9270_50	Bison_priscus_BS235_BlR_43400_900	Bison_priscus_BS592_Urals_42500_450	Bos_indicus_L27736_0_0
Bison_bison_BS432_AB_7310_45	Bison_priscus_BS236_SW_19420_100	Bison_priscus_BS605_NTC_20380_90	**Aurochs**
Bison_bison_BS433_AB_10450_55	Bison_priscus_BS237_AB_11240_70	Bison_priscus_BS660_Urals_29500_140	Bos_primigenius_DQ915522_ALL1_12030_52
Bison_bison_BS434_AB_809_32	Bison_priscus_BS243_SW_37550_400	Bison_priscus_BS662_SI_20000_0	Bos_primigenius_DQ915523_CAT1_5650_0
Bison_bison_BS439_AB_5845_45	Bison_priscus_BS244_LC_26210_170	Bison_priscus_BS674_Urals_29060_140	Bos_primigenius_DQ915524_CHWF_3905_185
Bison_bison_BS441_AB_1273_32	Bison_priscus_BS248_OCr_12350_70	Bison_priscus_BS708_Urals_47050_750	Bos_primigenius_DQ915537_CPC98_5936_34
Bison_bison_BS444_AB_636_29	Bison_priscus_BS249_F_39200_550	Bison_priscus_BS713_Urals_30970_180	Bos_primigenius_DQ915542_EIL06_5830_29
Bison_bison_BS445_AB_378_30	Bison_priscus_BS253_LC_12665_65	Bison_priscus_IB179_LC_12465_75	Bos_primigenius_DQ915543_EIL14_5830_29
Bison_bison_BS449_6195_45	Bison_priscus_BS254_CHL_10230_55	**European bison**	Bos_primigenius_DQ915554_LJU3_8020_50
Bison_bison_BS454_AB_287_29	Bison_priscus_BS258_F_22120_130	Bison_bonasus_AF083356_0_0	Bos_primigenius_DQ915558_NORF_3370_30
Bison_bison_BS456_AB_125_30	Bison_priscus_BS260_D_30750_290	Bison_bonasus_AY428860_0_0	Bos_primigenius_EF187280_PVL04_3204_56
Bison_bison_BS460_AB_10425_50	Bison_priscus_BS261_LC_12915_70	Bison_bonasus_EF693811_0_0	**Cattle**
Bison_bison_BS464_AB_5205_45	Bison_priscus_BS262_D_29150_500	Bison_bonasus_EU272053_0_0	Bos_taurus_DQ124372_T4_0_0
Bison_bison_BS465_AB_7115_50	Bison_priscus_BS281_BlR_40800_600	Bison_bonasus_EU272054_0_0	Bos_taurus_DQ124375_T4_0_0
Bison_bison_BS466_AB_3298_37	Bison_priscus_BS282_Si_56700_3200	Bison_bonasus_EU272055_0_0	Bos_taurus_DQ124381_T3_0_0
Bison_bison_BS503_BlR_2776_36	Bison_priscus_BS284_Y_13135_65	Bison_bonasus_U12953_0_0	Bos_taurus_DQ124383_T2_0_0
Bison_bison_BS560_AB_2807_28	Bison_priscus_BS286_Sim_49500_1300	Bison_bonasus_U12954_0_0	Bos_taurus_DQ124388_T3_0_0
Bison_bison_BS569_AB_3600_70	Bison_priscus_BS287_BlR_49100_1700	Bison_bonasus_U34294_0_0	Bos_taurus_DQ124394_T3_0_0
Bison_bison_BS570_AB_11300_290	Bison_priscus_BS289_BlR_2172_37	**Yak**	Bos_taurus_DQ124398_T3_0_0
Bison_bison_BS99_26_5	Bison_priscus_BS291_NS_49700_1400	Bos_grunniens_AY521140_0_0	Bos_taurus_DQ124400_T4_0_0
Bison_bison_U12935_0_0	Bison_priscus_BS292_NS_35710_730	Bos_grunniens_AY521149_0_0	Bos_taurus_DQ124401_T4_0_0
Bison_bison_U12936_0_0	Bison_priscus_BS294_BlR_58200_3900	Bos_grunniens_AY521150_0_0	Bos_taurus_DQ124412_T4_0_0
Bison_bison_U12941_0_0	Bison_priscus_BS297_NS_10990_50	Bos_grunniens_AY521151_0_0	Bos_taurus_EU177822_T3_0_0
Bison_bison_U12943_0_0	Bison_priscus_BS311_BlR_12425_45	Bos_grunniens_AY521152_0_0	Bos_taurus_EU177841_T1_0_0
Bison_bison_U12944_0_0	Bison_priscus_BS316_SI_57700_3000	Bos_grunniens_AY521154_0_0	Bos_taurus_EU177842_T1_0_0
Bison_bison_U12945_0_0	Bison_priscus_BS318_NS_12410_50	Bos_grunniens_AY521155_0_0	Bos_taurus_EU177845_T1_0_0
Bison_bison_U12946_0_0	Bison_priscus_BS320_SI_49600_1500	Bos_grunniens_AY521156_0_0	Bos_taurus_EU177847_T1_0_0
Bison_bison_U12947_0_0	Bison_priscus_BS321_AK_9506_38	Bos_grunniens_AY521160_0_0	Bos_taurus_EU177848_T1_0_0
Bison_bison_U12948_0_0	Bison_priscus_BS323_SI_37810_380	Bos_grunniens_AY521161_0_0	Bos_taurus_EU177853_T2_0_0
Bison_bison_U12955_0_0	Bison_priscus_BS327_D_31530_230	Bos_grunniens_DQ007210_0_0	Bos_taurus_EU177854_T2_0_0
Bison_bison_U12956_0_0	Bison_priscus_BS328_SIdy_31690_180	Bos_grunniens_DQ007221_0_0	Bos_taurus_EU177860_T2_0_0
Bison_bison_U12957_0_0	Bison_priscus_BS329_D_27060_190	Bos_grunniens_DQ007222_0_0	Bos_taurus_EU177861_T2_0_0
Bison_bison_U12958_0_0	Bison_priscus_BS337_CHL_10378_36	Bos_grunniens_DQ856594_0_0	Bos_taurus_EU177862_T5_0_0
Bison_bison_U12959_0_0	Bison_priscus_BS340_NS_24500_180	Bos_grunniens_DQ856599_0_0	Bos_taurus_EU177863_T5_0_0
Steppe bison	Bison_priscus_BS345_NS_39800_1200	Bos_grunniens_DQ856600_0_0	Bos_taurus_EU177864_T5_0_0
Bison_priscus_A3133_Yukon_26360_220	Bison_priscus_BS350_NS_38700_1000	Bos_grunniens_DQ856603_0_0	Bos_taurus_EU177865_T5_0_0
Bison_priscus_BS105_F_23380_460	Bison_priscus_BS351_BlR_57700_3200	Bos_grunniens_DQ856604_0_0	**Buffalo**
Bison_priscus_BS107_F_19570_290	Bison_priscus_BS359_NTC_20020_150	Bos_grunniens_EF494177_0_0	Bubalus_bubalis_AF197208_0_0
Bison_priscus_BS108_F_21020_360	Bison_priscus_BS364_NS_38800_1100	Bos_grunniens_EF494178_0_0	Bubalus_bubalis_AF475212_0_0
Bison_priscus_BS109_F_20730_350	Bison_priscus_BS365_NS_47000_2900	**Zebu**	Bubalus_bubalis_AF475256_0_0
Bison_priscus_BS111_F_21580_370	Bison_priscus_BS387_NS_33320_540	Bos_indicus_AB085923_0_0	Bubalus_bubalis_AF475259_0_0
Bison_priscus_BS121_F_19360_280	Bison_priscus_BS388_NS_27590_280	Bos_indicus_AB268563_0_0	Bubalus_bubalis_AF475278_0_0
Bison_priscus_BS123_BlR_1730_60	Bison_priscus_BS389_NS_17160_80	Bos_indicus_AB268564_0_0	Bubalus_bubalis_AY488491_0_0
Bison_priscus_BS124_BlR_11900_70	Bison_priscus_BS390_NS_31630_440	Bos_indicus_AB268566_0_0	Bubalus_bubalis_EF536327_0_0
Bison_priscus_BS125_F_27440_790	Bison_priscus_BS392_NS_36320_780	Bos_indicus_AB268571_0_0	Bubalus_bubalis_EF536328_0_0
Bison_priscus_BS126_F_19150_280	Bison_priscus_BS393_NS_39850_1200	Bos_indicus_AB268574_0_0	Bubalus_bubalis_EU268899_0_0
Bison_priscus_BS130_BlR_9000_250	Bison_priscus_BS394_NS_37460_890	Bos_indicus_AB268578_0_0	Bubalus_bubalis_EU268909_0_0
Bison_priscus_BS133_F_33800_1900	Bison_priscus_BS395_NS_40700_1300	Bos_indicus_AB268580_0_0	
Bison_priscus_BS145_NS_12270_50	Bison_priscus_BS396_NS_23680_170	Bos_indicus_AY378134_0_0	

Supplementary Table 5. List of published whole mitochondrial genome sequences used for phylogenetic analysis.

American bison	**Cattle**	**Yak**
GU947000_Bison_bison_Plains_Nebraska_0	FJ971080_Bos_Q_Italy_Romagnola_0	KJ704989_Bos_grunniens_ChinaGansu_Gannan_0
GU946976_Bison_bison_Plains_Montana_0	FJ971085_Bos_R_Italy_Cinisara_0	KR011113_Bos_grunniens_ChinaTibet_QinghaiPlateau_0
GU947004_Bison_bison_Plains_Wyoming_0	EU177841_Bos_T1_Italy_chianina_0	KR052524_Bos_grunniens_ChinaTibet_Pali_0
GU947006_Bison_bison_Wood_ElkIsland_0	DQ124383_Bos_T2_Korea_0	KJ463418_Bos_grunniens_ChinaQinghai_Dantong_0
GU946987_Bison_bison_Plains_Montana_0	EU177815_Bos_T3_Italy_piemontese_0	KM233417_Bos_mutus_ChinaTibet_Yakow_0
GU947005_Bison_bison_Wood_ElkIsland_0	DQ124372_Bos_T4_Korea_0	**Buffalo**
GU947002_Bison_bison_Plains_Texas_0	EU177862_Bos_T5_Italy_valdostana_0	GU947003_Bison_bison_Plains_Texas_0
GU947003_Bison_bison_Plains_Texas_0	**Aurochs**	AY488491_Bubalus_bubalis
Wisent	GU985279_Bos_P_England_6760	AY702618_Bubalus_bubalis
JN632602_Bison_bonasus_0	JQ437479_Bos_P_Poland_1500	AF547270_Bubalus_bubalis
HQ223450_Bison_bonasus_0	**Zebu**	
HM045017_Bison_bonasus_Poland_0	FJ971088_Bos_I1_Mongolia_0	
Steppe bison	EU177870_Bos_I2_Iran_0	
KM593920_Bison_priscus_SGE2_France_TroisFreres_19151		

Supplementary Table 6. f4 ratio estimates, f4(A,O,X,C) is the numerator, f4(A,O,B,C) is the denominator.

S6-A. Including heterozygotes

A	O	X	C	:	A	O	B	C	alpha	std.err	Z
AmericanBison	Ovis_aries	AllWisent+CladeX	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.890988	0.025788	34.551
AmericanBison	Ovis_aries	AllWisent+CladeX	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.109012	0.025788	4.227
AmericanBison	Ovis_aries	AllWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.884257	0.02918	30.304
AmericanBison	Ovis_aries	AllWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.115743	0.02918	3.967
AmericanBison	Ovis_aries	CladeX	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.893978	0.022763	39.273
AmericanBison	Ovis_aries	CladeX	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.106022	0.022763	4.658
AmericanBison	Ovis_aries	AncientWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.812638	0.054701	14.856
AmericanBison	Ovis_aries	AncientWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.187362	0.054701	3.425
AmericanBison	Ovis_aries	HistoricalWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.773802	0.032319	23.943
AmericanBison	Ovis_aries	HistoricalWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.226198	0.032319	6.999
AmericanBison	Ovis_aries	ModernWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.899149	0.031184	28.834
AmericanBison	Ovis_aries	ModernWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.100851	0.031184	3.234

S6-B. Haploidisation by randomly sampling an allele at heterozygous sites

A	O	X	C	:	A	O	B	C	alpha	std.err	Z
AmericanBison	Ovis_aries	AllWisent+CladeX	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.894329	0.027147	32.944
AmericanBison	Ovis_aries	AllWisent+CladeX	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.105671	0.027147	3.893
AmericanBison	Ovis_aries	AllWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.88342	0.030518	28.947
AmericanBison	Ovis_aries	AllWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.11658	0.030518	3.82
AmericanBison	Ovis_aries	CladeX	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.912424	0.025204	36.202
AmericanBison	Ovis_aries	CladeX	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.087576	0.025204	3.475
AmericanBison	Ovis_aries	AncientWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.813521	0.059078	13.77
AmericanBison	Ovis_aries	AncientWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.186479	0.059078	3.156
AmericanBison	Ovis_aries	HistoricalWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.786183	0.035363	22.232
AmericanBison	Ovis_aries	HistoricalWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.213817	0.035363	6.046
AmericanBison	Ovis_aries	ModernWisent	Aurochs	:	AmericanBison	Ovis_aries	Steppe	Aurochs	0.899281	0.032252	27.883
AmericanBison	Ovis_aries	ModernWisent	Steppe	:	AmericanBison	Ovis_aries	Aurochs	Steppe	0.100719	0.032252	3.123

Supplementary Table 7: Bootstrap resampling of genotypes for testing topologies using D statistics. The table shows the fraction of bootstrap replicates for which the original result was not recapitulated, from 10000 bootstraps, for 10%, 20%, etc. subsets of the genotypes. A topology is considered to be simple if it either has a non-significant D statistic (see Supplementary Figure 11), or has a D statistic closest to zero with confidence intervals that do not overlap the D statistic for the other two topologies.

Most parsimonious topology	Simple topology	10%	20%	30%	40%	50%	60%	70%	80%	90%
((CladeX, Steppe), ModernWisent)	True	0.0067	0.0001	0.0	0.0	0.0	0.0	0.0	0.0	0.0
((Steppe, HistoricalWisent), ModernWisent)	False	0.0575	0.0573	0.0284	0.0036	0.0005	0.0	0.0	0.0	0.0
((ModernWisent, CladeX), HistoricalWisent)	False	0.1753	0.371	0.485	0.4427	0.3039	0.1564	0.0549	0.0072	0.0

((CladeX, Steppe), HistoricalWisent)	True	0.0182	0.0174	0.0154	0.016	0.0113	0.0072	0.0022	0.0004	0.0
((AncientWisent, HistoricalWisent), ModernWisent)	True	0.0565	0.0152	0.0042	0.0012	0.0	0.0	0.0	0.0	0.0
((Steppe, HistoricalWisent), AncientWisent)	False	0.0151	0.0039	0.0001	0.0002	0.0	0.0	0.0	0.0	0.0
((AncientWisent, Steppe), ModernWisent)	True	0.0484	0.0086	0.0014	0.0002	0.0	0.0	0.0	0.0	0.0
((CladeX, Steppe), AncientWisent)	False	0.0304	0.0142	0.0086	0.0063	0.0033	0.0025	0.0015	0.0001	0.0
((AncientWisent, CladeX), ModernWisent)	True	0.0703	0.0213	0.0062	0.0015	0.0007	0.0	0.0	0.0	0.0
((HistoricalWisent, CladeX), AncientWisent)	False	0.0184	0.0053	0.001	0.0005	0.0	0.0	0.0	0.0	0.0
((ModernWisent, HistoricalWisent), Aurochs)	False	0.0591	0.0031	0.0005	0.0	0.0	0.0	0.0	0.0	0.0
((Aurochs, ModernWisent), CladeX)	False	0.2229	0.2476	0.0824	0.0115	0.0009	0.0	0.0	0.0	0.0
((HistoricalWisent, CladeX), Aurochs)	True	0.0061	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
((Steppe, CladeX), Aurochs)	True	0.0001	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
((Steppe, HistoricalWisent), Aurochs)	True	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
((Steppe, ModernWisent), Aurochs)	False	0.1362	0.0535	0.0048	0.0007	0.0002	0.0	0.0001	0.0	0.0
((Steppe, AncientWisent), Aurochs)	True	0.0441	0.0082	0.0001	0.0001	0.0	0.0	0.0	0.0	0.0
((AncientWisent, CladeX), Aurochs)	True	0.0276	0.0058	0.0004	0.0001	0.0	0.0	0.0	0.0	0.0

253 **Supplementary Table 8:** Hypergeometric test for shared derived steppe alleles. Steppe derived sites
254 were filtered for coverage depth in the wisent lineages 1 and 2, for which the test was performed. In the
255 last row, wisent represents all wisent other than CladeX.

1	2	Steppe	Derived 1	Derived 2	Common	P
Ancient Wisent	CladeX	161	111	133	108	1.72E-12
Ancient Wisent	Historical Wisent	174	115	119	108	1.37E-24
Ancient Wisent	Modern Wisent	178	124	108	95	5.12E-11
CladeX	Historical Wisent	529	448	385	370	3.09E-29
CladeX	Modern Wisent	556	469	350	326	2.79E-13
Historical Wisent	Modern Wisent	618	436	372	342	5.50E-48
Wisent	CladeX	557	357	468	332	4.18E-14

256

Supplementary Table 9: Hypergeometric test for shared derived aurochs alleles. Aurochs derived sites were filtered for coverage depth in the wisent lineages 1 and 2, for which the test was performed. In the last row, wisent represents all wisent other than CladeX.

1	2	Aurochs	Derived 1	Derived 2	Common	P
Ancient Wisent	CladeX	758	20	9	4	4.11E-05
Ancient Wisent	Historical Wisent	822	22	11	8	1.01E-11
Ancient Wisent	Modern Wisent	826	25	22	12	1.49E-14
CladeX	Historical Wisent	2517	36	47	16	7.34E-20
CladeX	Modern Wisent	2580	39	73	15	1.99E-14
Historical Wisent	Modern Wisent	2845	58	83	39	2.66E-50
Wisent	CladeX	2634	93	41	15	1.58E-12

Supplementary Table 10: The weighted sample median $\hat{M}$, the weighted sample mode $\hat{Mo}$, and the prediction error E_{pred}, for each ABC analysis.

Trio	$\hat{M}$	$\hat{Mo}$	E_{pred}
A875, 6A, Aurochs	0.8660	0.9204	0.4534
A3133, 6A, Aurochs	0.8480	0.9172	0.4881
A875, Historical Wisent, Aurochs	0.8636	0.9323	0.4187
A3133, Historical Wisent, Aurochs	0.8646	0.9384	0.4921
All	0.8250	0.9034	0.5111

Supplementary Table 11: Empirical posterior probabilities for levels of hybridisation 1%-5%, for each trio.

Trio	1%	2%	3%	4%	5%
A875, 6A, Aurochs	0.9620	0.9340	0.8720	0.8400	0.8120
A3133, 6A, Aurochs	0.9600	0.9600	0.8840	0.8440	0.7980
A875, Historical Wisent, Aurochs	0.9660	0.9340	0.8860	0.8520	0.7940
A3133, Historical Wisent, Aurochs	0.9580	0.9100	0.8580	0.8080	0.7640
All	0.9720	0.9440	0.9140	0.8760	0.8760

267
268

Supplementary Note 1:

Samples, DNA extraction and sequencing

Samples and radiocarbon dating

For clarity purposes we kept the most commonly used taxonomic nomenclature of bovine throughout the study. Although not yet widely accepted, it has been proposed to sink the genus *Bison* into *Bos* based on the shallow time depth of their evolutionary history [7]. The validity of such genetic separation is further tested in this study.

Samples from a total of 87 putative bison bones were collected from 3 regions across Europe: Urals, Caucasus, and Western Europe (Supplementary Data 1). As shown in the Supplementary Data 1, most of the samples were from bones identified as bison or bovid post-cranial samples, because cranial material is rare for this time period.

The main set of samples, from northeastern Europe, represents isolated bones excavated from a wide variety of cave deposits throughout the Ural Mountains and surrounding areas. These samples are housed at the Zoological Museum of the Institute of Plant and Animal Ecology (ZMIPAE) in Ekaterinburg, Russia.

In southeastern Europe, bovid bone fragments were excavated in Mezmaiskaya Cave in the Caucasus Mountains. Samples were obtained from the Laboratory of Prehistory in St Petersburg. Additional six samples from the Caucasus are identified as Caucasian bison (B. bonasus caucasicus, hereafter referred to as historical wisent): two of them are from the National History Museum (NHM) in London, and four come from hunts in the Kuban Oblast in the early 20th century (one collected by scientist Viktor Iwanovich Worobjew in 1906 and three hunted during the Kuban Hunt under the Grand Duke Sergei Mikhailovich of Russia), currently held at the Zoological Institute of the Russian Academy of Sciences (ZIRAS - Saint Petersburg, Russia). Four additional bones from the Caucasus region comes from the eastern border with Ukraine and are held at the Institute of Archeology (IAKiev), Ukrainian Academy of Sciences, Kiev.

Most western European bones come from late Pleistocene deposits on the North Sea bed. These specimens, now curated by the North Sea Network (NSN) in the Netherlands, were recovered by trawling operations and as such have little stratigraphic information. Specimens were selected on the basis of their morphological similarities with the 'small form' described by Drees and Post [8].

Three bones held in the collections of the Vienna Natural History Museum (VNHM), and three bones held in the Museum National d'Histoire Naturelle (Paris) come from central European Holocene sites.

Finally, one bone comes from the Monti Lessini rock-shelter site Riparo Tagliente in the North of Italy, one bone comes from the Swiss site of Le Gouffre de la combe de la racine in the Jura mountains (Swiss Institute for Speleology and Karst Studies, ISSKA), and one bone comes from l'Aven de l'Arquet in the Gard region of France (Musée de Préhistoire d'Orgnac).

In addition, two samples from the Beringian region were used: one sample, a steppe bison astragalus from the Yukon territory (Canada), has previously been used in a study of cytosine methylation in ancient DNA [9]; and another steppe bison from Alyoshkina Zaimka in Siberia.

All non-contemporaneous samples from which bison mitochondrial control region sequences were successfully amplified were sent for accelerator mass spectrometry (AMS) radiocarbon dating (except for seven samples from level 3 of the Mezmaiskaya cave, which were expected to be older than AMS dating capabilities [10,11]). The dating was performed by the AMS facility at the Oxford Radiocarbon Accelerator Unit at the University of Oxford (OxA numbers), the Eidgenössische Technische Hochschule in Zürich for a Ukrainian sample (ETH number), and the Ångström Laboratory of the University of Uppsala, Sweden, for the Swiss sample (Ua number). The results are shown in Supplementary Data 1, with all dates reported in kcal yr BP unless otherwise stated. The calibration of radiocarbon dates was performed using OxCal v4.1 with the IntCal13 curve [12].
In addition, two bones identified as bison were previously dated at the Centre for Isotope Research, Radiocarbon Laboratory, University of Groningen, Netherlands, with infinite radiocarbon age, consistently with the dating performed at Oxford (A2808-JGAC26=GrA-34533; A2809-JGAC27= GrA-34524).

Ancient DNA extraction

All ancient DNA work was conducted in clean-room facilities at the University of Adelaide's Australian Centre for Ancient DNA, Australia (ACAD), and at the University of Tuebingen, Germany (UT) following published guidelines [13].

University of Adelaide:

Samples were UV irradiated (260 nm) on all surfaces for 30 min. Sample surface was wiped with 3% bleach, then ~1 mm was removed using a Dremel tool and carborundum cutting disks. Each sample was ground to a fine powder using a Mikro-Dismembrator (Sartorius). Two DNA extraction methods were used during the course of the project (see Supplementary Data 1for the method used for specific samples):

- *Phenol-chloroform method*: Ancient DNA was extracted from 0.2-0.5g powdered bone using phenol-chloroform and centrifugal filtration methods according to a previously published method [2].

- *In solution silica based method*: Ancient DNA was extracted from 0.2-0.3g powdered bone according to a previously published method [14].

University of Tuebingen:

Samples were UV-irradiated overnight to remove surface contamination. DNA extraction was performed following a guanidinium-silica based extraction method [15] using 50mg of bone powder. A DNA library was prepared using 20µl of extract for each sample according to [16]. Sample-specific indexes were added to both library adapters to differentiate between individual samples after pooling and multiplex sequencing [17]. Indexed libraries were amplified in 100µl reactions, followed by purification over Qiagen MinElute spin columns (Quiagen, Hilden, Germany).

355 **Sequencing of the mitochondrial control region**

356 A ~600 bp fragment of the mitochondrial control region was amplified in one or up to
357 four overlapping fragments, depending on DNA preservation. PCR amplifications
358 were performed using primers designed for the bovid mitochondrial control region,
359 following the method described in [2].

360 One-step simplex PCR amplifications using Platinum *Taq* Hi-Fidelity polymerase
361 were performed on a heated lid thermal cycler in a final volume of 25 µl containing 1
362 µl of aDNA extract, 1mg/ml rabbit serum albumin fraction V (RSA; Sigma-Aldrich,
363 Sydeny, NSW), 2 mM $MgSO_4$ (Thermo Fisher, Scoresby VIC), 0.6 µM of each
364 primer (Supplementary Table 1), 250 µM of each dNTP (Thermo Fisher), 1.25 U
365 Platinum *Taq* Hi-Fidelity and 1 × Hi-Fidelity PCR buffer (Thermo Fisher). The
366 conditions for PCR amplification were initial denaturation at 95°C for 2 min,
367 followed by 50 cycles of 94°C for 20 sec, 55°C for 20 sec and 68°C for 30 sec, and a
368 final extension at 68°C for 10 min at the end of the 50 cycles.

369 Multiplex primer sets A and B were set up separately (Supplementary Table 1).
370 Multiplex PCR was performed in a final volume of 25 µl containing 2 µl of aDNA
371 extract, 1 mg/ml RSA, 6 mM $MgSO_4$, 0.2 µM of each primer (Supplementary Table
372 1), 500 µM of each dNTP, 2 U Platinum *Taq* Hi-Fidelity and 1 × Hi-Fidelity PCR
373 buffer. Multiplex PCR conditions were initial denaturation at 95°C for 2 min,
374 followed by 35 cycles of 94°C for 15 sec, 55°C for 20 sec and 68°C for 30 sec, and a
375 final extension at 68°C for 10 min at the end of the 35 cycles. Multiplex PCR
376 products were then diluted to 1:10 as template for the second step of simplex PCR.
377 The simplex PCR, using Amplitaq Gold (Thermo Fisher) or Hotmaster™ *Taq* DNA
378 polymerase (5Prime, Milton, Qld), was conducted in a final volume of 25 µl
379 containing 1 µl of diluted multiplex PCR product, 2.5 mM $MgCl_2$, 0.4 µM of each
380 primer (Supplementary Table 1), 200 µM of each dNTP, 1 U Amplitaq
381 Gold/Hotmaster *Taq* polymerase and 1 × PCR buffer. The PCR conditions were initial
382 denaturation at 95°C for 2 min, followed by 35 cycles of 94°C for 20 sec, 55°C for 15
383 sec and 72°C for 30 sec, and a final extension at 72°C for 10 min at the end of the 35
384 cycles. Multiple PCR fragments were cloned to evaluate the extent of DNA damage
385 and within-PCR template diversity.

386 PCR products were then checked by electrophoresis on 3.5-4.0% agarose TBE gels,
387 and visualized after ethidium bromide staining on a UV transilluminator. PCR
388 amplicons were purified using Agencourt® AMPure magnetic beads (Beckman
389 Coulter, Lane Cove, NSW) according to the manufacturer's instructions. Negative
390 extraction controls and non-template PCR controls were used in all experiments.

391 All purified PCR products were bi-directionally sequenced with the ABI Prism®
392 BigDye™ Terminator Cycle Sequencing Kit version 3.1 (Thermo Fisher). The
393 sequencing reactions were performed in a final volume of 10 µl containing 3.2 pmol
394 of primer (Supplementary Table 1), 0.25 µl Bigdye terminator premixture, and 1.875
395 µl of 5 × sequencing buffer. The reaction conditions included initial denaturation at
396 95°C for 2 min, 25 cycles with 95°C for 10 sec, 55°C for 15 sec, and 60°C for 2 min
397 30 sec. Sequencing products were purified using Agencourt® Cleanseq magnetic
398 beads (Beckman Coulter) according to the manufacturer's protocol. All sequencing
399 reactions were analysed on an ABI 3130 DNA capillary sequencer (Thermo Fisher).

400 Mitochondrial control region sequences (>400bp) were successfully amplified from
401 65 out of 87 analysed samples. Three samples produced a mixture of cattle and bison

402 amplification products; these were identified as contaminated and removed from all
403 analyses. Sequences from two individuals did not match bovid haplotypes and were
404 identified as brown bear and elk in BLAST searches (see Supplementary Data 1). This
405 is presumably due to the source postcranial elements being morphologically
406 ambiguous and misidentified.

407

408 **Sequencing of the whole mitochondrial genome**

409 To provide deeper phylogenetic resolution and further examine the apparent close
410 relationship between *Bos* and wisent mitochondria, full mitogenome sequences of 13
411 CladeX specimens, as well as one ancient wisent, one historical wisent, and one
412 steppe bison were generated using hybridisation capture with RNA probes.

413

414 *Samples A001, A004, A018, A4089 (CladeX)*

415 *DNA library preparation*

416 DNA repair and polishing were performed in a reaction that contained 20 µl DNA
417 extract, 1x NEB Buffer 2 (New England Biolabs, Ipswich, MA), 3U USER enzyme
418 cocktail (New England Biolabs), 20U T4 polynucleotide kinase (New England
419 Biolabs), 1mM ATP, 0.1 mM dNTPs (New England Biolabs), 8 µg RSA, and H_2O to
420 38.5 µl. The reaction was incubated at 37°C for 3 hours then 4.5U of T4 DNA
421 polymerase (New England Biolabs) was added and the reaction incubated at 25°C for
422 a further 30 min. Double-stranded libraries were then built with truncated Illumina
423 adapters containing dual 5-mer internal barcodes as in [16].

424

425 *Amplification of Bos taurus mitochondrial in vitro transcription (IVT) templates*

426 RNA probes were generated from long-range PCR products of *Bos taurus*
427 mitochondrial DNA. The NCBI Primer-Blast program
428 (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) was used to design primers to
429 amplify the *Bos taurus* mitochondrial genome (NC_006853.1) in three overlapping
430 sections: mito-1 (6568 bp), mito-2 (6467 bp), and mito-3 (5390 bp). Primer pairs
431 were designed with a high melting temperature to permit amplification with 2-stage
432 PCR and the T7 RNA promoter was attached to the 5' end of one primer from each
433 pair [18](Supplementary Table 1). Amplification of each mitochondrial section was
434 performed using a heated lid thermal cycler in multiple PCRs containing 1x Phire
435 Buffer (Thermo Fisher), 25 ng calf thymus DNA (Affymetrix, Santa Clara, CA), 200
436 µM dNTPs , 500 nM forward and reverse primers, 0.5 µl Phire Hot Start II DNA
437 polymerase (Thermo Fisher), and H_2O to 25 µl. The mito-1 and mito-2 sections were
438 amplified with a thermal cycler program of 1 cycle: 98°C for 30 sec; 26 cycles: 98°C
439 for 10 sec and 72°C for 70 sec; and 1 cycle: 72°C for 180 sec whilst the program for
440 mito-3 was 1 cycle: 98°C for 30 sec, 28 cycles: 98°C for 10 sec and 72°C for 60 sec,
441 and 1 cycle: 72°C for 180 sec. After amplification, 2 l of each PCR was agarose gel
442 electrophoresed and the product visualized with Gel-Red (Biotium, Hayward, CA)
443 staining and UV illumination. Amplification of mito-1 and mito-2 produced a single
444 band and the PCRs for these mitochondrial sections were separately pooled and then
445 purified with QiaQuick columns (Qiagen, Chadstone Centre, VIC) following the
446 provided PCR cleanup protocol. Amplification of mito-3 produced unwanted
447 products and the correct size amplicon was size selected using gel excision followed

448 by purification with QiaQuick columns using the gel extraction protocol. Purified
449 amplicons from each mitochondrial section were quantified using a NanoDrop 2000
450 Spectrophotometer (Thermo Fisher).

451

452 *Transcription of Bos taurus mitochondrial IVT templates*

453 Each of the three mitochondrial IVT templates were transcribed using a T7 High
454 Yield RNA Synthesis Kit (New England Biolabs) in multiple reactions containing
455 150-200 ng purified amplicon, 1x Reaction Buffer, 10 mM rNTPs, 2 µl T7 enzyme
456 mix, and H_2O to 20 µl. The IVT reactions were incubated for 16 hours at 37°C and
457 then the DNA template was destroyed by incubating for an additional 15 min at 37°C
458 with 2U Turbo Dnase (Thermo Fisher). IVT reactions for each mitochondrial section
459 were separately pooled and purified with Megaclear spin columns (Thermo Fisher)
460 except that H_2O was used to elute the RNA instead of the provided elution buffer. The
461 elution buffer provided with the Megaclear kit was found to inhibit fragmentation in
462 the next step. Integrity of the RNA was verified on an acrylamide gel and the mass
463 quantified with a Nanodrop 2000 Spectrophotometer.

464

465 *Fragmentation of mitochondrial IVT RNA*

466 RNAs from the IVT transcription were fragmented with a NEBNext Magnesium
467 RNA Fragmentation Module (New England Biolabs) in reactions that contained 1x
468 Fragmentation buffer, 45 µg RNA, and H_2O to 20 µl. Reactions were incubated at
469 94°C for 10 min and fragmentation stopped with the addition of 2 µl Stop Buffer.
470 After fragmentation, each reaction was purified with a RNeasy MinElute spin column
471 (Qiagen) by following the provided cleanup protocol except for the final elution. To
472 elute, 20 µL H_2O was pipetted into the column and the column was heated at 65°C for
473 5 min and then centrifuged at 15,000 g for 1 min. The flow-through was transferred
474 to a 1.5 ml tube and stored at -80°C. The fragmented RNA was quantified on a
475 NanoDrop 2000 Spectrophotometer and 100 ng was visualized on an acrylamide gel
476 producing a smear in the range of 80-300 bases.
477

478 *Biotinylation of fragmented RNA*

479 Biotinylation was performed in several reactions containing 6.7 µg each of mito-1,
480 mito-2, and mito-3 fragmented RNA, 40 µl Photoprobe Long Arm (Vector
481 Laboratories, Burlingame, CA), and H_2O to 80 µl in 200 µl PCR tubes. The tubes
482 were placed in a 4°C gel cooling rack and then incubated under the bulb of a UV
483 sterilization cabinet for 30 min. Organic extractions were performed on the labelling
484 reactions by adding 64 µl H_2O, 16 µl 1 M Tris buffer, and 160 µl sec-butanol to each
485 tube and shaking vigorously for 30 sec followed by centrifugation for 1 minute at
486 1000 g. The upper organic layers were discarded and the extraction repeated with an
487 additional 160 µl sec-butanol. After the second organic layers were discarded, the
488 remaining aqueous phases were purified with RNeasy MinElute spin columns
489 following the provided reaction cleanup protocol but with a modified elution
490 procedure described in the previous step. Elutions with similar RNA were pooled and
491 then quantified with a NanoDrop Spectrophotometer 2000 and the RNA, which will
492 now be called probe, was stored at -80°C in 5 µl aliquots at 100 ng/µl.

493

494 *Repetitive sequence blocking RNA*

495 RNA to block repetitive sequences in bison aDNA was transcribed from Bovine
496 HyBlock™ DNA (i.e. Cot-1 DNA, Applied Genetics Laboratories Inc., Melbourne,
497 FL) using a published linear amplification protocol [19]. Briefly, the HyBlock DNA
498 was polished in a reaction containing T4 polynucleotide kinase and T4 DNA
499 polymerase and purified with MinElute spin columns following the PCR cleanup
500 protocol provided. Tailing was performed on the polished DNA with terminal
501 transferase and a tailing solution containing 92 µM dTTP (Thermo Fisher) and 8 µM
502 ddCTP (Affymetrix). After tailing, the Hybloc DNA was purified with MinElute spin
503 columns as before. The HyBlock DNA was then heat denatured and the T7-A18B
504 primer (Supplementary Table 1), containing the T7 RNA polymerase promoter, was
505 allowed to anneal to the poly-T tail with slow cooling. A second-strand synthesis
506 reaction was then performed on the HyBlock DNA using DNA polymerase I Klenow
507 fragment (New England Biolabs) and the product was purified with MinElute spin
508 columns. The double stranded HyBlock DNA was transcribed using a T7 High Yield
509 RNA Synthesis Kit in multiple reactions containing 75 ng DNA, 1x Reaction Buffer,
510 10 mM rNTPs, 2 µl T7 enzyme mix, and H_2O to 20 µl. IVT reactions were incubated
511 for 16 hours at 37°C and then the DNA template was destroyed by adding 2U Turbo
512 Dnase and incubating for an additional 15 min at 37°C. The RNA was purified with
513 RNeasy MinElute spin columns as above. Purified RNA was quantified on a
514 NanoDrop 2000 and 100 ng visualized on an acrylamide gel, which produced a smear
515 80 to 500 bp in length.

516

517 *Primary mitochondrial hybridisation capture*

518 Truncated versions of the Illumina adapters were used for hybridisation capture
519 because full-length adapters reduce enrichment efficiency [20]. For the primary
520 hybridisation capture, three Reagent Tubes were prepared for each bison library with
521 the following materials: Reagent Tube #1- 3.5 µl of 35-55 ng/µl DNA library;
522 Reagent Tube #2- 5 µl probes, 1 µl HyBlock RNA, and 0.5 µl of 50 µM P5/P7 RNA
523 blocking oligonucleotides (Supplementary Table 1); Reagent Tube #3- 30 µl
524 Hybridisation Buffer [21]: 75% formamide (Thermo Fisher), 75 mM HEPES , pH 7.3, 3
525 mM EDTA (Thermo Fisher), 0.3% SDS (Thermo Fisher), and 1.2 M NaCl (Thermo
526 Fisher). Hybridisation capture was performed in a heated lid thermal cycler
527 programmed as follows: Step 1- 94°C for 2 min, Step 2- 65°C for 3 min, Step 3- 42°C
528 for 2 min, Hold 4- 42°C hold. To start hybridisation capture, Reagent Tubes were
529 placed in the thermal cycler at the start of each program Step in the following order:
530 Step 1- Reagent Tube #1; Step 2- Reagent Tube #2; Step 3- Reagent Tube #3. For
531 each library, once the Hold cycle started 20 µl of hybridisation buffer from Reagent
532 Tube #3 was mixed with the RNA in Reagent Tube #2. The entire content of Reagent
533 Tube #2 was then pipetted into Reagent Tube #1 and mixed with the bison library to
534 begin the hybridisation capture. Hybridisation capture was carried out at 42°C for 48
535 hours.
536 Magnetic streptavidin beads (New England Biolabs) were washed just prior to the end
537 of the hybridisation capture incubation. For each library, 50 µl of beads were washed
538 twice using 0.5 ml Wash Buffer 1(2X SSC+0.05% Tween-20, all reagents Thermo
539 Fisher) and a magnetic rack. We also saturated all magnetic bead sites that could
540 potentially bind nucleic acid in a non-specific fashion using yeast tRNA, to optimise
541 the expected and specific streptavidin-biotin binding. Briefly, the beads were blocked

542 by incubation in 0.5 ml Wash Buffer 1+ 100 µg yeast tRNA (Thermo Fisher) for 30
543 min on a rotor. Blocked beads were washed once as before and then suspended in 0.5
544 ml Wash Buffer. At the end of the hybridisation capture, each reaction was added to a
545 tube of blocked beads and incubated at room temperature for 30 min on a rotor. The
546 beads were then taken through a series of stringency washes as follows: Wash 1 - 0.5
547 ml Wash Buffer 1 at room temperature for 10 min; Wash 2 - 0.5 ml Wash Buffer 2
548 (0.75X SSC + 0.05% Tween-20) at 50°C for 10 min; Wash 3 - 0.5 ml Wash Buffer 2
549 at 50°C for 10 min; Wash 4 - 0.5 ml Wash Buffer 3 (0.2X SSC + 0.05% Tween-20) at
550 50°C for 10 min. After the last wash, the captured libraries were released from the
551 probe by suspending the beads in 50 µl of Release buffer (0.1 M NaOH, Sigma
552 Aldrich) and incubating at room temperature for 10 min. The Release buffer was then
553 neutralized with the addition of 70 µl Neutralization buffer (1 M Tris-HCl pH 7.5,
554 Thermo Fisher). Captured libraries were then purified with MinElute columns by first
555 adding 650 µl PB buffer and 10 µl 3 M sodium acetate to adjust the pH for efficient
556 DNA binding. Libraries were purified using the provided PCR cleanup protocol and
557 eluting with 35 µl EB+0.05% Tween-20.

558

559 *Primary hybridisation capture amplification*

560 Amplification of each primary hybridisation capture was performed in five PCRs
561 containing 5 µl of primary captured library, 1X Phusion HF buffer (Thermo Fisher),
562 200 µM dNTPs, 200 µM each of primers IS7_short_amp.P5 and IS8_short_amp.P7
563 (Supplementary Table 1), 0.25 U Phusion Hot Start II DNA polymerase (Thermo
564 Fisher), and H_2O to 25 µl. The five PCR products were pooled and DNA was purified
565 using AMPure magnetic beads.
566

567 *Secondary mitochondrial hybridisation capture*

568 Amplified primary libraries were taken through a second round of hybridisation
569 capture using the same procedure as describe in *Primary mitochondrial hybridisation*
570 *capture* step.

571

572 *Secondary hybridisation capture amplification*

573 Indexed primers were used to convert the DNA from the secondary hybridisation
574 capture to full length Illumina sequencing libraries. Each library was amplified in
575 three PCRs containing 5 µl secondary hybridisation capture library, 1X Phusion HF
576 buffer, 200 µM dNTPs, 200 µM each of primers GAII_Indexing_*x* (library specific
577 index) and IS4 (Supplementary Table 1), 0.25 U Phusion Hot Start II DNA
578 polymerase, and H_2O to 25 µl. Amplification was performed in a heated lid thermal
579 cycler programmed as follows 1 cycle: 98°C for 30 sec; 10 cycles: 98°C for 10 sec,
580 60°C for 20 sec, 72°C for 20 sec; and 1 cycle: 72°C for 180 sec. The five PCR
581 products were pooled and DNA was purified using AMPure magnetic beads.

582

583 <u>Samples A003, A005, A006, A007, A017, A15526, A15637, A15668 (CladeX),</u>
584 <u>A4093 (*ancient wisent*) and A15654 (*historical wisent*)</u>

585 *DNA library preparation*

586 Double-stranded Illumina libraries were built from 20 µl of each DNA extract using

587 partial UDG treatment [22] and truncated Illumina adapters with dual 7-mer internal
588 barcodes, following the protocol from [23].

589

590 *Hybridisation capture*

591 Commercially synthesised biotinylated 80-mer RNA baits (MYcroarray, MI, USA)
592 were used to enrich the target library for mitochondrial DNA. Baits were designed as
593 part of the commercial service using published mitochondrial sequences from 24
594 placental mammals, including *Bison bison* and *Bos taurus*.

595 One round of hybridisation capture was performed according to the manufacturer's
596 protocol (MYbaits v2 manual) with modifications. We used P5/P7 RNA blocking
597 oligonucleotides (Supplementary Table 1) instead of the blocking oligonucleotides
598 provided with the kit. We also incubated the magnetic beads with yeast tRNA to
599 saturate all potential non-specific sites on the magnetic beads that could bind nucleic
600 acids and increase the recovery of non-specific DNA and therefore decrease the final
601 DNA yield.

602 Indexed primers were used to convert the capture DNA to full length Illumina
603 sequencing libraries. Each library was amplified in eight PCRs containing 5 µl
604 hybridisation capture library, 1x Gold Buffer II, 2.5mM $MgCl_2$, 200 µM dNTPs, 200
605 µM each of primers GAII_Indexing_*x* (library specific index) and IS4
606 (Supplementary Table 1), 1.25 U Amplitaq Gold DNA polymerase, and H_2O to 25 µl.
607 Amplification was performed in a heated lid thermal cycler programed as follows 1
608 cycle: 94°C for 6 min; 15 cycles: 98°C for 30 sec, 60°C for 30 sec, 72°C for 40 sec;
609 and 1 cycle: 72°C for 180 sec. The PCR products were pooled and DNA was purified
610 using AMPure magnetic beads (Agencourt®, Beckman Coulter).

611

612 *Samples LE237, LE242 and LE257 (CladeX)*

613 Target DNA enrichment was performed by capture of the pooled libraries using DNA
614 baits generated from bison (*Bison bison*) mitochondrial DNA [24]. The baits were
615 generated using three primer sets (Supplementary Table 1, f) designed with the
616 Primer3Plus software package [25]. All extractions and pre-amplification steps of the
617 library preparation were performed in clean room facilities and negative controls were
618 included for each reaction.

619

620 *Sample A3133 (steppe bison)*

621 DNA repair and polishing were performed in a reaction that contained 20 µl bison
622 A3133 extract, 1x NEB Buffer 2, 3U USER enzyme cocktail, 20U T4 polynucleotide
623 kinase, 1mM ATP, 0.1 mM dNTPs, 8 µg RSA, and H_2O to 38.5 µl. The reaction was
624 incubated at 37°C for 3 hours then 4.5U of T4 DNA polymerase was added and the
625 reaction incubated at 25°C for a further 30 min. Double-stranded libraries were then
626 built with truncated Illumina adapters containing dual 5-mer internal barcodes as in [16]
627 with the final amplification with indexed primers using Phusion Hot Start II DNA
628 polymerase to obtain full length Illumina sequencing libraries.

629

630 **Nuclear locus capture**

631 Genome-wide nuclear locus capture was attempted on DNA repaired libraries of 13
632 bison samples (as described above - see Supplementary Supplementary Table 2). Two
633 different sets of probe were used (as described below), but ultimately, only the 9908
634 loci common to both sets were used for comparative analysis (see nuclear locus
635 analysis section).

636

637 Probe sets

638 *40k SNP probe set*

639 This probe set was originally designed to enrich 39,294 of the 54,609 BovineSNP50
640 v2 BeadChip (Illumina) bovine single nucleotide polymorphism (SNP) loci used in a
641 previous phylogenetic study [26], allowing for a direct comparison of the newly
642 generated data to published genotypes. The discrepancy in the number of surveyed
643 targets was due to manufacturing constraints, as the flanking sequences surrounding
644 certain bovine SNP were too degenerate for synthesis with the MyBaits technology.
645 Probes (MYcroarray, Ann Arbor, MI) were 121-mer long, centred on the targeted
646 bovine SNP and with no tiling, as per the original design of the BovineSNP50 v2
647 BeadChip [27].

648 The BovineSNP50 v2 BeadChip assay targets SNPs that are variable in *Bos taurus* in
649 order to genotype members of cattle breeds. Consequently, SNPs are heavily
650 ascertained to be common in cattle, and their use in phylogenetic studies of other
651 bovid species results in levels of heterozygosity that decrease rapidly with increased
652 genetic distance between cattle and the species of interest. Decker et al. (2009) found
653 the average minor allele frequency in plains bison and wood bison for the 40,843
654 bovine SNPs used in the phylogenetic analysis was 0.014 and 0.009, respectively.
655 Average minor allele frequencies ranged from 0.139 to 0.229 in breeds of taurine
656 cattle.

657

658 *10k SNP probe set*

659 A second set of probes was ordered from MyBaits that targeted a 9,908 locus subset
660 of the previous 39,294 bovine SNPs selected for enrichment. This smaller subset was
661 chosen to minimise ascertainment bias during phylogenetic and population analyses
662 based on their polymorphism within the diversity of available modern genotypes of
663 bison (American and European), Yak, Gaur and Banteng (total of 72 individuals). All
664 of these taxa belong to a monophyletic clade, outside of the cattle diversity, and are
665 consequently all equidistant from the cattle breeds that were used to ascertain the SNP
666 [27], therefore reducing the impact of ascertainment bias when conducting comparisons
667 within the clade. The exclusion of monomorphic sites across specie allows focusing
668 the capture on loci that are more likely to be phylogenetically informative within the
669 bison diversity. Furthermore, singleton sites (only variable for one modern individual,
670 and therefore not informative for the modern phylogeny) were retained on the
671 principle that they might capture some of the unknown ancient diversity of bison
672 when genotyping ancient individuals.

673 We designed 70-mer probes, and this short length, as well as the limited number of
674 targets, allowed for a tiling of 4 different probes for each targeted locus, within the
675 same MYcroarray custom kit of 40,000 unique probes. Among all potential 70-mer

676 sequences within the original 121-mer probe sequence set, only those containing the
677 targeted bovine SNP no fewer than 10 nucleotides from either end were retained as
678 potential probes. Four probes were then designed using the following criteria: i)
679 Estimated melting temperature closest to the average from the 40k SNP probe set; ii)
680 Optimum proportion of guanine based on the efficiency of the 40k SNP probe set; iii)
681 No two probes can be closer than 7 nucleotides from one another; iv) All 'GGGG'
682 and 'CTGGAG' motifs were modified to 'GTGT' and 'CTGTAG', respectively. The
683 former change was incorporated on the recommendation from MyBaits to avoid poly
684 G stretches because their synthesis technology has difficulty with this type of motif
685 and the latter variation was included to remove a restriction site that will be used in a
686 future protocol to produce these probes from an immortalized DNA oligo library [28].

687

688 DNA library preparation

689 All DNA libraries were used for capture of both the mitochondrial genome and
690 genome-wide nuclear loci. See Supplementary Information "Whole mitochondrial
691 genome sequencing" for protocols.

692

693 Hybridisation capture

694 One round of hybridisation capture was performed according to the manufacturer's
695 protocol (MYbaits v2 manual) with modifications. We used P5/P7 RNA blocking
696 oligonucleotides (Supplementary Table 1) instead of the blocking oligonucleotides
697 provided with the kit. We also incubated the magnetic beads with yeast tRNA (see
698 above) to saturate all potential non-specific sites on the magnetic beads that could
699 bind nucleic acids and increase the recovery of non-specific DNA.

700 Indexed primers were used to convert the capture DNA to full length Illumina
701 sequencing libraries. Each library was amplified in eight PCRs containing 5 µl
702 hybridisation capture library, 1C Gold Buffer II, 2.5mM $MgCl_2$, 200 µM dNTPs, 200
703 µM each of primers GAII_Indexing_*x* (library specific index) and IS4
704 (Supplementary Table 1), 1.25 U Amplitaq Gold DNA polymerase, and H_2O to 25 µl.
705 Amplification was performed in a heated lid thermal cycler programed as follows 1
706 cycle: 94°C for 6 min; 15 cycles: 98°C for 30 sec, 60°C for 30 sec, 72°C for 40 sec;
707 and 1 cycle: 72°C for 180 sec. The PCR products were pooled and DNA was purified
708 using AMPure magnetic beads.

709

710 **NGS and data processing**

711 *Whole mitochondrial genomes*

712 All libraries enriched for the mitochondrial genome were sequenced in paired-end
713 reactions on Illumina machines (HiSeq 2500 for LE237A, LE242B and LE247B –
714 MiSeq for the rest), except for A017 and A15526 from which the final concentration
715 of DNA obtained after capture was insufficient for sequencing. The mitochondrial
716 genome of the steppe bison A3133 was recovered from shotgun sequencing on an
717 Illumina HiSeq, performed in the context of another study (see Supplementary Table
718 3).

719 All NGS reads were processed using the pipeline Paleomix v1.0.1[29]. AdapterRemoval
720 v2[30] was used to trim adapter sequences, merge the paired reads, and eliminate all

721 reads shorter than 25 bp. BWA v0.6.2[31] was then used to map the processed reads to
722 the reference mitochondrial genome of the wisent (NC_014044) or the American
723 bison (NC_012346, only for the steppe bison A3133). Minimum mapping quality was
724 set at 25, seeding was disabled and the maximum number or fraction of gap opens
725 was set to 2.

726

727 MapDamage v2[32] was used to check that the expected contextual mapping and
728 damage patterns were observed for each library, depending on the enzymatic
729 treatment used during library preparation (see Supplementary Table 3 and Figures S1-
730 3 for examples), and re-scale base qualities for the non-repaired libraries.

731 Finally nucleotides at the position of the bovine SNP were called using samtools and
732 bcftools, setting the minimum base quality at 30 and the minimum depth of coverage
733 at 2. Consensus sequences were then generated using the Paleomix script
734 vcf_to_fasta.

735

736 *Nuclear*

737 Nuclear DNA from historical (historical wisent: A15654) and ancient (ancient wisent:
738 A4093; CladeX: A15526, A001, A003, A004, A005, A006, A007, A017, A018;
739 steppe: A3133, A875) samples, containing HiSeq data (A3133 and A875) and MiSeq
740 data (all samples), was processed using Paleomix v1.0.1[29] to map reads against the
741 *Bos taurus* reference UMD 3.1[33]. Paleomix was configured to use BWA v0.6.2[31] for
742 mapping, with seeding disabled and -n 0.01 -o 2 (see Supplementary Table 2).
743 MapDamage v2[32] was used to check that the expected contextual mapping and
744 damage patterns were observed for each library, and empirically re-scale base
745 qualities at the end of the fragments.

746 Variants were called using the consensus caller of samtools/bcftools v1.2[34] limiting
747 calls to the 9908 capture sites. Variant calls with a QUAL value lower than 25 were
748 removed. The genotypes for historical and ancient samples were merged with
749 previously published extant bovid 40k capture data[26], and *Bos primigenius* (aurochs)
750 sample CPC98[35]. Only genotypes for the 9908 loci common among all data were
751 retained.
752

753 **Supplementary Note 2:**

754 **DNA analyses**

755

756 **Phylogenetic analysis**

757 *Mitochondrial control region phylogeny*

758 The 60 newly sequenced bovid mitochondrial regions (Supplementary Data 1) were
759 manually aligned, using SeaView v4.3.5[36]. These sequences were aligned with 302
760 published sequences (Supplementary Table 4) representing the following bovid
761 mitochondrial lineages: European bison or wisent (*Bison bonasus*), American bison
762 (*Bison bison*), steppe bison (*Bison priscus*), zebu (*Bos indicus*), and cattle (*Bos*
763 *taurus*). Among these published sequences, 5 were from steppe bison collected in the
764 Urals (Shapiro et al. 2004, Supplementary Data 1).

765 The TN93+G6 model of nucleotide substitution was selected by comparison of
766 Bayesian information criterion (BIC) scores in ModelGenerator v0.85[37]. A
767 phylogenetic tree was then inferred using both maximum-likelihood and Bayesian
768 methods (Figure 2A). Bayesian analyses were performed using the program MrBayes
769 v3.2.3[38]. Posterior estimates of parameters were obtained by Markov chain Monte
770 Carlo sampling with samples drawn every 1000 steps. We used 2 runs, each of four
771 Markov chains, comprising one cold and three heated chains, each of 10 million steps.
772 The first 50% of samples were discarded as burn-in before the majority-rule
773 consensus tree was calculated. A maximum-likelihood analysis was performed with
774 the program PhyML v3[39], using both NNI and SPR rearrangements to search for the
775 tree topology and using approximate likelihood-ratio tests to establish the statistical
776 support of internal branches. Complete phylogenies inferred using both methods are
777 shown in Supplementary Figure 4.

778 *Whole mitochondrial genome phylogeny*

779 The 16 newly sequenced bison whole mitochondrial genomes (Supplementary Data 1)
780 were aligned with 31 published sequences (Supplementary Table 5) representing the
781 following bovid mitochondrial lineages: 3 wisent (*Bison bonasus*), 8 American bison
782 (*Bison bison*), 1 steppe bison (*Bison priscus*), 5 yaks (*Bos grunniens – Bos mutus*), 2
783 zebus (*Bos indicus*), 7 cattle (*Bos taurus*), 2 aurochsen (*Bos primigenius*), and 4
784 buffalo (*Bubalus bubalis*).

785 We used the same methods as described above for the control region to align and
786 estimate the phylogeny. The HKY+G6 model of nucleotide substitution was selected
787 through comparison of BIC scores (Figures 2B and S5).

788 *Estimation of evolutionary timescale*

789 To estimate the evolutionary timescale, we used the program BEAST v1.8.1[40] to
790 conduct a Bayesian phylogenetic analysis of all radiocarbon-dated samples from
791 CladeX and wisent (Figure 1C). The GMRF skyride model[41] was used to account for
792 the complex population history, and a strict clock was assumed. We found support for
793 a strict molecular clock based on replicate analyses using a relaxed uncorrelated
794 lognormal clock[42], which could not reject the strict clock assumption.

795 Mean calibrated radiocarbon dates associated with the sequences were used as
796 calibration points. Some samples appear to be older than 55 ky: one from the Urals,
797 four from the North Sea and five from the Caucasus (Supplementary Data 1). Because

798 these dates have effectively infinite radiocarbon error margins, we allowed them to
799 vary in the analysis by treating them as distinct parameters to be estimated in the
800 model[43]. The dated samples from Mezmaiskaya Cave are from stratigraphic layers
801 2B4 and 2B3, which lie atop of layer 3. All these lower Middle Palaeolithic layers at
802 Mezmaiskaya have 14C results beyond the radiocarbon limit, reflected in the
803 predominance of greater-than or near-background limit ages[11], and therefore are
804 consistent with the electron spin resonance (ESR) chronology for these levels[10], which
805 suggests mean ages in the range from 53 to 73 ky BP (including error margins).
806 Consequently, for each Caucasian sample, we specified a lognormal prior age
807 distribution (mean=8,000) with an offset of 50 ky and with 95% of the prior
808 probability less than 80 ky. A similar prior distribution (mean=26,000) was used for
809 the five remaining samples that had infinite radiocarbon dates, with a 95% prior
810 probability less than 150 ky. Based on the results of all four phylogenetic analyses
811 described above, which showed strong support for the reciprocal monophyly of
812 CladeX and wisent when outgroups were included, this monophyly was constrained
813 for the BEAST runs.

814 All parameters showed sufficient sampling (indicated by effective sample sizes above
815 200) after 5,000,000 steps, with the first 10% of samples discarded as burn-in. In
816 addition, a date-randomization test was conducted to check whether the temporal
817 signal from the radiocarbon dates associated with the ancient sequences was sufficient
818 to calibrate the analysis[44]. This test randomizes all dates and determines whether the
819 95% high posterior density (HPD) intervals of the rates estimated from the date-
820 randomized data sets include the mean rate estimated from the original data set
821 (Supplementary Figure 6).
822
823
824 The time to the most recent common ancestor (tMRCA) between wisent and
825 CladeX mitochondrial lineages was estimated at 121.6 kyr (92.1 – 152.3) (Figure 2C).
826 The tMRCAs for the two lineages was inferred to be 69.3 kyr (53.4 – 89.4) for wisent
827 and 114.9 kyr (89.2 – 143.1) for CladeX. Furthermore, there is some
828 phylogeographical structure within CladeX, with all individuals from the North Sea
829 forming a basal group, which existed before the population replacement with steppe
830 bison, but complete mixture of genetic diversity between all locations after re-
831 colonization. In addition, the tMRCA of the MIS 3 diversity of CladeX was estimated
832 to be about 53.1 kyr (41.5 – 67.5). This date closely matches the ages of the last
833 observed MIS 4 CladeX individuals across all sampled locations, supporting the idea
834 of a population movement and contraction of wisent individuals towards a refugium
835 during the warmer period of MIS 3 in Europe.

836

837 *Nuclear phylogeny from bovine SNP locus data*

838 Phylogenetic trees were inferred from nuclear locus data (see next section for
839 information about the data sets). First, a phylogenetic tree of modern representatives
840 of bovid species, and with sheep as an outgroup, was inferred from published 40,843
841 data[26] (Supplementary Figure 7). Using RAxML v8.1.21[45], the three characters
842 (genotype states AA, AB and BB) from the BovineSNP50 chip were considered as
843 different states in an explicit analogue of the General Time Reversible (GTR)
844 substitution model, with separate substitution parameters for the three possible
845 transformations. For all analyses, 20 maximum likelihood searches were conducted to

find the best tree, and branch support was estimated with 500 bootstrap replicates using the rapid bootstrapping algorithm[46].

This species tree, estimated from genome-wide nuclear locus data, shows that the extant bison species (wisent and American bison) are sister taxa, contrary to the phylogenetic signal from the maternally inherited mitochondrial genome. This topology also clearly shows the paraphyletic status of the genus *Bos* (banteng, gaur, yak, zebu and cattle), as it also includes the genus *Bison* (wisent and American bison).

Using the same method, we reconstructed the phylogeny of bison with the inclusion of five pre-modern samples (for which the highest number of nuclear loci were called amongst the ~10k nuclear bovine SNPs). When only the two steppe bison specimens are included they form a sister-lineage to modern American bison (Supplementary Figure 8A). Similarly, when the steppe bison and pre-modern wisent (including ancient, historical and CladeX) are included, all five pre-modern specimens form a clade most closely related to American bison (Supplementary Figure 8C). However, when only the pre-modern wisent is included, the three specimens (ancient, historical and CladeX) form a clade that is most closely related to modern wisent (Supplementary Figure 8B). These conflicting results reflect the complex non-tree like relationships among the modern and pre-modern taxa, and are consistent with the hybridisation origin of wisent/CladeX and the severe bottleneck in the recent history of the wisent. Hence, we used population genomics statistics to study this nuclear locus dataset (see next section). Finally, these topologies are robust to the removal of transitions (see Supplementary Figure 8D), a minimum depth of 2 for variant calling, and haploidisation (data not shown).

Genome wide nuclear locus analysis

Captured nuclear loci corresponding to bovine SNPs for ancient samples were analysed with published genotypes from modern populations: 20 American bison were selected on the criterion that they do not display any detectable signal of recent introgression from cattle (unpublished data); 2 Yak (*Bos gruniens*); 10 water buffalo (*Bubalus bubalis*); and 10 Sheep (*Ovis aries*). Additionally, 7 modern wisent were selected (among 50 sequenced – [47]) as non-related individuals on a known five-generation pedigree (as shown in Supplementary Figure 9).

Principal Component Analysis

PCA (Figures 3A and S10) was performed using EIGENSOFT version 6.0.1 [48]. In Figure 3A, CladeX sample A006 was used as the representative of CladeX, as this sample contained the most complete set of nuclear loci called at the bovine SNP loci (see Supplementary Table 2). Other CladeX individuals, as well as ancient wisent, cluster towards coordinates 0.0, 0.0 (see Supplementary Figure 10), most likely due to missing data.

Topology testing with the D statistic

For three bison populations, assuming two bifurcations and no hybridisations, there are three possible phylogenetic topologies. For this simple case, the D statistic is expected to be significantly different from zero for exactly two of the three topologies, and not significantly different from zero for the most parsimonious topology. We therefore calculate a D statistic [49] for each of these three topologies, using the sheep (*Ovis aries*) as an outgroup.

When D statistics for the set of three topologies do not indicate zero for one topology and non-zero for the other two, the true phylogeny is not treelike. However, the most parsimonious topology may still be apparent when considering only small amounts of introgression from populations of similar size. The interpretation of a most parsimonious tree topology is not valid where confidence intervals around the D statistic closest to zero, contain one or more of the other D statistics.

In this manner, the D statistic was used to indicate the most parsimonious topology for phylogenies including CladeX, ancient wisent, historical wisent, modern wisent, steppe bison and aurochs (Supplementary Figure 11). D statistics were calculated using ADMIXTOOLS version 3.0, git~3065acc5 [50].

Following concern over the limited amount of data for CladeX, particularly in samples other than 6A, we calculated the D statistics with sample 6A omitted from the analysis (Supplementary Figure 12). The most parsimonious topologies match in both cases.

Sensitivity to other factors were also investigated, such as setting a bovine SNP site coverage depth threshold of two (Supplementary Figure 13), changing the outgroup to *Bubalus bubalis* (Asian water buffalo, Supplementary Figure 14), and haploidisation by randomly sampling an allele at heterozygous sites (Supplementary Figure 15). None of these factors had notable influences on the outcome.

We also considered that the obtained topologies may have been caused by the small number of observed loci. To determine how sensitive the topology testing was missing data, we performed bootstrap resampling of the locus calls on decreasingly sized subsets of the data (Supplementary Table 7). For 10,000 bootstraps, we counted how often we obtained a result other than shown in Supplementary Figure 11.

For this bootstrap, a topology is considered to be simple if: (1) It has a D statistic which, uniquely amongst the set of three, is not significantly different from zero, or (2) All three are significantly different from zero but one has a D statistic closest to zero, with confidence intervals that do not overlap the D statistic for the other two topologies.

For simple topologies, we counted how often the bootstrap replicate suggested a simple topology that did not match the most parsimonious topology in Supplementary Figure 11. For non-simple topologies, we counted how often the result suggested any simple topology. In both cases, a lack of support for any simple topology (such as multiple topologies having a D statistic not significantly different from zero) was not counted.

This bootstrapping shows that the D statistics are robust to the small number of observed genotypes.

936 *Admixture proportion determination using an f4 ratio*
937
938 The proportion of the wisent's ancestry differentially attributable to the steppe bison
939 and the aurochs, was estimated with AdmixTools using an f4 ratio, as described in [50]
940 with sheep (*Ovis aries*) as the outgroup. For the admixture graph shown in
941 Supplementary Figure 16, the admixture proportion, α, is the ratio of two f4 statistics.

$$\alpha y = F4(A, O; X, C)$$

$$y = F4(A, O; B, C)$$

$$\alpha = \frac{\alpha y}{y} = \frac{F4(A, O; X, C)}{F4(A, O; B, C)}$$

942 For the estimation of admixture proportions using an f4 ratio, it is intended that the
943 ingroup A, while closely related to B, has diverged from B prior to the admixture
944 event. However, in the context of steppe ancestry for wisent, no such population
945 matching ingroup A was available. The admixture graph for wisent is shown in
946 Supplementary Figure 17.

$$\alpha y = F_4(AmericanBison, O; Wisent, Aurochs)$$

$$x + y = F_4(AmericanBison, O; Steppe, Aurochs)$$

$$\alpha \approx \frac{\alpha y}{x + y} = \frac{F_4(AmericanBison, O; Wisent, Aurochs)}{F_4(AmericanBison, O; Steppe, Aurochs)}$$

947 Where α in Supplementary Figure 17 is approximately determined by the f4 ratio for
948 small branch lengths *x*. The f4 ratio we calculate therefore represents a lower bound
949 on the proportion of steppe bison present in the wisent populations. The steppe
950 ancestry was found to be at least 0.891, with a standard error of 0.026 (Supplementary
951 Table 6-A).

952 Sensitivity to haploidisation was checked by randomly sampling an allele at
953 heterozygous sites (Supplementary Table 6-B), which had no notable influence on the
954 outcome.

955

956 *Hypergeometric test for shared derived alleles*
957
958 To test whether the wisent lineages (including CladeX) have a common hybrid
959 ancestry (Supplementary Figure 18A), or whether multiple independent hybridisation
960 events gave rise to distinct wisent lineages (Supplementary Figure 18B), we identify
961 nuclear loci which have an ancestral state in the aurochs lineage, but a derived state in
962 the steppe lineage (see next section 'identification of derived alleles'). Under the
963 assumption of a single hybrid origin, we expect a common subset of derived steppe
964 alleles to be present in the various wisent lineages. In contrast, multiple hybridisation
965 events would result in different subsets of derived steppe alleles being present in
966 different wisent lineages. Likewise, we expect the subset of derived aurochs alleles to
967 indicate either one, or multiple hybridisation events.

968 If the total number of derived steppe alleles is *s*, the number of derived steppe alleles
969 observed in one wisent lineage is *a*, and the number in a second wisent lineage is *b*,
970 then under model B, the number of sites which are found to be in common is a
971 random variable X~HGeom(*a*, *s*-*a*, *b*). Where HGeom is the hypergeometric

distribution, having probability mass function:

$$P(X = k) = \frac{\binom{a}{k}\binom{s-a}{b-k}}{\binom{s}{b}}$$

For the number of derived steppe alleles in common between two wisent lineages, c, we calculate $P(X \geq c)$. This indicates the likelihood of having observed c or more derived steppe alleles in common, if independent hybridisation events gave rise to both wisent and CladeX lineages.

Likelihoods were calculated for steppe derived alleles on all pairwise combinations of wisent lineages (Supplementary Table 8), and then repeated for derived aurochs alleles (Supplementary Table 9). This provides strong support for an ancestral hybridisation event occurring prior to the divergence of the wisent lineages.

We note that parallel genetic drift may also result in a pattern of alleles observed to be derived in the steppe lineage and the wisent lineages, however this is only a confounding factor where the parallel drift occurred in the post hybridisation lineage common to wisent and CladeX in Supplementary Figure 18A. Therefore, this only confounds the determination of genomic positions from a specific parent population, not that the wisent and CladeX lineages have shared ancestry post hybridisation. Alleles under strong selection following distinct hybridisation events would also be shared between lineages more often than if they were randomly distributed. We consider this situation unlikely, as it would require that the same alleles were randomly introgressed repeatedly, and then a strong selective advantage of the alleles at all times and in all environments.

Although we cannot reject the hypothesis that the modern European bison morph may be recent, and only appeared after the LGM as an adaptation to the Holocene environment in Europe, it would mean that the *Bos* mitochondrial lineage has been maintained in the steppe bison diversity throughout the late Pleistocene, and that only individuals carrying this mitochondrial lineage survived in Europe. Therefore, a hybrid origin of the European morph prior to 120 kyr, and maintained during the late Pleistocene, is more parsimonious with the current data.

Identification of derived alleles

The identification of a derived allele in the B lineage of Supplementary Figure 16, for the above analysis, can be performed in a simple way. If the ancestral allele is fixed in both C and the outgroup O, and the derived allele is fixed within B, then the site may be readily identified as derived. However, such fixed alleles are likely to be rare, especially in large populations, and therefore in limited number in our 10K SNP subset. Furthermore, a steppe bison derived allele observed in a wisent population may not be fixed in the wisent, as the population may also contain the ancestral allele from the aurochs lineage.

Relaxing the criterion of allele fixation in any lineage, we identify differential ancestry using the difference in allele frequencies between populations. An ancestral site is one in which the allele frequency closely matches that of the outgroup and a derived site has an allele frequency differing from the outgroup.

For the admixture graph in Supplementary Figure 16, where population X has ancestry from both B and C lineages, with outgroup O, we define an allele frequency shift in B, analogous to a derived state, if

$\hat{F}_2(C,O) < \hat{F}_2(X,C)$ and $\hat{F}_2(C,O) < \hat{F}_2(X,O)$,

where $\hat{F}_2(M,N)$is an unbiased estimate of $(m-n)^2$, for populations M and N with population allele frequencies *m* and *n* at a single locus, as in Appendix A of [50]. Similarly, we define the allele frequency shift in B to have the same shift in X if, in addition to the shift in B:

$\hat{F}_2(B,X) < \hat{F}_2(B,C)$ and $\hat{F}_2(B,X) < \hat{F}_2(B,O)$ and

$\hat{F}_2(B,X) < \hat{F}_2(X,C)$ and $\hat{F}_2(B,X) < \hat{F}_2(X,O)$ and

$\hat{F}_2(C,O) < \hat{F}_2(B,C)$ and $\hat{F}_2(C,O) < \hat{F}_2(B,O)$.

By observing a shared allele frequency shift instead of shared fixed alleles, we obtain greater sensitivity to the phylogenetic signal that is specific to one ancestral lineage. As for fixed derived alleles, the specific sites showing an allele frequency shift are identified, and can then be compared between multiple daughter populations.

Admixture proportion determination using ABC and simulated data

As the f4 ratio test is giving an upper limit to the amount of aurochs introgression (due to the branch length uncertainty shown in Supplementary Figure 17), we independently test the admixture proportions using simulated data and an ABC approach.

Approximate Bayesian Computation (ABC) is a likelihood-free methodology employed when calculating likelihood functions is either impossible or computationally expensive[51]. The methodology relies on being able to efficiently simulate data, and then compare simulated data to observed data. When simulated data is sufficiently close to the observed data, the parameters used to simulate the data are retained in a posterior distribution.

Consider a single locus, which for three individuals A, B, and C, two different genotypes are observed. The three possible patterns that can be observed are AB, BC, and AC, denoted by the tree tips with shared state. The observed pattern results from a single mutation somewhere on the gene tree, where the position of the mutation relative to the internal node defines which pattern is observed. For example, from the un-rooted gene tree in Supplementary Figure 19c, if a mutation occurs on the branch between C and the internal node, the pattern AB is observed. We assume the relevant time scales are short enough that multiple mutations at a single locus are rare (infinite sites model[52]).

Under the assumption of neutral and independent mutations, the number of fixed mutations accumulating on a branch is Poisson distributed with mean $\mu \times t$, where μ is mutations per locus per generation, and time t is in units of $2N_e$ generations[53,54]. The counts $\boldsymbol{n} = (n_{ab}, n_{bc}, n_{ac})$, of observed site patterns AB, BC, and AC, are random variables, which for topology X_1 (Supplementary Figure 19c),

$$n_{ab} \sim Pois(T_m + T_c),$$

$$n_{bc} \sim Pois(T_a),$$

$$n_{ac} \sim Pois(T_b),$$

and topology X_2 (Supplementary Figure 19d),

$$n_{ab} \sim Pois(T_c),$$
$$n_{bc} \sim Pois(T_m + T_a),$$
$$n_{ac} \sim Pois(T_b),$$

where $\boldsymbol{T} = (T_a, T_b, T_c, T_m)$ are branch lengths in units of evolutionary time of $2N_e\mu$ generations, and the total number of observed patterns is $N = n_{ab} + n_{bc} + n_{ac}$. Thus for a locus where two genotypes are observed, the probability of patterns AB, BC, AC, is given by $\boldsymbol{p^T} = (p_{ab}^{\boldsymbol{T}}, p_{bc}^{\boldsymbol{T}}, p_{ac}^{\boldsymbol{T}}.)$, where for topology X_1 (Supplementary Figure 19c),

$$\begin{aligned} P(\mathrm{AB}|\boldsymbol{T}, X_1) &= p_{ab}^{\boldsymbol{T},X_1} = (T_m + T_c)/(T_m + T_c + T_a + T_b) \\ P(\mathrm{BC}|\boldsymbol{T}, X_1) &= p_{bc}^{\boldsymbol{T},X_1} = T_a/(T_m + T_c + T_a + T_b) \\ P(\mathrm{AC}|\boldsymbol{T}, X_1) &= p_{ac}^{\boldsymbol{T},X_1} = T_b/(T_m + T_c + T_a + T_b) \end{aligned}$$

and for topology X_2 (Supplementary Figure 19d),

$$\begin{aligned} P(\mathrm{AB}|\boldsymbol{T}, X_2) &= p_{ab}^{\boldsymbol{T},X_2} = T_c/(T_m + T_c + T_a + T_b) \\ P(\mathrm{BC}|\boldsymbol{T}, X_2) &= p_{bc}^{\boldsymbol{T},X_2} = (T_a + T_m)/(T_m + T_c + T_a + T_b) \\ P(\mathrm{AC}|\boldsymbol{T}, X_2) &= p_{ac}^{\boldsymbol{T},X_2} = T_b/(T_m + T_c + T_a + T_b). \end{aligned}$$

We simulate site pattern counts for each of the two species trees in Supplementary Figure 19 by drawing from a Multinomial distribution, where for tree topology X_1, $\boldsymbol{n}^{X_1} \sim \mathrm{Mult}(N, \boldsymbol{p}^{\boldsymbol{T},X_1})$, and for tree topology $X_2, \boldsymbol{n}^{X_2} \sim \mathrm{Mult}(N, \boldsymbol{p}^{\boldsymbol{T},X_2})$.

Given a collection of site pattern counts from a hybrid tree with hybridisation parameter $\gamma \in [0,1]$ (Figure S19e), we expect that the combined site pattern counts will be a linear combination of the counts for the different topologies X_1 and X_2. This assumption is reasonable for a large number of total observations N. The simulated counts, $\boldsymbol{n}^{\gamma}$, of site patterns for the hybridised tree is then given by

$$\begin{aligned} \boldsymbol{n}^{\gamma} &= \gamma \boldsymbol{n}^{X_1} + (1-\gamma)\boldsymbol{n}^{X_2} \\ &= (n_{ab}^{\gamma}, n_{bc}^{\gamma}, n_{ac}^{\gamma}). \end{aligned}$$

As branch lengths are not known (μ, N_e and number of generations are all unknown), we use uninformative priors for the branch lengths. Furthermore, we only require relative branch lengths, so branch lengths $\boldsymbol{T}$ used for simulation were scaled such that $T_b = 1$. Hence we can meaningfully simulate counts of site patterns $\boldsymbol{n}^{\gamma}$ under hybridisation, for comparison to observed site pattern counts.

We perform ABC using the R package 'abc', with a ridge regression correction for comparison of the simulated and observed data using the "abc" function[55]. The distance between the observed and simulated data sets is calculated as the Euclidean distance in three-dimensional space. A tolerance $\epsilon = 0.005$ was chosen so that the closest $\ell \times \epsilon$ simulated data sets are retained. For each analysis we had $\ell = 100000$, resulting in 500 posterior samples.

We performed leave-one-out cross-validation using the function "cv4abc" on $\ell' = 250$ randomly selected simulations, and report the prediction error, calculated as

$$E_{\text{pred}} = \frac{\sum_{i=1}^{\ell'} (\hat{\gamma}_i - \gamma_i)^2}{\text{Var}(\gamma_i)}$$

for each analysis. At most the prediction error was 0.5111 standard deviations away from zero, and so we observe that the ridge regression has performed well (see Supplementary Table 11).

Similarly, on inspection of the cross-validation plots, we observe that the ridge regression performs well for γ, as the true simulated values of γ are well estimated by the ridge regression correction. Hence the correction has strengthened the parameter inference methodology when compared to a simple rejection algorithm.

We avoid reporting sample means due to the heavy negative skew in the posterior distributions of γ, and hence report the median (the most central ordered observed value) and mode of each distribution. The mode is estimated using a kernel density estimate of the posterior distribution. Not all simulated data is equally 'close' to the observed data, and the median and mode are weighted according to these distances[56].

The weighted posterior median was between 0.8250 and 0.8660, and the weighted posterior mode was between 0.9034 and 0.9384. These measures of centre indicate evidence for some non-zero level of hybridisation from the Aurochs genome. Evidence against hybridsation must be indicated by overwhelming support for either $\gamma = 0$ or $\gamma = 1$ (no mixing of the tree topologies). However, these values lie on either end of the support for the prior distribution of γ, and hence any resulting posterior distribution for γ. There- fore, classical highest probability density (HPD) intervals cannot be used to indicate uncertainty in the estimates of these measures of centre, as any interval of density less than 100% will result in zero and one being artificially omitted by construction. This is not evidence for or against hybridisation, but rather a consequence of the way in which we calculate HPD intervals.

Supplementary Table 11 gives empirical posterior probabilities for different levels of hybridisation. For example, the first column gives the empirical posterior probability of observing at least 1% hybridisation. This is found for each trio by calculating the total proportion of posterior samples where $0.01 \leq \gamma \leq 0.99$. In general, for some percentage of hybridisation α, Supplementary Table 11 reports

$$[P(\frac{\alpha}{100} \leq \gamma \leq 1 - \frac{\alpha}{100})]$$

for α = 1%, 2%, 3%, 4% and 5%, from the posterior distribution of γ.

As there is no accepted value of γ for which we can claim that significant hybridisation has occurred, we leave it to the reader to consider what they consider to be a significant level of hybridisation, and to find the appropriate probability. However, if one considers 1% hybridisation to be significant, then the observed data indicates that the data has between a 95.80% and 97.20% chance of being from a hybridised topology. Similarly, if one considers 5% hybridisation to be significant, then the observed data has between a 76.40% and 85.00% chance of being from a hybridised topology.

Asymmetrical hybridisation

In this study, we show that wisent and CladeX are of hybrid origin, certainly between ancient aurochs and steppe bison forms. This is consistent with the population structure of most bovids, where a single bull usually breeds with different females of multiple generations. As explained in[57], this usually results in asymmetrical hybridization when males of one species (steppe bison here) dominate males of the other species (aurochs here), therefore preferentially mating with female aurochs, as well as their offspring, potentially over several generations. In addition, male F_1 hybrids are usually sterile or sub-fertile, increasing the amount of steppe bison genomic contribution to the offspring. As illustrated in Supplementary Figure 20, after just a few generations, this mating process results in individuals that are essentially steppe bison for their nuclear genome, but with an aurochs mitochondrial genome (strictly maternally inherited), which is the result that we obtained from the genotyping of historical and ancient wisent individuals (including CladeX).

Supplementary Note 3:

Paleoenvironment reconstruction and stable isotope analyses in the Ural region

The Urals are a well sampled region, with the highest number of genotyped bones through time (Figure 5 and S22). We generated a convex hull based on geo-referenced site locations for all genotyped ancient samples collected from the Urals (Supplementary Figure 21). We used the HadCM3 global circulation model and BIOME4 model to reconstruct paleoclimate and environmental conditions for the Ural region throughout the period from 70,000 years ago to the present day.

We used the HadCM3 global circulation model to reconstructed paleoclimate proxies for the Ural region. The HadCM3 consists of linked atmospheric, ocean and sea ice models at a spatial resolution of 2.5° latitude and 3.75° longitude, resampled at a 1° x 1° latitude/longitude grid cell resolution [58]. The temporal resolution of the raw data is 1,000 year slices back to 22,000BP and 2,000 year slices from 22,000 to 80,000BP [58] We used these palaeo-climate simulations to derive estimates of annual mean daily temperature and Köppen-Geiger climate classifications [59] throughout the period from 70,000 years ago to the present day. We intersected each grid cell in the Ural study region (n = 51) with the derived climate estimates, at each point in time, using ArcGIS 10. We calculated the mean temperature for the region and change in the proportion of the study region represented by four Köppen climate classes, each differing temperature: Dfa (hot summers), Dfb (warm summers), Dfc (cool summers), Dfd (continental temperatures). These are shown in Supplementary Figure 22. Interestingly, our reconstructions for the Urals show a decrease in area with hot and warm summer conditions (Dfa and Dfb) after 35kya.

BIOME4 was used to infer paleovegetation types. BIOME4 is a coupled biogeographical and biogeochemical model that simulates the distribution of 28 plant functional types (PFT) at a global scale [60]. Model inputs for each grid cell are monthly climate (mean annual temperature, mean annual precipitation and mean annual sunshine hours), atmospheric [CO_2], and soil texture class. Ecophysiological constraints determine which PFT is likely to occur in each grid cell. A coupled carbon and water flux model calculates the leaf area index that maximizes net primary production (in gC m^{-2} $year^{-1}$) for each PFT. Competition between PFTs was simulated by using the optimal net primary production of each PFT as an index of competitiveness. Global maps of BIOME4 PFTs were accessed at the same spatial and temporal resolution as the paleoclimate data (http://www.bridge.bris.ac.uk/resources/simulations/). We grouped PFTs into three categories: Grassland (PFT identify numbers = 18-20); Tundra (ID = 22-26); and Forest (ID = 7-11). For each grid cell in the Ural study region, at each point in time, we determined whether the dominant PFT was grassland, tundra or forest. Interestingly the vegetation shift between an all forest-like landscape to a landscape represented by a large proportion of tundra and grassland-like vegetation occurred after 35kya, which coincides with a decrease in hot and warm summer conditions (see above).

These results from the paleovegetation and climate inferences agree with previous landscape reconstructions of the region: In the Middle Urals, where almost all the samplings sites were located, the areas covered with arboreal vegetation underwent

changes during MIS3. Spruce and birch open forests were widespread during coolings, and spruce and birch forest-steppe with occurrence of pine formed during warmings. Mesophilic meadows dominated by forbs and grasses were also prevalent during warm climatic events (Lapteva, 2008; 2009; Pisareva and Faustova, 2008). In the south, where one of the sites (Gofmana) is situated, steppe landscapes dominated by Asteraceae, Artemisia, and Poaceae were widespread. Spruce, birch and pine forests covered the areas along the rivers (Smirnov, Bolshakov, Kosintsev et al., 1990). The following was reconstructed for the territory of the Irtysh River: forest-steppe landscapes with pine (Pinus s/g Haploxylon) and spruce forests, as well as meadows with a predominance of Cyperaceae and Poaceae and small quantities of Artemisia and Chenopodiaceae (Araslanov *et al*. 2009).

During MIS2, periglacial forest-steppes dominated by herbaceous communities were typical of the Last Glacial Maximum. Larch, pine and birch covered the river-valleys. Herbaceous vegetation was dominated by goosefoot, sagebrush and grass (Grichuk 2002). Periglacial forest-steppes with arboreal vegetation, including pine-birch forests and small quantities of spruce have been reconstructed for the Last Glacial Termination. Areas covered with sagebrush-goosefoot steppes with small quantities of grass were widespread (Lapteva, 2007).

At later stages of MIS2, periglacial forb-grass forest-steppes with pine, birch and small quantities of spruce have been reconstructed for the Sur'ya 5 and Rasik 1 sites [61]. Periglacial steppes dominated by Artemisia, Rosaceae, Chenopodiaceae, Cichorioideae and Poaceae have been reconstructed for the Voronovka site. Pinus sylvestris and Betula pubescens with occurrence of spruce (Picea), oak (Quercus) and teil (Tilia) covered the river-valleys [62].

The palynological analyses and landscape reconstruction suggest that both bison forms inhabited semi-open landscapes of forest-steppe type, where arboreal vegetation was represented by birch, spruce, pine and sometimes larch, while steppe and meadow herbaceous communities were observed. However, only CladeX (specifically from the Gofmana site, during MIS 3, Rasik 1 and Sur'ya 5, and Voronovka sites, during MIS2) also inhabited steppe-like landscapes, showing a more diverse ecological niche than steppe in this region.

In addition to the paleo-climate and -vegetation reconstructions, stable isotope values (δ13C and δ15N) obtained for all the genotyped bison individuals from the Ural region were compared between steppe bison and wisent (Supplementary Figure 23). Wisent individuals displayed more diverse stable isotope ratios than the steppe bison individuals. This observation is consistent with feeding in more diverse vegetations communities, which correlates well with the reconstructed paleo-environments for the region in the time periods they are found.

Modelled paleo-climate and -vegetation reconstruction at the sampling locations in the southern Urals suggest drastic shifts, which coincide in time with the observed population replacements between steppe bison and wisent. More specifically, between 14 and 31 kya wisent were likely to exist in environmental condition characterised by relatively cold average temperatures, open landscapes with tundra-like flora, and the absence of warm summers. Although modern wisent are found today in wood-like habitats, it has been suggested that they are living in sub-optimal habitat, and paleodiet reconstructions have placed ancient wisent in tundra-like environments, in agreement with our observations [63].

Interestingly, the steppe bison was only recorded when forest vegetation was inferred to dominate the landscape, adding to the evidence that this form of bison might not have been exclusively steppe-adapted [63,64].

Supplementary Note 4:

Cave painting

The present survey, placing wisent across Europe (from the Urals/Caucasus to Ukraine/Italy) during MIS2 and late MIS3, suggests that depictions of bison in European Palaeolithic art, such as cave painting, carving and sculptures, are likely to include representations of wisent. Paleolithic art representations have often been used to infer the morphological appearance of steppe bison, sometimes in great detail [64,4,65–67]. And until now, the steppe bison (i.e., direct ancestor of modern American bison) has always been assumed to be the unique model present at the time of cave painting, and therefore, the diversity within the representations of bison was mainly explained by putative cultural and individual variations of style through time [68–70]. However, in the vast diversity of bison representations (820 pictures representing 20.6% of all known cave ornamentation, according to [71]), two consistent morphological types can be distinguished (see Fig 1 and Fig S24-27). The first type, abundant prior to the last glacial maximum, is characterized by long horns (with one curve), a very oblique dorsal line and a very robust front part of the body (solid shoulders versus hindquarters), all these traits being similar to the modern American bison. The second type, dominating the more recent paintings between 18 and 15 kya, displays thinner sinuous horns (often with double curve), a smaller hump and more balanced dimensions between the front and the rear of the body, similar to the modern wisent lineage, and to some extant the *Bos* lineage. The imposing figure of the steppe bison, with its high hump and long horns stepping out the head profile, certainly was a very strong influence on the artists painting in the cave in Europe before the last glacial maximum. However, later generations thoroughly depicted the slender shape of the more recent form of bison. Considering the geographical and temporal distribution of genotyped steppe bison and wisent presented here, particularly the ~16,000 years old wisent B individual from Northern Italy, it is likely that the variety of bison representations in Paleolithic art does not just come from stylistic evolution, but actually represents different forms of bison (i.e., pre and post-hybridisation) through time.

Supplementary References

1. Wolff, E. W., Chappellaz, J., Blunier, T., Rasmussen, S. O. & Svensson, A. Millennial-scale variability during the last glacial: The ice core record. *Quaternary Science Reviews* **29,** 2828–2838 (2010).
2. Shapiro, B. *et al.* Rise and Fall of the Beringian Steppe Bison. *Science* **306,** 1561–1565 (2004).
3. Leroi-Gourhan, A. & Allain, J. *Lascaux inconnu*. (CNRS, 1979).
4. Capitan, L., Breuil, H. & Peyrony, D. *La caverne de Font-de-Gaume, aux Eyzies (Dordogne)*. (Imprimerie du Chêne, 1910).
5. Lorblanchet, M. *La grotte ornée de Pergouset (Saint-Géry, Lot). Un sanctuaire secret paléolithique*. (Maison des Sciences de l'Homme, 2001).
6. Barrière, C. L'art pariétal de Rouffignac, la grotte aux cent mammouths. *Bulletins et Mémoires de la Société d'anthropologie de Paris* **10,** 144–145 (1983).
7. Groves, C. & Grubb, P. *Ungulate Taxonomy*. (Johns Hopkins University Press, 2011).
8. Drees, M. & Post, K. Bison bonasus from the North Sea, the Netherlands. *Cranium* **24,** 48–52 (2007).
9. Llamas, B. *et al.* High-Resolution Analysis of Cytosine Methylation in Ancient DNA. *PLoS ONE* **7,** e30226 (2012).
10. Skinner, A. R. *et al.* ESR dating at Mezmaiskaya Cave, Russia. *Applied Radiation and Isotopes* **62,** 219–224 (2005).
11. Pinhasi, R., Higham, T. F. G., Golovanova, L. V. & Doronichev, V. B. Revised age of late Neanderthal occupation and the end of the Middle Paleolithic in the northern Caucasus. *PNAS* **108,** 8611–8616 (2011).

1292 12. Reimer, P. J. *et al.* IntCal13 and Marine13 Radiocarbon Age Calibration Curves
1293 0–50,000 Years cal BP. *Radiocarbon* **55,** 1869–1887 (2013).
1294 13. Willerslev, E. & Cooper, A. Ancient DNA. *Proc Biol Sci* **272,** 3–16 (2005).
1295 14. Brotherton, P. *et al.* Neolithic mitochondrial haplogroup H genomes and the
1296 genetic origins of Europeans. *Nat Commun* **4,** 1764 (2013).
1297 15. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat.*
1298 *Protocols* **2,** 1756–1762 (2007).
1299 16. Meyer, M. & Kircher, M. Illumina Sequencing Library Preparation for Highly
1300 Multiplexed Target Capture and Sequencing. *Cold Spring Harb Protoc* **2010,**
1301 pdb.prot5448 (2010).
1302 17. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in
1303 multiplex sequencing on the Illumina platform. *Nucl. Acids Res.* **40,** e3–e3 (2012).
1304 18. Cone, R. W. & Schlaepfer, E. Improved In Situ Hybridization to HIV with RNA
1305 Probes Derived from PCR Products. *J Histochem Cytochem* **45,** 721–727 (1997).
1306 19. Liu, C., Bernstein, B. & Schreiber, S. *DNA linear amplification*. (Scion Publishin
1307 Ltd, 2005).
1308 20. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing
1309 libraries for multiplexed target capture. *Genome Res.* gr.128124.111 (2012).
1310 doi:10.1101/gr.128124.111
1311 21. Konietzko, U. & Kuhl, D. A subtractive hybridisation method for the enrichment
1312 of moderately induced sequences. *Nucleic Acids Res.* **26,** 1359–1361 (1998).
1313 22. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil–
1314 DNA–glycosylase treatment for screening of ancient DNA. *Philosophical*
1315 *Transactions of the Royal Society of London B: Biological Sciences* **22,** 939–949
1316 (2015).

1317 23. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-
1318 European languages in Europe. *Nature* **522,** 207–211 (2015).
1319 24. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA Sequence Capture of
1320 Mitochondrial Genomes Using PCR Products. *PLoS ONE* **5,** e14004 (2010).
1321 25. Untergasser, A. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucl.*
1322 *Acids Res.* **35,** W71–W74 (2007).
1323 26. Decker, J. E. *et al.* Resolving the evolution of extant and extinct ruminants with
1324 high-throughput phylogenomics. *PNAS* **106,** 18644–18649 (2009).
1325 27. Matukumalli, L. K. *et al.* Development and Characterization of a High Density
1326 SNP Genotyping Assay for Cattle. *PLoS ONE* **4,** e5350 (2009).
1327 28. Shankaranarayanan, P. *et al.* Single-tube linear DNA amplification (LinDA) for
1328 robust ChIP-seq. *Nat Meth* **8,** 565–567 (2011).
1329 29. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP
1330 detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat.*
1331 *Protocols* **9,** 1056–1082 (2014).
1332 30. Lindgreen, S. AdapterRemoval: Easy Cleaning of Next Generation Sequencing
1333 Reads. *BMC Research Notes* **5,** 337 (2012).
1334 31. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–
1335 Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
1336 32. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L.
1337 mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage
1338 parameters. *Bioinformatics* **29,** 1682–1684 (2013).
1339 33. Zimin, A. V. *et al.* A whole-genome assembly of the domestic cow, Bos taurus.
1340 *Genome Biology* **10,** R42 (2009).

1341 34. Li, H. A statistical framework for SNP calling, mutation discovery, association
1342 mapping and population genetical parameter estimation from sequencing data.
1343 *Bioinformatics* **27,** 2987–2993 (2011).
1344 35. Park, S. D. E. *et al.* Genome sequencing of the extinct Eurasian wild aurochs, Bos
1345 primigenius, illuminates the phylogeography and evolution of cattle. *Genome*
1346 *Biology* **16,** 234 (2015).
1347 36. Gouy, M., Guindon, S. & Gascuel, O. SeaView Version 4: A Multiplatform
1348 Graphical User Interface for Sequence Alignment and Phylogenetic Tree
1349 Building. *Mol Biol Evol* **27,** 221–224 (2010).
1350 37. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & Mclnerney, J. O.
1351 Assessment of methods for amino acid matrix selection and their use on empirical
1352 data shows that ad hoc assumptions for choice of matrix are not justified. *BMC*
1353 *Evolutionary Biology* **6,** 29 (2006).
1354 38. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and
1355 Model Choice Across a Large Model Space. *Syst Biol* **61,** 539–542 (2012).
1356 39. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-
1357 Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* **59,**
1358 307–321 (2010).
1359 40. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by
1360 sampling trees. *BMC Evolutionary Biology* **7,** 214 (2007).
1361 41. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth Skyride through a
1362 Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics.
1363 *Mol Biol Evol* **25,** 1459–1471 (2008).
1364 42. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed
1365 Phylogenetics and Dating with Confidence. *PLoS Biol* **4,** e88 (2006).

1366 43. Shapiro, B. *et al.* A Bayesian Phylogenetic Method to Estimate Unknown
1367 Sequence Ages. *Mol Biol Evol* **28,** 879–887 (2011).
1368 44. Ho, S. Y. W. *et al.* Bayesian Estimation of Substitution Rates from Ancient DNA
1369 Sequences with Low Information Content. *Syst Biol* **60,** 366–375 (2011).
1370 45. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
1371 analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690
1372 (2006).
1373 46. Stamatakis, A., Hoover, P. & Rougemont, J. A Rapid Bootstrap Algorithm for the
1374 RAxML Web Servers. *Syst Biol* **57,** 758–771 (2008).
1375 47. Pertoldi, C. *et al.* Phylogenetic relationships among the European and American
1376 bison and seven cattle breeds reconstructed using the BovineSNP50 Illumina
1377 Genotyping BeadChip. *Acta Theriol* **55,** 97–108 (2010).
1378 48. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis.
1379 *PLoS Genet* **2,** e190 (2006).
1380 49. Durand, E. Y., Patterson, N., Reich, D. & Slatkin, M. Testing for Ancient
1381 Admixture between Closely Related Populations. *Mol Biol Evol* **28,** 2239–2252
1382 (2011).
1383 50. Patterson, N. *et al.* Ancient Admixture in Human History. *Genetics* **192,** 1065–
1384 1093 (2012).
1385 51. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian
1386 Computation in Population Genetics. *Genetics* **162,** 2025–2035 (2002).
1387 52. Kimura, M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite
1388 Population Due to Steady Flux of Mutations. *Genetics* **61,** 893–903 (1969).
1389 53. Watterson, G. A. On the number of segregating sites in genetical models without
1390 recombination. *Theor Popul Biol* **7,** 256–276 (1975).

54. Hudson, R. in *Oxford Surveys in Evolutionary Biology* **7,** 1–44 (Oxford University Press, 1990).

55. Csilléry, K., François, O. & Blum, M. G. B. abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution* **3,** 475–479 (2012).

56. Blum, M. G. B. & François, O. Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20,** 63–73 (2009).

57. Groves, C. Current taxonomy and diversity of crown ruminants above the species level. *Zitteliana* **B 32,** 5–14 (2014).

58. Singarayer, J. S. & Valdes, P. J. High-latitude climate sensitivity to ice-sheet forcing over the last 120 kyr. *Quaternary Science Reviews* **29,** 43–55 (2010).

59. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **11,** 1633–1644 (2007).

60. Kaplan, J. O. Geophysical Applications of Vegetation Modeling. (Lund University, 2001).

61. Lapteva, E. G. Landscape-climatic changes on the eastern macroslope of the Northern Urals over the past 50000 years. *Russ J Ecol* **40,** 267–273 (2009).

62. Lapteva, E. G. & Korona, O. M. Holocene vegetation changes and anthropogenic influence in the forest-steppe zone of the Southern Trans-Urals based on pollen and plant macrofossil records from the Sukharysh cave. *Veget Hist Archaeobot* **21,** 321–336 (2011).

63. Bocherens, H., Hofman-Kamińska, E., Drucker, D. G., Schmölcke, U. & Kowalczyk, R. European Bison as a Refugee Species? Evidence from Isotopic

Data on Early Holocene Bison and Other Large Herbivores in Northern Europe. *PLoS ONE* **10,** e0115090 (2015).

64. Guthrie, R. D. *Frozen fauna of the Mammoth Steppe : the story of Blue Babe*. (University of Chicago Press, 1990).
65. Bandi, H.-G. ; H., W. ;. Sauter, M. R. ;. Sitter, B. *La Contribution de la Zoologie et de L'Ethologie a L'Interpretation de L'Art des Peuples Chasseurs Prehistoriques*. (Editions Universitaires, 1984).
66. Guthrie, R. D. *The nature of Paleolithic art*. (University of Chicago Press, 2005).
67. Paillet, P. *Le bison dans les arts magdaléniens du Périgord*. (CNRS éd, 1999).
68. Breuil, H. *Quatre cents siècles d'art pariétal; les cavernes ornées de l'âge du renne.* (Centre d'études et de documentation préhistoriques, 1952).
69. Leroi-Gourhan, A. *Préhistoire de l'art occidental*. (1965).
70. Petrognani, S. *De Chauvet à Lascaux: l'art des cavernes, reflet de sociétés préhistoriques en mutation*. (Editions Errance, 2013).
71. Sauvet, G. & Wlodarczyk. L'art pariétal, miroir des sociétés paléolithiques. *Zephyrus: Revista de prehistoria y arqueología* **53,** 217–240 (2000).

References in Russian:

Arslanov KH, Laukhin SA, Maksimov FE, *et al*. (2009) Radiocarbon Chronology and Landscapes of Western Siberian Lipovsk-Novoselovsky Interstadial (on evidence of study section near V. Lipovka) // Fundamental Problems of Quaternary: Resultats and Trends of Further Researches. (Ed. A.E. Kantorovich). Novosibirsk. P. 44 – 47. (in Russian).

Grichuk VP (2002) Vegetation of the Late Pleistocene. In: A.A.Velichko (ed.), Dynamics of terrestrial landscape components and inner marine basins of Northern Eurasia during the last 130 000 years. Moscow: GEOS Publishers, pp. 64-88. (in Russian).

Lapteva EG (2007) Реконструкция ландшафтно-климатических изменений на территории Среднего Зауралья в позднеледниковье и голоцене на основе палинологических данных из рыхлых отложений пещеры Першинская-1 // Эколоия древних и традиционных обществ. Вып. 3. (Ред. Н.П. Матвеева). С. 30 – 36. (in Russian).

Lapteva EG (2008) Major palaeogeographical stages and specific landscape-climatic changes on the eastern slope of the Urals during the last 50 kyrs (inferred from palynological data) // Problems of Pleistocene palaeogeography and stratigraphy. (Eds. N.S. Bolikhovskaya and P.A. Kaplin). Vol. 2. P. 196 – 204. (in Russian).

Pisareva VV, Faustova MA (2008) Reconstruction of Landscapes of Northern Russia during the Middle Valday Mega-Interstadial // Way to North: Paleoenvironment and Inhabitants of Arctic and Subarctic (Eds. A.A. Velichko and S.A. Vasil'ev). Moscow.P. 53 – 62. (in Russian).

Chapter 6

Conclusions

6.1 Summary

In this thesis we have presented two projects under the unifying theme of detecting departures from simplifying assumptions when analysing genetic data. These two projects yielded two new methods that were very different in the way they approached our theme. Both methods were also used for other purposes to address questions of interest in their own right.

Our first method is a powerful tool for exploring single-copy DNA, and is analogous to spectral decomposition of the classical allele correlation values for detecting linkage disequilibrium. It calculates the coordinates in gene space of sequenced individuals, while simultaneously calculating coordinates for informative sites in the genome. Due to our choice of scaling factor for the column scores, researchers are able to visualise the relationships between both individuals and sites of interest in the same coordinate-space. Researchers may also use our method for dimension reduction, reducing potentially massive numbers of SNPs into far fewer dimensions with potentially little reduction in information.

Our method allows for the coordinates of supplementary variables to be calculated

and visualised in the same coordinate-space as individuals and informative sites, and for the relationships between the supplementary variables and principal dimensions to be quantified. Our method also allows for additional sequences to be projected onto this coordinate-space.

To demonstrate the biological interpretability of our method, we identified known haplotype structure in human mitochondrial DNA (mtDNA). We were also able to efficiently visualise the strength of the relationships between supplementary variables and empirical sequence data. Our first analysis showed a strong geographic structure to genetic diversity for the extinct thylacine, both in Tasmania and on the mainland of Australia. Using polynomial regression we also detected a potential migration route for the thylacine radiating from New South Wales. In our second analysis, we showed that ghost bat populations are highly structured, with respect to genetic diversity, in colonies. This highlighted a particular ecological vulnerability of the ghost bats to mining practices that disrupt entire breeding colonies.

We then applied the method to a novel data set containing Aboriginal Australian mtDNA. This data is unique for its reliable provenance of the geographical history of the Aboriginal Australians prior to the post-European resettlement. We showed a strong relationship between genetic diversity and geographic location. Coupled with phylogenetic analyses of the macrohaplogroups, our method showed strong evidence that Aboriginal Australians inhabited the same discrete geographic areas, dating back to the original colonisation of Australia approximately 50,000 years before present.

Our second method which identifies proportions of admixture is mainly focussed on the simplifying assumption of interest: the departure from a tree-like evolutionary history. Identifying a significant departure would indicate the need to consider fitting an admixture graph. However, many publications appear specifically interested in the problem of estimating the proportion of ancestry in a hybrid species attributable to a parent species of interest, such as the proportion of Neanderthal ancestry in

modern humans.

This method estimates the posterior distribution of the proportion of ancestry of a parent species for a hybrid species, denoted γ. We used two methods, approximate Bayesian computation (ABC) and numerical integration, to estimate the posterior distribution of γ. We showed via a simulation study that our method performed well for a range of biologically reasonable scenarios. Naturally, our method was upwardly biased for very small values of γ, and this bias was more pronounced for the ABC method.

We applied our method to the genomes of pre-ice age European humans to detect the proportion of Neanderthal ancestry for nine ancient samples. We compared our results to those of Fu *et al.* obtained using the popular ratio of f_4 statistics, and found that our method consistently estimates similar results [18].

Finally, we used our method to investigate the evolutionary history of bovids in Europe, prior to the Holocene (11.7 thousand years before present). Our method, in concert with the ratio of f_4 statistics, showed that the wisent inhabiting Europe was a hybrid offspring of Steppe bison and aurochs, and that this hybridisation occurred approximately 120,000 years before present.

Our two projects addressed the need for statistically-rigorous methods to detect departures from simplifying assumptions for the complex evolutionary histories of individuals in sequence alignments. As the fields of statistical phylogenetics and population genetics continue to grow, and as the amount of genetic data available to researchers also grows, there will be a need to find new ways of analysing genetic data. As whole genome studies of organisms become more commonplace, methods will need to adapt to the increased amount of data, and to the increased complexity of the underlying models we are wish to answer. In these cases, methods to be able to validate simplifying assumption must also continue to grow and adapt.

6.2 Future Work

For the project concerned with spectral decompositions of single-copy DNA alignments, we aim to broaden the type of data the algorithm can accept. It would be trivial to allow the analysis of more complex molecules, such as microsatellite markers, amino acids or nuclear DNA. In the case of nuclear DNA, one could effectively visualise sites under linkage disequilibrium, and the populations to which they belong. It would also be natural to extend our method to analyse pseudo-haploid DNA by relaxing the need for zero-one indicator variables in the contingency table of frequency counts, and instead replacing them with the empirical proportions of observed nucleotides.

We also aim to further investigate the performance of our method to identify migration gradients in gene space. This is an active field of research for principal components analysis of nuclear DNA. Through simulation and empirical data from model species, we aim to test the performance of our method for detecting migration from provenanced single-copy alignments.

Finally, we wish to develop an `R-package` to implement this method to make it readily available for use by researchers.

For the project concerned with estimating proportions of admixture, we identify the need to include a correction for the effect of incomplete lineage sorting (ILS), a mechanism by which gene trees and sequence trees may have differing topologies. In Chapter 4, we avoid the effect of ILS, by assuming that sufficient time has passed since the divergence of the parent species. Using classical population genetics theory we aim to include the probability of discordant gene trees due solely to ILS.

For our numerical integration approach to find the marginal posterior distribution of the mixing parameter, γ, we used an equally spaced grid of points for the parameters. We aim to include a preprocessing step to find the optimal set of grid points such that we invest more computational effort in evaluating the integral for regions of the

joint distribution that are of greatest interest, and to simultaneously avoid evaluating the joint distribution for regions of near-zero probability density.

Finally, we wish to develop an `R-package` to implement our method to make it readily available for use by researchers.

Bibliography

[1] Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., *et al.* 2013. Hybridization and speciation. *Journal of Evolutionary Biology*, 26(2): 229–246.

[2] Alexander, D. H., Novembre, J., and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9): 1655–1664.

[3] Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics*, 23(2): 147.

[4] Beaumont, M. A., Zhang, W., and Balding, D. J. 2002. Approximate Bayesian Computation in population genetics. *Genetics*, 162(4): 2025–2035.

[5] Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8): 4563–4568.

[6] Blum, M. G. and François, O. 2010. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20(1): 63–73.

[7] Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., and Drummond, A. J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4): e1003537.

[8] Brisbin, A. 2010. *Linkage analysis for categorical traits and ancestry assignment in admixed individuals*. Ph.D. thesis, Cornell University.

[9] Bruening, G., Lyons, J., *et al.* 2000. The case of the FLAVR SAVR tomato. *California Agriculture*, 54(4): 6–7.

[10] Cabreros, I. and Storey, J. D. 2017. A nonparametric estimator of population structure unifying admixture models and principal components analysis. *bioRxiv*, page 240812.

[11] Clark, P. U., Dyke, A. S., Shakun, J. D., Carlson, A. E., Clark, J., Wohlfarth, B., Mitrovica, J. X., Hostetler, S. W., and McCabe, A. M. 2009. The last glacial maximum. *Science*, 325(5941): 710–714.

[12] Consortium, . G. P. *et al.* 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68.

[13] Coyne, J. A. and Orr, H. A. 1998. The evolutionary genetics of speciation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 353(1366): 287–305.

[14] Darwin, C. 2004. *On the origin of species, 1859*. Routledge.

[15] Disotell, T. R. 1999. Human evolution: origins of modern humans still look recent. *Current Biology*, 9(17): R647–R650.

[16] Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8): 2239–2252.

[17] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.

[18] Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., Mittnik, A., *et al.* 2016. The genetic history of ice age Europe. *Nature*, 534(7606): 200.

[19] Green, R. E., Malaspinas, A.-S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., *et al.* 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3): 416–426.

[20] Hellenthal, G., Busby, G. B., Band, G., Wilson, J. F., Capelli, C., Falush, D., and Myers, S. 2014. A genetic atlas of human admixture history. *Science*, 343(6172): 747–751.

[21] Huerta-Sánchez, E., Jin, X., Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., Somel, M., Ni, P., *et al.* 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512(7513): 194.

[22] Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408(6813): 708.

[23] Kingman, J. F. 2000. Origins of the coalescent: 1974-1982. *Genetics*, 156(4): 1461–1463.

[24] Kuhner, M. K. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, 22(6): 768–770.

[25] Lanfear, R., Calcott, B., Ho, S. Y., and Guindon, S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29(6): 1695–1701.

[26] Lawson, L. P., Bates, J. M., Menegon, M., and Loader, S. P. 2015. Divergence at the edges: peripatric isolation in the montane spiny throated reed frog complex. *BMC evolutionary biology*, 15(1): 128.

[27] Liebers, D., De Knijff, P., and Helbig, A. J. 2004. The Herring Gull complex is not a ring species. *Proceedings of the Royal Society B: Biological Sciences*, 271(1542): 893.

[28] Martin, S. H., Davey, J. W., and Jiggins, C. D. 2014. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, 32(1): 244–257.

[29] Menozzi, P., Piazza, A., and Cavalli-Sforza, L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science*, 201(4358): 786–792.

[30] Mersmann, O., Beleites, C., Hurling, R., Friedman, A., and Ulrich, J. 2004. microbenchmark: accurate timing functions; 2015. *URL http://CRAN. R-project. org/package= microbenchmark. R package version*, pages 1–4.

[31] Meyer, M. and Kircher, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 2010(6): pdb–prot5448.

[32] Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1): 268–274.

[33] Patterson, N., Price, A., and Reich, D. 2006. Population structure and eigenanalysis. *PLoS genet*, 2(12): e190.

[34] Patterson, N. J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient admixture in human history. *Genetics*, pages genetics–112.

[35] Posth, C., Renaud, G., Mittnik, A., Drucker, D. G., Rougier, H., Cupillard, C., Valentin, F., Thevenet, C., Furtwängler, A., Wißing, C., *et al.* 2016. Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a Late Glacial population turnover in Europe. *Current Biology*, 26(6): 827–833.

[36] Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, 5(6): e1000519.

[37] Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959.

[38] Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., De Filippo, C., *et al.* 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481): 43.

[39] Reich, D., Price, A. L., and Patterson, N. 2008. Principal component analysis of Genetic data. *Nature Genetics*, 40(5): 491–492.

[40] Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L., *et al.* 2010. Genetic history of an archaic hominin group from Denisova Cave in siberia. *Nature*, 468(7327): 1053.

[41] Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., *et al.* 2001. Linkage disequilibrium in the human genome. *Nature*, 411(6834): 199.

[42] Rhymer, J. M. and Simberloff, D. 1996. Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics*, 27(1): 83–109.

[43] Sankararaman, S., Sridhar, S., Kimmel, G., and Halperin, E. 2008. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2): 290–303.

[44] Sankararaman, S., Mallick, S., Dannemann, M., Prüfer, K., Kelso, J., Pääbo, S., Patterson, N., and Reich, D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492): 354.

[45] Slatkin, M. and Racimo, F. 2016. Ancient DNA and human history. *Proceedings of the National Academy of Sciences*, page 201524306.

[46] Soubrier, J., Gower, G., Chen, K., Richards, S. M., Llamas, B., Mitchell, K. J., Ho, S. Y., Kosintsev, P., Lee, M. S., Baryshnikov, G., *et al.* 2016. Early cave Art and ancient DNA record the origin of European bison. *Nature Communications*, 7: 13158.

[47] Sutton, P. 2004. *Native title in Australia: An ethnographic perspective*. Cambridge University Press.

[48] Tattersall, I. 2009. Human origins: out of Africa. *Proceedings of the National Academy of Sciences*, 106(38): 16018–16021.

[49] Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17: 57–86.

[50] van de Loosdrecht, M., Bouzouggar, A., Humphrey, L., Posth, C., Barton, N., Aximu-Petri, A., Nickel, B., Nagel, S., Talbi, E. H., El Hajraoui, M. A., *et al.* 2018. Pleistocene North African genomes link Near Eastern and sub-saharan African human populations. *Science*, 360(6388): 548–552.

[51] Yang, Z. 2010. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome Biology and Evolution*, 2: 200–211.