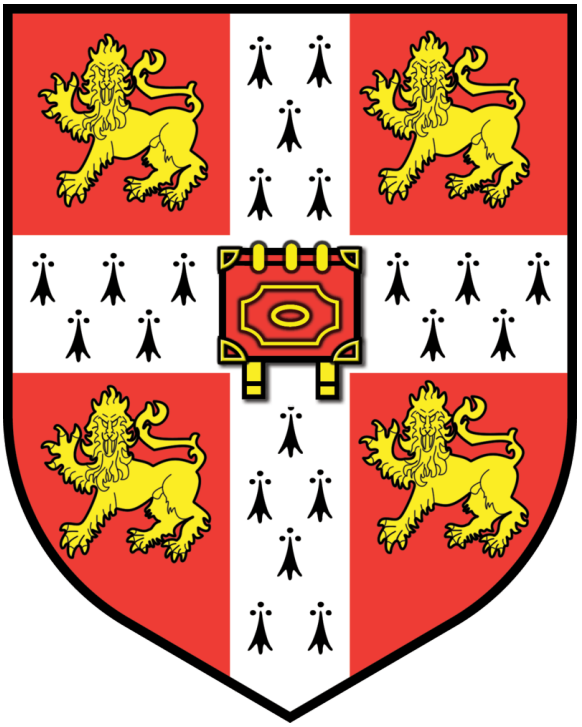
Painting the Past:

Uncovering Ancestral Contributions to Complex Human Phenotypes in Western Eurasia

William Barrie

Pembroke College

University of Cambridge



This thesis is submitted for the degree of

Doctor of Philosophy

January, 2023

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the Faculty of Biology Degree Committee.

William Barrie

January 2023

Abstract

Painting the Past: Uncovering Ancestral Contributions to Complex Human Phenotypes in Western Eurasia
William Barrie

The high mutation load within and phenotypic differences between modern human populations remain poorly understood. Understanding these phenomena would lead to a better understanding of the origins of complex phenotypes, including genetically-influenced diseases, and their geographic distributions.

The aim of this thesis is to uncover the genetic origins of complex human phenotypes, from the Last Glacial Maximum until the Bronze Age in western Eurasia. Specifically, it aims to assess the contributions of differentiated genetic ancestries which existed in this period, and link this to modern-day differences in disease susceptibility.

To achieve this, methods were developed to infer local ancestry in a large modern panel, the UK Biobank, using new ancient reference genomes. Modern samples were selected based on a ‘typical ancestral profile’ for each country represented in the UK Biobank. This dataset was then used to infer the genome-wide ancestry components of modern populations, and the contribution of each ancestry to a polygenic phenotype using a new statistic analogous to a polygenic risk score based on local ancestry probabilities. An in-depth investigation into the origins of Multiple Sclerosis (MS) was performed.

This project was the first to use ancient DNA to infer local ancestry in a very large modern panel to assess ancestral contributions to polygenic phenotypes. Simulations showed that the accuracy of ancestry assignment was good. Differences in average ancestry components were calculated per-country within Eurasia and north Africa, and per-county within Britain, reflecting past episodes of migration and admixture. Aggregate ancestral contributions to phenotypes known to be over-dispersed in ancient populations were then calculated, including height, BMI and some psychiatric traits. Finally, the origins of the genetic risk for MS were traced to the Bronze Age Steppe populations; positive selection drove these variants to higher frequency, likely in response to novel pathogen exposure resulting from lifestyles changes and leading to a heterogeneous risk profile across Europe today.

These results demonstrate the power of combining large ancient and modern DNA panels, using local ancestry assignment methods, to investigate the histories of genetic variants and associate them with selection due to differing ancient lifestyles, or drift. This can explain geographic differences in genetic risk, and highlights the importance of the Bronze Age as a determinant of modern immune response. This may have clinical implications for the treatment of auto-immune diseases, for example concerning childhood pathogen exposure.

Acknowledgements

I am grateful beyond words to all those who have helped me on this long journey, many of whom are not mentioned here.

My first thanks goes to my supervisor, Eske Willerslev, and my co-supervisors Dan Lawson and Rasmus Nielsen. Thank you for your guidance, patience, and trust in me. I would like to thank everyone in the GeoGenetics Group in Copenhagen, and specifically Line Olsen and Anna Razeto Richter, without whom almost nothing would happen, and Evan Irving-Pease, who has been a constant source of knowledge and guidance, always in good humour. Thank you also to Anders Rosengren and Andrés Ingason; it has been a pleasure to work with you both.

In Cambridge, our small but formidable GeoGenetics family was a constant source of comfort, insight and laughter. Thank you to Alison Sutherland, Aramish Fatima, Helia Naji, Angeliki Kanouta, Ana Prohaska, Yucheng Wang, and Ruairidh Macleod. Andrea Manica and the entire DAB 1st floor adopted me and invited me to their events, and Richard Durbin and Alice Pearson helped me with regular calls during lockdown - thank you all. Going further back in my time at Cambridge, Maanasa Raghavan first inspired me to pursue a PhD in ancient genomics, for which I will always be grateful.

In the last year I have worked closely with a small group to study the evolutionary origins of multiple sclerosis. I am constantly amazed by your intellect, passion, and generosity. I believe we have done a lasting good of which I am immensely proud. Thank you to Lars Fugger, Astrid Iversen, Yaoling Yang, Lise Torp Jensen, Kate Attfield, and Gabriele Scorrano.

Research of this type cannot take place without institutional support over many years, and steady and generous funding. I would like to thank Clare College and Pembroke College for their pedagogical and pastoral support during my undergraduate and graduate studies; the Department of Zoology for fostering an environment which is friendly, welcoming and encompasses an astonishingly and admirably broad intellectual scope; and Hanne and Torkel Weis-Fogh, whose generous bequeathment funded this PhD.

It is hard to summarise the support I received outside academic circles for the duration of my PhD. I must mention in particular my Cambridge family, Hani, Harry, Tai, Cuti, Ignacy,

Melissa, George, Diego, and many others; thank you to Brucie, for putting up with me in a small flat through winter lockdowns (‘get binned, learn things’); to Ben, for being a source of happiness and entertainment in that period, and much more later; and to every member of the Cambridge lacrosse community, which has given me so much over the last seven years.

Words in an acknowledgement will never do justice to my gratitude to my family. I completed a significant part of the work of this thesis at home during the scary and uncertain period of the Covid-19 pandemic, during lockdowns that seemed never to end. Thank you Elspeth, Chris, Tom and Gabe for your support, then and always. This work is dedicated to you.

Finally, to Dr Thomas Craig Sinclair (1932 - 2009), who would have been proud of this grandson.

Publications

Versions of Chapters One, Two, Three and Four have previously been published, or are currently in review, as the following papers (§ denotes joint first authors, @ denotes joint last authors).

Population Genomics of Stone Age Eurasia

Morten E. Allentoft§, Martin Sikora§, Alba Refoyo-Martínez§, Evan K. Irving-Pease§, Anders Fischer§, William Barrie§, Andrés Ingason§, Jesper Stenderup, Karl-Göran Sjögren, Alice Pearson, Bárbara Sousa da Mota, Bettina Schulz Paulsson, Alma Halgren, Ruairidh Macleod, Marie Louise Schjellerup Jørkov, Fabrice Demeter, Maria Novosolov, Lasse Sørensen, Poul Otto Nielsen, Rasmus H.A. Henriksen, Tharsika Vimala, Hugh McColl, Ashot Margaryan, Melissa Ilardo, Andrew Vaughn, Morten Fischer Mortensen, Anne Birgitte Nielsen, Mikkel Ulfeldt Hede, Peter Rasmussen, Lasse Vinner, Gabriel Renaud, Aaron Stern, Theis Zetner Trolle Jensen, Niels Nørkjær Johannsen, Gabriele Scorrano, Hannes Schroeder, Per Lysdahl, Abigail Daisy Ramsøe, Andrei Skorobogatov, Andrew Joseph Schork, Anders Rosengren, Anthony Ruter, Alan Outram, Aleksey A. Timoshenko, Alexandra Buzhilova, Alfredo Coppa, Alisa Zubova, Ana Maria Silva, Anders J. Hansen, Andrey Gromov, Andrey Logvin, Anne Birgitte Gotfredsen, Bjarne Henning Nielsen, Borja González-Rabanal, Carles Lalueza-Fox, Catriona J. McKenzie, Charleen Gaunitz, Concepción Blasco, Corina Liesau, Cristina Martinez-Labarga, Dmitri V. Pozdnyakov, David Cuenca-Solana, David O. Lordkipanidze, Dmitri En'shin, Domingo C. Salazar-García, T. Douglas Price, Dušan Borić, Elena Kostyleva, Elizaveta V. Veselovskaya, Emma R. Usmanova, Enrico Cappellini, Erik Brinch Petersen, Esben Kannegaard, Francesca Radina, Fulya Eylem Yediay, Henri Duday, Igor Gutiérrez-Zugasti, Inna Potekhina, Irina Shevnina, Isin Altinkaya, Jean Guilaine, Jesper Hansen, Joan Emili Aura Tortosa, João Zilhão, Jorge Vega, Kristoffer Buck Pedersen, Krzysztof Tunia, Lei Zhao, Liudmila N. Mylnikova, Lars Larsson, Laure Metz, Levon Yepiskoposyan, Lisbeth Pedersen, Lucia Sarti, Ludovic Orlando, Ludovic Slimak, Lutz Klassen, Malou Blank, Manuel González-Morales, Mara Silvestrini, Maria Vretemark, Marina S. Nesterova, Marina Rykun, Mario Federico Rolfo, Marzena Szmyt, Marcin Przybyła, Mauro Calattini, Mikhail Sablin, Miluše Dobisíková, Morten Meldgaard, Morten Johansen, Natalia Berezina, Nick Card, Nikolai A. Saveliev, Olga Poshekhonova, Olga Rickards, Olga V. Lozovskaya, Olivér Gábor, Otto Christian Uldum, Paola Aurino, Pavel Kosintsev, Patrice Courtaud, Patricia Ríos, Peder Mortensen, Per Lotz, Per Persson, Pernille Bangsgaard, Peter de Barros Damgaard, Peter Vang Petersen, Pilar Prieto Martinez, Piotr Włodarczak, Roman V. Smolyaninov, Rikke Maring, Roberto

Menduiña, Ruben Badalyan, Rune Iversen, Ruslan Turin, Sergey Vasilyiev, Sidsel Wåhlin, Svetlana Borutskaya, Svetlana Skochina, Søren Anker Sørensen, Søren H. Andersen, Thomas Jørgensen, Yuri B. Serikov, Vyacheslav I. Molodin, Vaclav Smrcka, Victor Merz, Vivek Appadurai, Vyacheslav Moiseyev, Yvonne Magnusson, Kurt H. Kjær, Niels Lynnerup, Daniel J. Lawson, Peter H. Sudmant, Simon Rasmussen, Thorfinn Korneliussen@, Richard Durbin@, Rasmus Nielsen@, Olivier Delaneau@, Thomas Werge@, Fernando Racimo@, Kristian Kristiansen@, Eske Willerslev@
bioRxiv 2022.05.04.490594; doi: https://doi.org/10.1101/2022.05.04.490594
In review at Nature, January 2023.

The Selection Landscape and Genetic Legacy of Ancient Eurasians
Evan K. Irving-Pease§, Alba Refoyo-Martínez§, Andrés Ingason§, Alice Pearson§, Anders Fischer§, William Barrie§, Karl-Göran Sjögren, Alma S. Halgren, Ruairidh Macleod, Fabrice Demeter, Rasmus A. Henriksen, Tharsika Vimala, Hugh McColl, Andrew Vaughn, Aaron J. Stern, Leo Speidel, Gabriele Scorrano, Abigail Ramsøe, Andrew J. Schork, Anders Rosengren, Lei Zhao, Kristian Kristiansen, Peter H. Sudmant@, Daniel J. Lawson@, Richard Durbin@, Thorfinn Korneliussen@, Thomas Werge@, Morten E. Allentoft@, Martin Sikora@, Rasmus Nielsen@, Fernando Racimo@, Eske Willerslev@
bioRxiv 2022.09.22.509027; doi: https://doi.org/10.1101/2022.09.22.509027
In review at Nature, January 2023.

Genetic risk for Multiple Sclerosis originated in Pastoralist Steppe populations
William Barrie§, Yaoling Yang§, Kathrine E. Attfield§, Evan Irving-Pease§, Gabriele Scorrano§, Lise Torp Jensen§, Angelos P. Armen, Evangelos Antonios Dimopoulos, Aaron Stern, Alba Refoyo-Martinez, Abigail Ramsøe, Charleen Gaunitz, Fabrice Demeter, Marie Louise S. Jørkov, Stig Bermann Møller, Bente Springborg, Lutz Klassen, Inger Marie Hyldgård, Niels Wickmann, Lasse Vinner, Thorfinn Sand Korneliussen, Martin Sikora, Kristian Kristiansen, Santiago Rodriguez, Rasmus Nielsen, Astrid K. N. Iversen@, Daniel J. Lawson@, Lars Fugger@, Eske Willerslev@
bioRxiv 2022.09.23.509097; doi: https://doi.org/10.1101/2022.09.23.509097
In review at Nature, January 2023.

Contents

Introduction

Motivation for work

This thesis concerns the genetic origins of complex human phenotypes, and in particular diseases, very roughly focussed on the period from the Last Glacial Maximum (LGM) until the Bronze Age (~16 thousand years ago (kya) until ~2.5 kya). Geographically, it focusses on western Eurasia and the people who inhabited this region in this period.

Two broad and related observations motivate the study of the genetic origins of complex human phenotypes. The first is that, somewhat counterintuitively, human populations harbour a high mutation load (Henn et al., 2015); how and why highly heritable human-specific disorders, which severely impact human fitness, are still segregating with our species, remains unanswered. The second is that differences in phenotypes within and between modern human populations, driven by environmental variation, genetic variation, and interaction between the two (Plomin et al., 1977), including differences in genetic susceptibilities to diseases, remain poorly understood (e.g. Kenney et al., 2017; Brinkworth, 2017).

There are multiple competing explanations for a high mutation load which address the fate of deleterious mutations, i.e. those which reduce (inclusive) fitness. Specifically, when the rate of creation of deleterious mutations in a population (introduced at random in the germline) exceeds the rate at which these are eliminated by negative selection, the mutation load will increase. This can happen for a number of reasons, for example: weak negative selection, in which harmful mutations have not been purged from the population due to selection acting weakly, often due to very small effect sizes or recessive effects; genetic drift, in which random sampling of a population's set of alleles, for example when the population size is small, results in a random increase in frequency of harmful alleles (Henn et al., 2015); or balancing selection, in which multiple alleles are actively maintained in a population at a frequency above that expected by genetic drift (Lenz et al., 2016). Conversely, mutations which are now deleterious may have been under positive natural selection in the past due to environment-specific effects, leading to a high frequency in modern populations (Graves & Weinreich, 2017). Regardless of its origins, the high mutation load in modern human populations continues to cause a high prevalence of diseases, including chronic and autoimmune diseases.

Most studies have focussed on genetic drift to explain differences in phenotypes between modern human populations. For example, the mutation load in Out-of-Africa populations is higher than African populations (Henn et al., 2015). Other studies have shown that differing

demographic events, such as population bottlenecks, divergence, and isolation, have shaped the human genome differently between populations, resulting in phenotypic diversity (Prohaska et al., 2019). This is particularly important for recessive disorders, which are much more likely to be homozygous in these populations. For example, this has been studied in Ashkenazi Jews (e.g. Rivas et al., 2018), who have a high prevalence of rare alleles associated with Mendelian disorders such as Tay-Sachs disease and Gaucher disease due to a past effective population size of approximately 350 (Carmi et al., 2014), and in Greenlandic Inuit (Pedersen et al., 2017) among others. In Eurasia, Stone Age hunter-gatherers show low genomic diversity, indicative of small mobile populations that underwent repeated bottlenecks (Allentoft et al., 2015; Sikora et al., 2017; Skoglund et al., 2014). To what extent these bottlenecks in Eurasian prehistory shape modern health and disease variation remains an open question.

Relatedly, differential selection in these populations would leave a similar signature of genetic risk heterogeneity, but has generally been understudied due to little information on the genetics and phenotypes of past populations, the difficulty in exporting polygenic risk scores (PRS) to ancient populations, and the high numbers of individuals required to investigate polygenic traits using association studies (Prohaska et al., 2019). In nearly all studies of past selection, the signal has not been decomposed on a regional or population level.

Explaining both the high prevalence of genetically-influenced disease and the geographic distribution of these prevalences is of basic scientific interest. This is the aim of this thesis. Fundamentally, this enquiry attempts to address *why* genetically-influenced diseases exist as they do rather than *how* they are expressed - in Tinbergen's terminology, the *ultimate* rather than the *proximate* cause (Tinbergen, 1963). Answering this sheds light not just on why a particular genetic disorder persists in the present day, but also the conditions under which it might have originated. The full implications for these findings are covered in the Discussion Chapter.

Having outlined the motivations for the study, there are two further justifications necessary. The first question which might reasonably be posed here is why study humans ahead of any other given species. There are several answers. Humans represent a perhaps unique example among extant species of genetically isolated populations undergoing independent evolution in the context of highly divergent cultural, ecological and environmental niches (Laland et al., 2001). As a result, we might a priori expect that these different ancient populations have contributed differently to genetic risk in modern populations. Humans

therefore represent a promising test case. There are also practical reasons: for example, humans are the only species for which we have more than a very rudimentary understanding of the genetic contributions to polygenic traits, enabled through large-scale association studies in the past 15 years (Visscher et al., 2017, Loos, 2020), and are the only species with large numbers of sequenced ancient genomes available. And finally, there are good ethical justifications for focussing on human disease: through studying the genetic factors in disease aetiology, there is the potential to reduce human suffering.

The second justification required here is in answer to the question of 'why study Europe?', especially when so much has been written about the bias currently inherent in the field of genetics towards white populations over non-white and indigenous groups, and the subsequent aggravation of pre-existing health disparities between these groups (e.g. Martin et al., 2019); and also given the long and often explicitly racist history of our field, including its origins in eugenics (Adams, 1990) and in justifying racial categorisations and differential treatment (e.g. Burchard et al., 2003). The primary answer to this is that at the outset of my PhD, it simply was not possible to carry out this research on anything other than a modern European population, because we did not have sufficient numbers of ancient samples to capture the genetic variation inherent to the ancient ancestries for other geographic regions, or large enough modern panels of genomes required to generate power for our tests. These limitations are rapidly changing, and it is my sincere hope that this work will act as a foundation, aiding the much needed elucidation of population-specific genetic outcomes to improve people's lives in other populations, and that the methods developed here might be applied similarly to populations across the world, either by me or others.

Finally, a quick note on the terminology used. The field of genetics and ancient DNA (aDNA) in particular has rightly been criticised for the conflation of genetics with culture, and using confusing terminology either inconsistently or misleadingly. Terms like 'ancestry' and 'population' can have both specific technical meanings as well as a more widely understood 'lay' meaning. In the absence of alternative terminology, I use the term 'ancestry' throughout this thesis to mean genetic ancestry, i.e. the specific path that genetic material takes backwards in time; I do not mean to conflate this with any form of identity associated with ancestry, which is a complex mixture of cultural, historical and personal factors. There cannot be a contradiction between genetic ancestry and someone's identity based on their ancestry. I also use the term 'population' to mean a grouping of individuals based on some biological factor, such as a measure of distance in DNA sequences; this may or may not accord with cultural practices. This is discussed more below. Finally, although we associate genetic ancestries with broad cultural practices such as hunting and gathering or farming, I

recognise that these associations are simplistic and individuals often defy these expectations, and I urge the reader to keep this in mind.

Outline of Introduction

Having outlined the aims of the thesis and their motivation, some background introduction is necessary. This will form the remainder of this Introductory chapter. The first area concerns the demographic history of Europe: this describes which people lived where and when, and what sort of lifestyles they led. Much of our understanding comes from traditional archaeology and absolute dating techniques, but has been added to in recent years by the addition of data generated from aDNA. This latter source will be the focus of this introduction, as it forms the basis of our understanding of the genetic ancestry of modern Europeans, from which we can begin to ask questions about the ancestry-specific origins of genetic risk. There will then be an overview of current and previous approaches to questions of the genetic origins of complex phenotypes, both with and without aDNA data, first with a focus on methods for detecting natural selection and then with a focus on methods based on Local Ancestry Inference (LAI) as these form the basis of my work. Finally, there will be a more detailed overview of the datasets, approaches used, and chapter contents.

Demographic history of Europe

Since the application of high-throughput sequencing to well-preserved human subfossils yielded the first ancient human genome in 2010, the number of sequenced ancient human genomes has grown exponentially (Liu et al., 2021). This has revolutionised our understanding of the past genetic history of human populations, as well as providing insights into adaptation and admixture with archaic populations. Nowhere has this process been more pronounced than in West Eurasia, which currently has the highest density of sampled remains (Allentoft et al., 2022).

Anatomically modern humans were widely distributed throughout Europe by at least 42-45 kya (Higham et al., 2011). There were subsequently several major repopulations, with the first around 30-35 kya associated with the replacement of the Aurignacian with the Gravettian culture (Fu et al., 2016). Of the very early Pleistocene lineages, some showed no genetic continuity to modern populations at all (e.g. Fu et al., 2015), although from 37 kya onwards all individuals share some affinity with modern European groups (Olalde and Posth, 2020). Broadly over time there was increased population structure and migration rates across Eurasia (Yang & Fu, 2018), and populations before the Last Glacial Maximum (LGM) showed a lower mutation load and higher diversity compared with populations after the LGM, particularly in northern latitudes (Svensson et al., 2021). There were further major repopulations after the LGM around 19-25 kya, from southern European and central Eurasian refugia; and around 14.5 kya to form the Mesolithic populations of Europe (the so-called *Villabruna* cluster) (Fu et al., 2016). There was an east-west cline in these hunter-gatherer populations, the origins of which are unclear (Haak et al., 2015), which is traditionally split into western hunter-gatherers (WHG) and eastern hunter-gatherers (EHG) stretching up to the Samara region of western Russia (de Barros Damgaard et al., 2018).

Starting around 8.5 kya, this Mesolithic ancestry was marginalised as ancestry related to that found in Neolithic northwest Anatolia (Skoglund et al., 2012; Haak et al., 2015), and ultimately early farmers from the Levant and Iran (Broushaki et al., 2016), expanded throughout Europe. Over the next 4,000 years the Neolithic farmers and remnant hunter-gatherers merged, with admixed populations typically having 10-25% hunter-gatherer ancestry (Haak et al., 2015; Mathieson et al., 2015; Skoglund et al., 2014) with significant local differences. For example, some regions showed continuous population admixture (Nikitin et al., 2019; Betti et al., 2020), while others displayed almost total population replacement (Brace et al., 2019). In Denmark, there was a 1,000 year delay in the introduction of farming to the region (Allentoft et al., 2022). Moreover, east of a boundary

zone between the Black Sea and the Baltic, there was no shift in genetic ancestry until ~5 kya, in remarkable contrast to the western region and in congruence with archeological records (Allentoft et al., 2022).

At around 5 kya, ancestry associated with the Pontic-Caspian Steppe appeared abruptly in Europe (Allentoft et al., 2015; Haak et al., 2015) within a ~1,000 year timeframe; this ancestry is itself a mixture of at least two ancestries: EHG from the Middle Don River region and Caucasus hunter-gatherers (CHG) (Allentoft et al., 2022). Steppe ancestry has been associated with the Yamnaya culture and then with the expansion westwards of the Corded Ware Complex (characterised by cord-decorated ceramics) and later the Bell Beaker culture (characterised by bell-shaped grave goods), and the eastwards expansion in the form of the Afanasievo culture in the Altai region. However, these later groups are not necessarily direct descendents of the Yamnaya population (Heyd, 2017). The Steppe expansion is believed to have brought Indo-European languages into Europe (Allentoft et al., 2015; Haak et al., 2015), and again there were significant local regional differences. A male-biased migration of the Steppe ancestry from the Pontic Steppe during the Bronze Age was suggested for some of the Central European populations belonging to the Corded Ware groups (Goldberg et al. 2017a; Scorrano et al., 2021).

Thus present-day western Europeans can be modelled as mixtures of diverged genetic ancestry components that existed in the Mesolithic and Neolithic: Mesolithic hunter-gatherers (WHG, EHG, CHG), Neolithic farmers, and Steppe ancestry.

Great Britain has a distinct but related demographic history. In the period 10.5 - 6 kya all British Mesolithic individuals cluster with the Western Hunter Gatherer group, most closely resembling Lochsbour (Brace et al., 2019). In Britain, one of the furthest part of Europe from the Aegean origin of the migrating farmers (Broushaki et al., 2016), there had been universal agreement among archaeologists that there was a dramatic change around 6 kya with the introduction of farming, but the extent to which this was caused by cultural or demographic processes was uncertain (Sheridan, 2010). Given the isolation of Great Britain, and differing climate, it was thought the adoption of farming may have differed to the rest of Europe. Studies in the last few years have overwhelmingly indicated that the appearance of Neolithic practices was mediated by an immigration of farmers from continental Europe 6 kya, descended from Iberian Neolithic-related populations who were themselves descended from Aegean farmers who had followed a Mediterranean dispersal route (Brace et al., 2019).

Individuals from Wales retain the lowest levels of WHG admixture, followed by those from South-West and Central England; Neolithic individuals from Scotland and South-East England show higher levels (Figure 1, from Brace et al., 2019). This probably reflects differing degrees of admixture between farmers and local foragers, and multiple continental source populations which were variable in WHG ancestry themselves. Unlike in other parts of Europe, there was no detected increase in hunter-gatherer ancestry in the Neolithic populations of Britain (Brace et al., 2019).

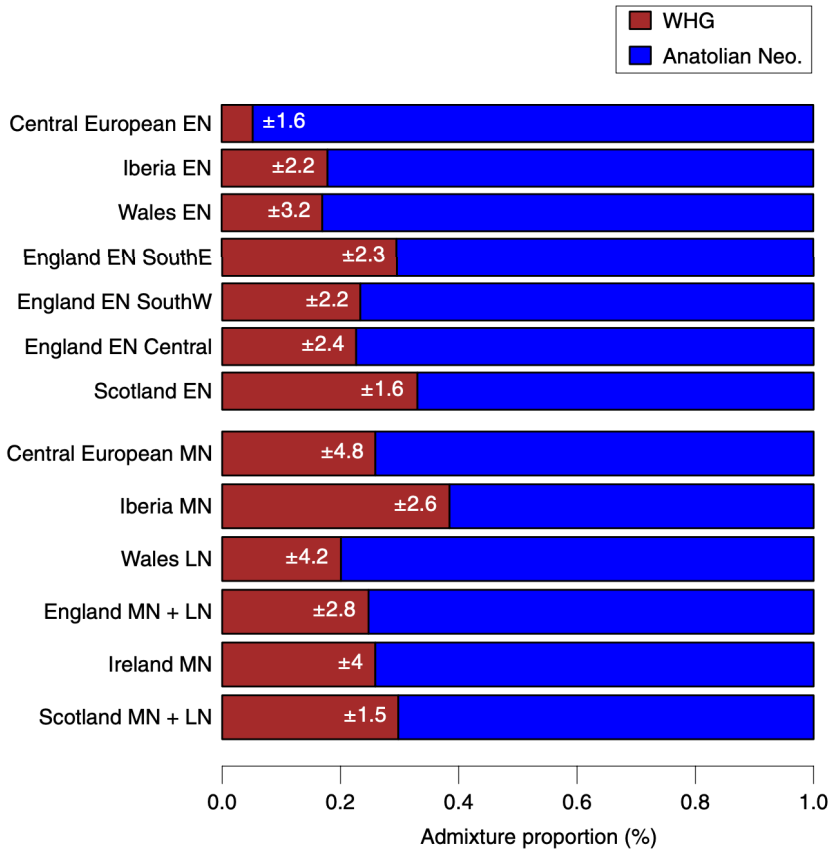


Figure 1 | WHG and Aegean Neolithic Farmer ancestry components of British, Irish and Continental European Neolithic populations.
Relative WHG and ANF ancestry in Early and Middle Neolithic British, Irish and continental European populations quantified by qpAdm. Percentages indicate error estimates computed by block jack-knifing with a block size of 5 centimorgans (cM). **From Brace et al. (2019).**

After 4.5 kya people associated with the Bell Beaker cultural complex arrived in Britain, carrying large amounts of Steppe-derived ancestry (Figure 2, from Olalde et al., 2018) and resulting in substantial replacement of the gene pool (Olalde et al., 2018). This pervasive Steppe-related ancestry, absent during the Neolithic period, remains predominant in Britain today (Haak et al., 2015). The Beaker-associated individuals from southern Britain are most closely related to the Beaker-associated individuals from Oostwoud (the Netherlands), who

had an almost identical proportion of Steppe-related ancestry, though this was variable; while not necessarily direct ancestors, this group was closely related to the population (perhaps so far unsampled) that moved into Britain from continental Europe (Olalde et al., 2018). Copper Age and Bronze Age individuals from Britain show qpADM mixture proportions from this group of between 59-100% (Olalde et al., 2018, supplementary information) when modelled as a mixture of Oostwoud and Neolithic British (Figure 3, from Olalde et al., 2018). After an initial period of variability, after roughly 4 kya individuals were more homogeneous and slightly increased in Neolithic-related ancestry. The results from Olalde et al. (2018) imply a minimum of 90 +/- 2% local population turnover by the Middle Bronze Age, although these results could also be explained by gene flow from continental Europe; thus for British individuals living during and after the Beaker period, a very high fraction of their DNA comes from continental Europe.

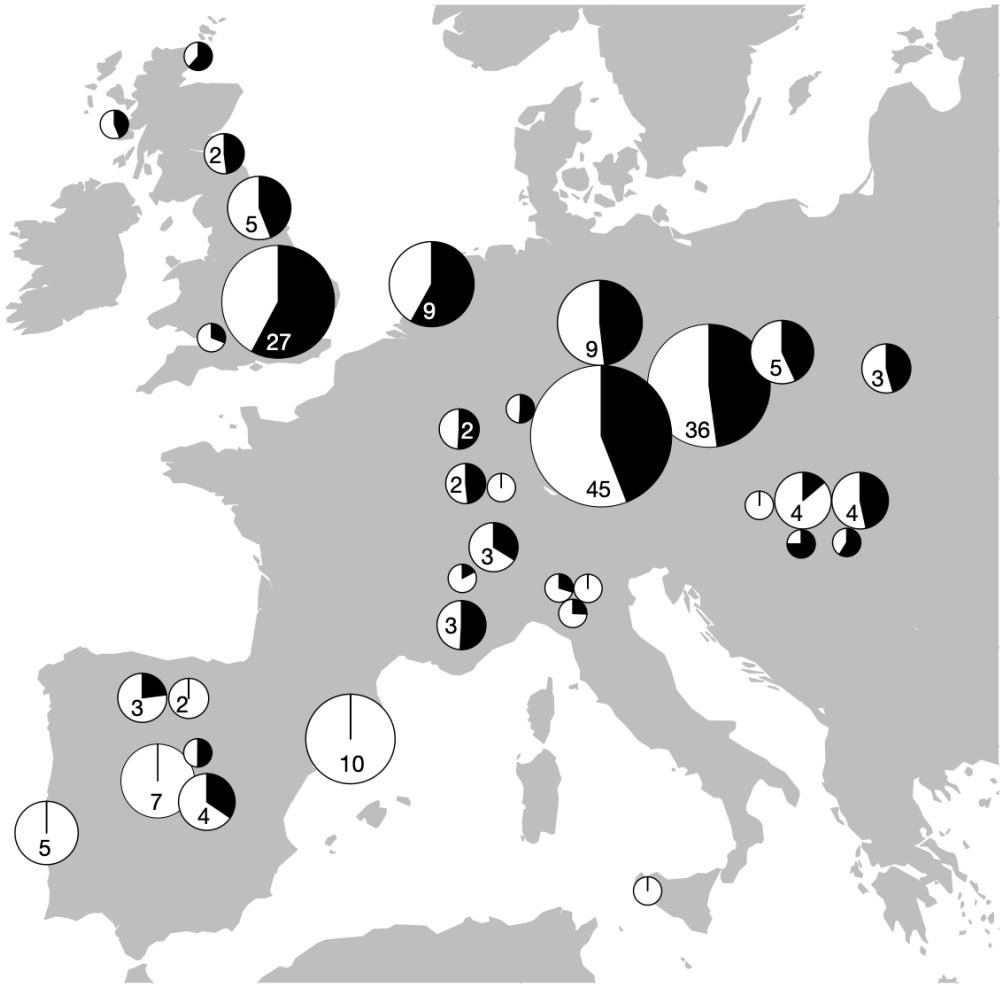


Figure 2 | Proportion of steppe-related ancestry (in black) in Beaker-complex-associated groups.

Computed with qpAdm under the model ‘Steppe_EBA + Anatolia_N + WHG’ (WHG, Mesolithic western European hunter-gatherers). The area of the pie is proportional to the number of individuals (number shown if more than one). **From Olalde et al. (2018).**

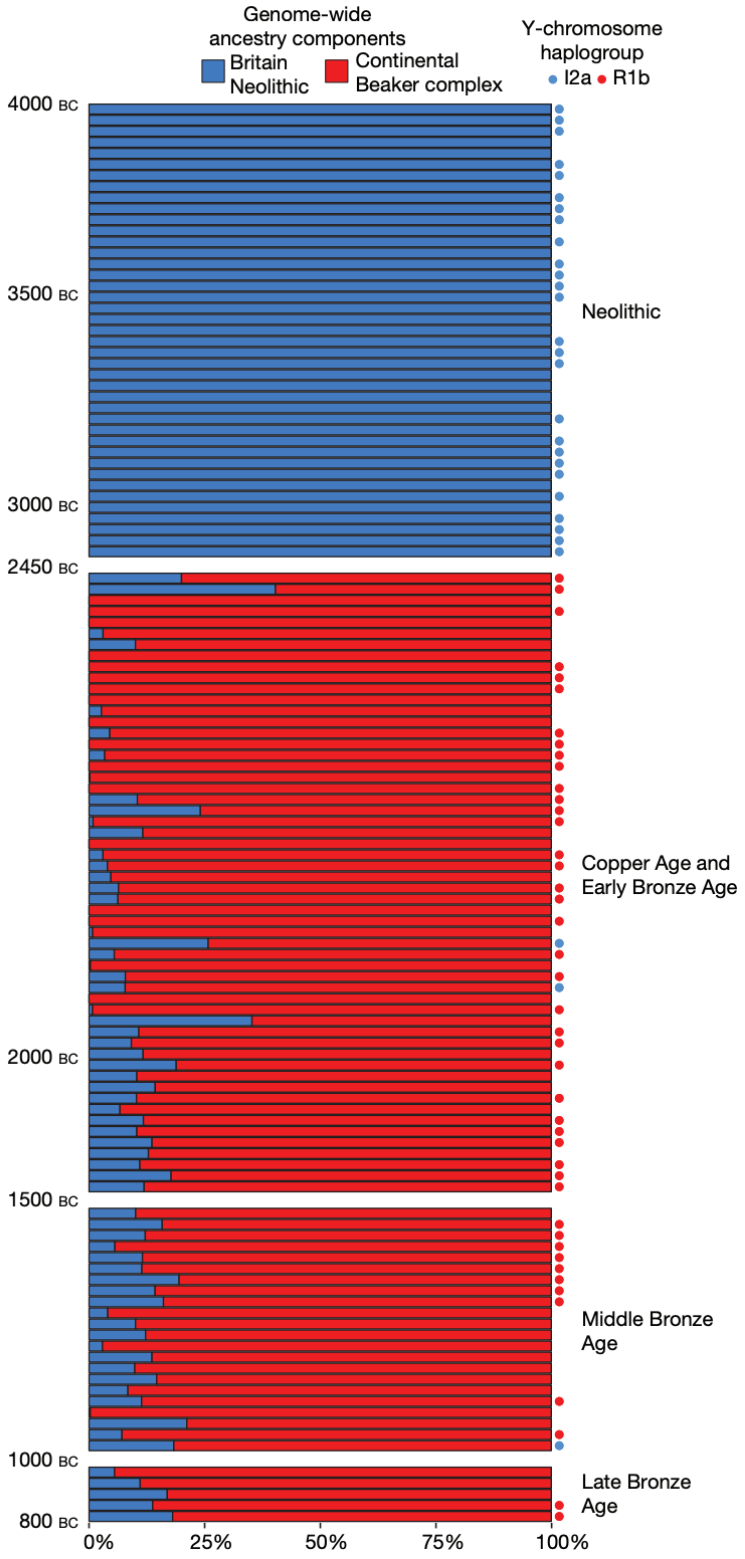


Figure 3 | Population transformation in Britain associated with the arrival of the Beaker complex.

Modelling Neolithic, Copper and Bronze Age (including Beaker-complex-associated) individuals from Britain as a mixture of continental Beaker-complex-associated individuals (red) and the Neolithic population from Britain (blue). Each bar represents genome- wide mixture proportions for one individual. Individuals are ordered chronologically and included in the plot if represented by more than 100,000 SNPs. Circles indicate the Y-chromosome haplogroup for male individuals. **From Olalde et al. (2018).**

In the Late Bronze Age, Iron Age and Anglo-Saxon period, there is evidence that migrations continued to play an important role in shaping the genetic make-up of British people, both within Britain (Martiniano et al., 2016), from northern European populations (Schiffels et al., 2016) and from further afield, such as the Middle East (Martiniano et al., 2016). A surge in Neolithic farmer-related ancestry during the late Bronze Age (1000 - 875 BC) in people in England and Wales led to the detection of a large-scale migration from the continent by people relatively enhanced in this ancestry, which may be linked to the spread of Celtic languages into Britain (Patterson et al., 2022). In the Iron Age there was reduced continental migration (Patterson et al., 2022), though the adoption of cultural practices from mainland Europe continued (Guggisberg, 2018). More recent targeted sampling of Medieval individuals has revealed a substantial increase in ancestry from continental northern Europe, which the authors explain by migration from the ‘continental North Sea zone’ - i.e. northern Netherlands to southern Sweden, with most samples in Lower Saxony and Denmark (Gretzinger et al., 2022). This study claimed to confirm the Anglo-Saxon ‘migration theory’ which had largely been rejected since the 1960s in favour of models of small numbers of elites moving to Britain from the continent accompanied by local acculturation.

Modern British population structure is limited (O’Dushlaine et al., 2010) but methodology developed based on haplotype-sharing patterns, fineSTRUCTURE, is able to detect fine-scale population structure. This approach revealed highly localised clustering, with many clusters occupying non-overlapping regions (Figure 4, from Leslie at al., 2015). These clusters are notable for differentiation over small distances and the stability of some clusters over very large distances – for example there is a large cluster that covers most of central and southern England that extends up the east coast, which remains intact even at the finest level of differentiation within fineSTRUCTURE. These clusters have been proposed to reflect geographical barriers, cultural barriers, and historical events such as invasions or migrations (Leslie et al., 2015; Margaryan et al., 2020).

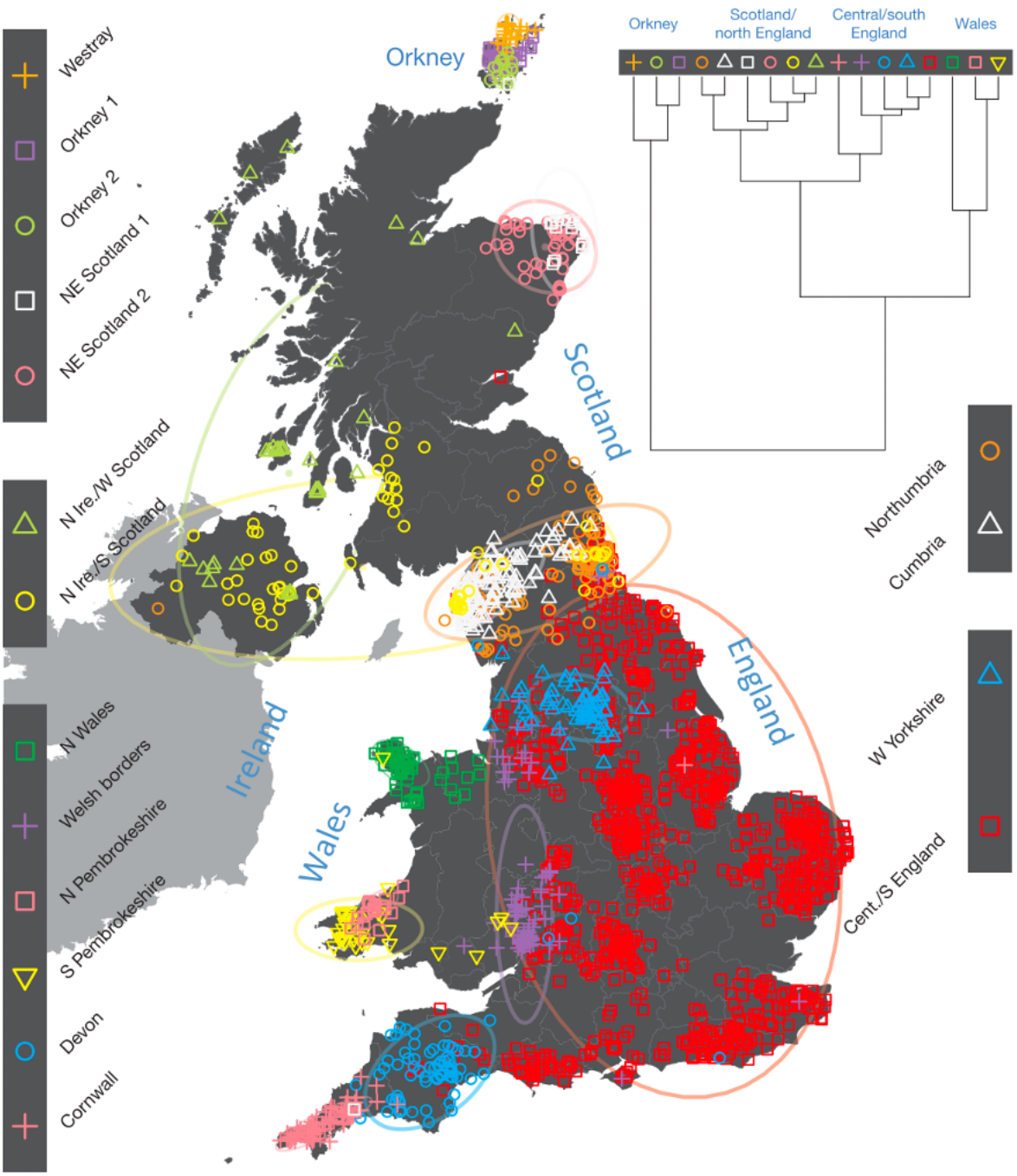


Figure 4 | Clustering of 2,039 UK individuals into 17 clusters based only on genetic data. For each individual, the coloured symbol representing the genetic cluster to which the individual is assigned is plotted at the centroid of their grandparents' birthplaces. Cluster names are in side-bars. The tree (top right) depicts the order of the hierarchical merging of clusters. **From Leslie et al. (2015).**

Tests for selection

A variety of methods have been developed to detect evidence of natural selection using population genetic data. Although many types of selection have been recognised, most research has focussed on positive selection - the favouring of an allele due to enhanced inclusive fitness resulting in an increase in frequency in the population over time - because it leaves strong signatures within the genome and because it is thought to be the primary mode of adaptation. The latter justification also means that many of the loci identified are associated with resistance to infectious disease (Fumagalli et al., 2011), or non-infectious diseases like autoimmune and metabolic disorders (Hancock et al., 2008). Furthermore, although earlier approaches focussed initially on a phenotype hypothesised to be adaptive and worked 'forwards' to identify the loci under selection, it is now possible to scan the entire genome for selective events (Vitti et al., 2013). Methods developed for interspecies comparisons are used to detect old and strong selective episodes that reflect macroevolutionary trends; here, the focus will be on intraspecies tests which can detect more recent selective events in the timeframe of human pre-history.

Positive selection which causes an allele to rise to high frequency rapidly also brings surrounding alleles to high frequency due to linkage disequilibrium (LD), resulting in a region of low diversity: the so-called selective sweep. This signal can persist for hundreds of thousands of years in the case of humans, until recombination and mutation restore genetic diversity. Many measures have been developed to detect this, such as Tajima's D, which measures an excess of rare alleles at a selective sweep which result from new mutations on a homogeneous background. A different set of measures exploit the fact that selection causes long-range LD, or long haplotypes, around the selected variant. These are particularly useful for detecting incomplete or soft sweeps. A statistic that is widely used is extended haplotype homozygosity (EHH) (Sabeti et al., 2002), which measures the LD from a core region with surrounding regions in a population. The longer and more common the haplotype, the stronger and more recent the putative selection. The integrated haplotype score (iHS) is the integral of the curve of EHH for the derived and ancestral variants travelling further away from the core region (Vitti et al., 2013).

It has been recognised that demographic events can mimic selection, though traditionally studies have tried to rule this out by comparing local signals with genome-wide data, under the assumption that demographic events would affect the whole genome equally (Vitti et al., 2013). However, this assumption must be questioned if selection is pervasive across the genome, as shown in *Drosophila* (Li and Stephan, 2006). Alternative approaches have

therefore explicitly modelled parameters such as population structure (e.g. Excoffier et al., 2009).

The incorporation of aDNA into tests for positive selection has increased the power of such tests, as time-series data allow the estimation of genetic diversity and population parameters before, during and after the selection event (Dehasque et al., 2020). Initially, this yielded insights into selection on small numbers of high-effect variants (Olalde et al., 2014; Allentoft et al., 2015; Mathieson et al., 2015; Jablonski et al., 2017). However, most phenotypes are polygenic in nature (Hill, 2001; Visscher et al., 2017; Martin et al., 2019), and more recent work has focussed on assessing polygenic adaptation: small changes in allele frequency at many sites across the genome (Irving-Pease et al., 2021). These tests face multiple difficulties.

The first challenge in detecting polygenic adaptation is in identifying the variants associated with the trait. Genome-wide association studies (GWAS) are used to find statistical associations between variants and traits of interest, while controlling for variables such as age, sex, socio-demographic status, and ‘ancestry’ in the form of principal components (PCs). These statistically associated variants can then be either analysed jointly to test for collective directional changes in allele frequency (e.g. Stern et al., 2021) or used to construct PRS for ancient populations, in which each variant contributes to an additive score for the genetic contribution to a continuous trait or probability of a binary trait (Irving-Pease et al., 2021).

However, GWAS studies usually only explain a small proportion of the heritability of a trait estimated from twin studies (the so-called ‘missing heritability’ problem) (Maniolo et al., 2009), suffer from widely-documented problems with lack of portability between populations (Duncan et al., 2019), fail to tag true biological signals (Boyle, Li & Pritchard, 2017), and exhibit variable prediction accuracy even within ancestry groups due to the socio-economic status, age or sex of the individuals in which the GWAS and the prediction were conducted, as well as on the GWAS design (Mostafavi et al., 2020).

The main explanations for the ‘missing heritability’ in GWAS studies have historically been (1) small sample sizes mean there is insufficient power to detect all associations (Tam et al., 2019); (2) strict p-value thresholds for significance of association to account for multiple testing, therefore missing SNPs of modest effect. This is traditionally accounted for using a Bonferroni correction maintaining the genome-wide false positive rate of 5% assuming 1 million independent tests, therefore requiring a threshold of $p < 5e\text{-}8$ (Tam et al., 2019); and

(3) twin studies have overestimated heritability (Young, 2019). However, the missing heritability problem may not be as problematic as previously thought: research has shown that when the sample size is large enough it is possible to account for nearly all common SNP-based heritability for a complex trait within an ancestry group, in this case height in Europeans (Yengo et al., 2022).

While the missing heritability problem may be solvable, there are other reasons for poor prediction accuracy, both within similar populations and when exporting significance values and effect sizes to other populations. The failure to tag a true biological signal rather than a tag SNP in high LD with the true causal variant means differing LD structures in populations will reduce the accuracy of a PRS. To combat this, many studies now employ 'fine mapping', a technique to identify causal variants even when multiple causal variants are in high LD (e.g. Gaulton et al., 2015). Common genetic variants are likely to be old and shared across ancestries. Although differing patterns of LD and allele frequencies between populations contribute to poor portability if causal SNPs are not identified, increasingly trans-ethnic cohorts are being used which can aid in fine mapping causal variants (Visscher et al., 2017; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, 2014) which should help with portability. Furthermore, recent methods can take account of the underlying LD structure when generating PRS, such as LDpred (Vilhjálmsson et al., 2015) and LDpred2 (Privé et al., 2020), rather than performing the traditional LD-based marker pruning and p-value thresholding, which discards information.

Poor GWAS design can result in population stratification leading to spurious signals of association driven by covariates that have not been sufficiently controlled for. This is particularly a problem as cohorts get larger, when it is almost impossible to perfectly match case and control cohorts (McClellan & King, 2010). Ancestry is included as a covariate to account for this because it is correlated with multiple measures of cultural and socio-demographic factors as well as differences in allele frequencies. However, assessment of ancestry is usually self-assigned and then confirmed using Principal Component Analysis (PCA), with the first *n* PCs used as covariates (Duncan et al., 2019). This is often insufficient to account for population stratification (Liu et al., 2013). Because the ancestry of a genome differs along its length it is expected that genome-wide assessment of ancestry will not capture all possible information. For example, until recently the strongest evidence for polygenic adaptation had come from European PRS for height, based on the GIANT consortium; however, it was shown that these effects are strongly attenuated or absent when using analysis based on the UK Biobank, and could be explained by insufficient methods for correction for population stratification in GWAS (Berg et al., 2019a; Sohail et al., 2019).

Potential solutions to population stratification involve linear mixed models (LMMs), although these are computationally intensive (Liu et al., 2013). Efforts have also been made to begin to use the local ancestry at each SNP as a covariate in GWAS studies (Atkinson et al., 2021), though this relies on accurate local ancestry inference and has not been attempted using genetically closer ancestries.

The second major challenge in using aDNA for testing for positive selection derives from calculating PRS for ancient populations using GWAS summary statistics in modern populations, and therefore exporting the effect sizes backwards in time (reviewed in Irving-Pease et al., 2021). Ancient populations have differing LD patterns compared to the modern populations in the GWAS studies, so if tag SNPs are used rather than true causal variants this will cause problems. However, even assuming that the true signal has been found, SNP effects are known to be dependent on both the physical environment (gene-environment interactions) (Johnsen et al., 2021) and genomic environment (gene-gene interactions, epistasis) (Visscher et al., 2017) and therefore unlikely to export well backwards in time. Despite these criticisms, this has been a popular approach for tracing the changes in the genetic scores for a phenotype over time (Allentoft et al., 2015; Haak et al., 2015; Mathieson et al., 2015; Esteller-Cucala et al., 2020; Ju & Mathieson, 2021), for example testing for over-dispersion of these scores among populations compared to a null model under drift (Berg and Coop, 2014), which was subsequently extended to work with admixture graphs (Racimo et al., 2018).

Of particular focus has been standing height, which shows evidence of negative selection in modern populations (O'Connor et al., 2019) and has been the subject of multiple studies using aDNA (Mathieson et al., 2015; Martiniano et al., 2017; Cox et al., 2019). A benefit of studying this trait is that PRS predictions can be compared to height estimates based on skeletal remains, which has shown that these predictions offer approximately one quarter the predictive power compared to PRS in modern populations (Cox et al., 2022).

Given these limitations, it is unclear how to interpret PRS for ancient individuals, or even to compare ancient populations, given that within group variance in PRS is often higher than between group variance (Irving-Pease et al., 2021). Most traits are inherently unverifiable, and gene-environment interactions mean expression of a trait may have differed substantially in the past.

A third challenge in detecting polygenic adaptation using aDNA is presented by population migration and admixture in the past, which is pervasive (reviewed in Pickrell and Reich,

2014; Leonardi et al., 2017; Skoglund and Mathieson, 2018; Orlando et al., 2021). Where the admixing populations have different allele frequencies, sampling of the populations pre- and post-admixture can result in an apparent shift in frequency, i.e. directional selection. Explicitly modelling local ancestry across the genome can negate this phenomenon, and has increased power to detect patterns of selection in admixed populations: for example in Cabo Verde (European and West African sources) (Hamid et al., 2021) and Malagasy (Asian and African sources) (Pierron et al., 2018) populations. However, this approach has not been attempted incorporating aDNA, and has only included very diverged ancestries for which local ancestry inference is relatively straightforward.

The work presented in this thesis consists of two main steps: firstly, accurate local ancestry inference in a large modern panel using aDNA reference populations; and secondly, using this data to infer ancestral contributions to modern complex phenotypes. It therefore attempts to overcome challenges two and three outlined above - firstly by avoiding exporting risk scores over space and time, making no explicit claims about the phenotypes of ancient individuals or populations, and secondly by using local ancestry information to control for admixture when looking for signals of selection. The first challenge outlined above (inherent to GWAS studies) is more difficult to overcome, but with increasing study sizes and more advanced computational methods it can be minimised.

Local ancestry inference

It is beyond the scope of this thesis to report a full overview of local ancestry inference (LAI) methods, but a brief overview will be given with particular emphasis on the methods used in this work. To do this, coalescent theory and ancestral recombination graphs are introduced, with an outline of some computational approaches for their inference. I will then discuss some applications of LAI in admixture mapping and GWAS design, both areas related to work in this thesis.

After two populations mix, recombination breaks down the admixture tracts (haplotypes) each generation to produce a mosaic of ancestry across the genome. Over time these haplotypes get shorter, but theoretically every locus can be assigned to one of the two ancestries. Because this ancestry is inherently unobservable, it has to be inferred. This assignment is the aim of local ancestry inference methods, a class of algorithms that experienced a publication peak in the period 2006-2016 (Wu et al., 2021).

Typically, LAI methods use putative ancestral populations as ‘sources’ or ‘references’, which may either be inferred or explicitly specified. Originally, allele frequency differences at certain markers between source populations were used to infer ancestry segments. If a so-called ancestry informative marker (AIM) was at frequency 100% in one source and 0% in another, its presence in an admixed individual could be used to infer ancestry back to the first source. This method was more feasible when source populations have admixed relatively recently, as is the case in African-Americans (~8 generations), but as admixture times move backwards the number of markers required to infer all ancestry segments increases, so that twice as many are required for Latino Amercians (admixture ~16 generations ago) as for African Americans (Shriner, 2013). Because there are so few *fully* informative markers between source populations (i.e. markers with frequency 0% and 100%), it is necessary to include markers with multiple alleles in each source population. Thus each marker becomes probabilistic rather than deterministic. This probabilistic modelling is the basis of modern LAI algorithms.

The gold standard of local ancestry inference is the ancestral recombination graph (ARG) (Griffiths, 1991), which represents the full genetic history of a set of related samples: it defines a set of recombination points (and therefore haplotypes), mutation points, and genealogies for each locus in the genome. Recombination events in between loci will cause changes in their genealogies, therefore each locus has a local (or marginal) tree annotated with mutations, which may or may not be shared with neighbouring positions depending on

the historical recombination patterns. It is therefore possible to determine the local ancestry of an allele by examining its history via the ARG by inferring from which ancient population it ‘moves through’ backwards in time.

As we trace a given set of samples backwards in time at a particular genetic locus, samples will eventually coalesce into shared ancestral lineages, until all samples have coalesced; this is the basis of coalescent theory. Coalescent theory was conceptualised in the the 1980s primarily by Kingman (Kingman 1982a, Kingman 1982b, Kingman 1982c) and also Hudson (Hudson, 1983a) and Tajima (1983) building on the Wright–Fisher model (Sewall-Wright, 1931) and work by Malecot in the 1940s (Wakeley, 2009). It was subsequently extended to incorporate population structure (Donnelly and Tavaré, 1987), migration and recombination (Wakeley, 1996), and natural selection (Nordberg et al., 2001). Coalescent theory is a framework for how and when the sampled alleles coalesce: the patterns of coalescences depend on all of the factors listed above.

Many ARG inference methods leverage an underlying coalescent model in which the probability of observing the real data can be calculated under varying sample histories. This can be used to, for example, sample from the posterior distributions of ARGs (Rasmussen et al., 2014). The space of plausible sample histories conditioned on the observed data is often very large, leading to computational difficulties (Speidel et al., 2019); however, this approach has been successfully applied in specific circumstances, outlined below.

Although the coalescent process can be viewed as a stochastic process over time, it can also be viewed as stochastic along the length of the genome (Wiuf & Hein, 1999), in which ARGs are simplified as sequences of local trees along the genome, which captures nearly all the information in a full ARG (Rasmussen et al., 2014; Hubisz & Siepel, 2020). The sequence of genealogies along the genome are then considered to be approximately Markovian (Wiuf & Hein, 1999), which is known as the Sequentially Markov Coalescent (SMC) (McVean & Cardin, 2005). This insight allows the employment of the well-understood algorithms for hidden Markov models (HMMs). The simplest of these is perhaps the pairwise SMC (PSMC) (Li and Durbin, 2011), used to infer effective population sizes over time: when considering just two homologous chromosomes, there is only one tree topology at each locus, and time is discretized to reduce the state space for the Markov chain (consisting of all possible genealogies); using the resulting n-state HMM (where n is the number of time slices), it is possible to integrate over all possible ARGs to perform the inference (Ramussen et al., 2014). However, this requires a complete characterisation of the SMC state-space, which even in discretized time units is not currently possible for more than a small number of

samples. Later methods such as ARGweaver (Ramussen et al., 2014) improved on this, enabling inference of full ARGs for dozens of individuals by considering multiple sequences simultaneously, and subsequent sampling from the posterior distribution of ARGs, under a set of modelling assumptions, to infer the most likely ARG given the observed genetic data (Hubisz & Siepel, 2020). ARGweaver is considered to be state-of-the-art but computationally intensive, limited to tens of samples.

When full ARG inference is not a requirement, a popular alternative LAI approach uses the output of the Li and Stephens copying model (Li & Stephens, 2003), motivated primarily by computational considerations. In this model, which is itself an HMM, a sequence is reconstructed as noisy 'copies' of other sequences in the sample, with the source for the copied fragment representing the genealogically closest sequence present (i.e. the nearest neighbour in a tree). The noise allows for mutations in between divergence and sampling. The output of this algorithm - often termed a 'painting' of a chromosome - can be calculated over an infinite number of paintings to estimate the probability that a given haplotype acts as a donor to another as a function of position along the genome. This approach has been widely used for haplotype phasing, genotype imputation and LAI (e.g. HAPMIX, Price et al., 2009). However, because it does not reconstruct the full ARG it is more limited in its information and inferences (Rasmussen et al., 2014).

More recent advances have used the Li and Stephens model to infer full ARGs. For example Relate (Speidel et al., 2019) uses the Li and Stephens algorithm to generate locus-specific distance matrices at intervals along the genome, which are then fed to a tree builder under an approximate coalescent model, estimating coalescence times and effective population times. This method is more than four times faster than ARGweaver (Speidel et al., 2019), though the latter remains more accurate. Alternatively, *tsinfer* (Kelleher et al., 2019) employs a method in which haplotypes are painted using the Li and Stephens HMM from inferred ancestral sequences, sequentially building up a copying path for each segment in the focal chromosome. This is possible on a Biobank scale, aided by more efficient storage formats for tree sequences.

Overall, approximately 70% of LAI methods use HMMs to infer local ancestry probabilistically, where the hidden states correspond to local ancestry along the length of the genome and are used to generate the observed genotypes (Wu et al., 2021). These HMM-based algorithms can be split into two types: the first relies on allele frequencies for each of the source populations (e.g. LAMP (Sankararaman et al., 2008) which employs a sliding window approach); the second uses haplotype information in the source populations

(e.g. HAPMIX (Price et al., 2009) and Chromopainter (Lawson et al., 2012)). The latter contains more information than allele frequencies, and can therefore distinguish genetically closer populations (Lawson et al., 2012). Other LAI approaches include the use of PCA (e.g. Price et al., 2006, PCAdmix (Brisbin et al., 2012)) or machine learning classification tools (e.g. RF-Mix (Maples et al., 2013)).

There are therefore numerous methods for LAI, both with and without inferring a full ARG. Some of these now allow the incorporation of aDNA data directly, enabling better specification of ancestral populations rather than using inferred donor groups or modern surrogates as donors, which are often only distantly related to the real ancient sources. As these LAI methods progress ancestry assignment will improve. LAI has been used to study population evolution, natural selection, and to map disease associations.

Admixture mapping is one such application of LAI, designed to utilise recent admixture to test the correlation between local ancestry and a given phenotype, and can be performed using generalised linear models which allow for different phenotypic data (i.e. continuous and discrete) and the inclusion of covariates and interaction terms (Shriner, 2013). Admixture mapping is used to assess differential risk by ancestry, which is useful for inferences about ancestry-based health disparities and ancestry-specific evolution. For example, if a locus has excess ancestry in cases but not controls, this can be used to infer the presence of a disease locus; this disease allele will be at higher frequency in the ancestry displaying an excess. It is also used to reduce the set of credible variants driving an association signal by leveraging the different source populations' different LD patterns; this is more effective the shorter the admixture tracts are.

Some recent use cases of admixture mapping include studies on BMI and Type 2 diabetes in African Americans (Wu et al., 2022), breast cancer risk in US Latinas (Fejerman et al., 2012), and multiple sclerosis in African Americans, Asian Americans and Hispanics (Chi et al., 2019). Broadly, admixture mapping has been successful for phenotypes that are highly stratified between ethnic groups, but is subject to false positives where local ancestry increases are unrelated to the phenotype of interest and cannot be used for phenotypes which are observed at similar rates across groups.

Admixture mapping relies on accurate LAI, and has therefore mostly been applied to recent admixture events between distantly related populations. The older the date of admixture, the higher the resolution of admixture mapping in locating a disease allele but the more difficult the LAI. As LAI methods have improved and DNA from ancient samples becomes available,

it has become possible to elucidate the local ancestry of more closely related populations and older admixture events. This is a promising field of research.

Association studies, which test genotype-phenotype correlations rather than ancestry-phenotype correlations, are premised on the presence of similar allele frequencies between ancestries. Where this assumption does not hold false positives are expected, because differences in allele frequencies between cases and controls is due to differences in ancestry rather than associations of the alleles with a disease; this phenomenon is termed population stratification. Most studies aim to combat this by removing admixed individuals and controlling for confounding due to population stratification, the latter usually by including principal components as covariates. However, this has the effect of (1) significantly reducing the power of GWAS studies as the numbers of individuals included is lowered; (2) biassing studies in favour of larger (usually white) ethnic populations and therefore exacerbating existing health disparities (Martin et al., 2019); and (3) failing to effectively control for the confounding effect because PCs capture well genome-wide ('global') ancestry fractions but not 'local' ancestry - there could still be differences in local ancestry between cases and controls even if their genome-wide ancestry is identical, leading to false positives. There have therefore been recent advances in incorporating admixed individuals in GWAS studies, and in using local ancestry as a covariate at the genotype level, producing ancestry-specific effect sizes and p-values (Atkinson et al., 2021). This approach increases power, expands the number of populations studied, and improves signal localisation.

Furthermore effect sizes may be ancestry-specific, either due to (1) differential tagging of a true effect by the tag SNP because of differing LD patterns; (2) different rare variants in different populations; or (3) gene-gene interactions (epistasis) depending on the ancestral background of a variant. This means that polygenic risk scores (PRS), which are anticipated to be widely used in some precision medicine in the very near future (Polygenic Risk Score Task Force of the International Common Disease Alliance, 2021), have lower accuracy in admixed individuals or individuals of a different ancestry from the GWAS study (Marnetto et al., 2020; Ding et al., 2022). There has been some work in developing methods for generating PRS for admixed individuals using local ancestry deconvolution, but again this has only been applied to the extreme scenario of recent admixture between very diverged populations, i.e. individuals of recent joint African and West Eurasian descent (Marnetto et al., 2020; Bitarello & Mathieson, 2020). This approach has improved trait predictability for admixed individuals, but is limited by poor ancestry-specific effect size estimates. As outlined above, the estimation of ancestry-specific effect sizes is improving, including through the leveraging of admixed individuals (Atkinson et al., 2021).

To my knowledge, large-scale LAI, admixture mapping, and the inclusion of local ancestry in admixed GWAS studies have only been performed using individuals of relatively recent admixture, predominantly African Americans, rather than using local ancestry inferred for non recently admixed individuals, or for more closely related ancestries like those found in modern Europeans. To do so would require accurate LAI methods, large panels of modern admixed samples, and a sufficiently large ancient panel with representatives of each ancient ancestry to perform the LAI. However, the benefits of such an undertaking would be significant: elucidating ancestry-specific risk at loci associated with a phenotype, including subsequent conclusions about the evolution of phenotypes in past populations; increased accuracy of PRS in admixed individuals; fine-mapping of variants using differing LD patterns between ancestries; and better control of population structure and therefore avoidance of false positives in GWAS studies. Each of these has the potential to significantly extend the use and application of very large modern genome panels like the UK Biobank, enabling the inclusion of the ~100,000 admixed individuals in that panel. This approach is now feasible in western Europe thanks to advances in computational LAI techniques, very large panels of ancient and modern genomes, a detailed understanding of the past demography of the continent and therefore the ancestries under consideration, and large GWAS studies which are able to accurately infer many small effects.

Approach and Datasets

The work in Chapters One, Two, Three and Four is based on a painting of the UK Biobank using a large panel of ancient genomes partitioned into ancestral populations: Western Hunter-Gatherer, Eastern Hunter-Gatherer, Caucasus Hunter-Gatherer, FarmerAnatolia, FarmerEarly, FarmerMiddle, FarmerLate, Yamnaya (Steppe), African and East Asian. This panel is designed to capture all of the major ancestral components from the Mesolithic and Neolithic contributing to a modern European population, as outlined in this Introduction. Significant work was done to make this computationally feasible on a Biobank scale, and to efficiently store the local ancestry painting probabilities: ~425,000 individuals painted at ~550,000 sites for ten ancestries results in ~2.3375e+12 painting probabilities.

I have described the motivations for generating such a dataset in this Introductory chapter, not all of which are capitalised on here but which I hope will be in the future. Briefly, these motivations include the ability to trace genome-wide ancestry components in modern populations, a question of historical interest; the ability to better control for population stratification in GWAS studies, where the inclusion of aDNA has the potential to significantly improve recent attempts at this, both with genome-wide components and at the local ancestry level; the estimation of differing risk by ancestry for any phenotype where there is sufficient understanding of the underlying risk variants, and the consequent conclusions about the evolution of specific variants and specific phenotypes, including tracing their origins to one or more ancestral populations; and relatedly, the detection of selection through the excess of an ancestry at a locus of interest. Because the ancestral populations are here explicit and relatively well studied, and in collaboration with archaeologists and others, it is also possible to draw stronger (and more interesting) conclusions about the possible causes of the findings, including linking them to lifestyle and cultural changes, pathogenic disease exposure, or environmental factors such as sunlight or temperature.

An outline of the thesis: Chapter One deals with the painting process, including simulations to test the accuracy of the inferred local ancestry; it includes results from painting the ‘white British’ individuals in the UK Biobank, including how various ancestry components vary geographically across Great Britain. Chapter Two describes new methods to select individuals in the UK Biobank not born in Britain but of a ‘typical ancestral background’ for their country of birth, based on density-based clustering of PCs; it includes results describing ancestry variation across other countries in Eurasia and north Africa, as well as new relationships between some PCs and the defined ancestry components. Chapter Three describes methods to estimate aggregate ancestral contributions to modern polygenic

phenotypes, utilising the local ancestry paintings and GWAS summary statistics, in a new statistic analogous to a PRS. Chapter Four describes the application of this dataset to the specific question of the origin of the genetic risk for multiple sclerosis, which is traced back to the Bronze Age Steppe.

Chapter One: ChromoPainting the UK Biobank

Preface

Much of the contents of this chapter was previously published as (§ denotes joint first authors, @ denotes joint last authors):

Population Genomics of Stone Age Eurasia

Morten E. Allentoft§, Martin Sikora§, Alba Refoyo-Martínez§, Evan K. Irving-Pease§, Anders Fischer§, William Barrie§, Andrés Ingason§, Jesper Stenderup, Karl-Göran Sjögren, Alice Pearson, Bárbara Sousa da Mota, Bettina Schulz Paulsson, Alma Halgren, Ruairidh Macleod, Marie Louise Schjellerup Jørkov, Fabrice Demeter, Maria Novosolov, Lasse Sørensen, Poul Otto Nielsen, Rasmus H.A. Henriksen, Tharsika Vimala, Hugh McColl, Ashot Margaryan, Melissa Ilardo, Andrew Vaughn, Morten Fischer Mortensen, Anne Birgitte Nielsen, Mikkel Ulfeldt Hede, Peter Rasmussen, Lasse Vinner, Gabriel Renaud, Aaron Stern, Theis Zetner Trolle Jensen, Niels Nørkjær Johannsen, Gabriele Scorrano, Hannes Schroeder, Per Lysdahl, Abigail Daisy Ramsøe, Andrei Skorobogatov, Andrew Joseph Schork, Anders Rosengren, Anthony Ruter, Alan Outram, Aleksey A. Timoshenko, Alexandra Buzhilova, Alfredo Coppa, Alisa Zubova, Ana Maria Silva, Anders J. Hansen, Andrey Gromov, Andrey Logvin, Anne Birgitte Gotfredsen, Bjarne Henning Nielsen, Borja González-Rabanal, Carles Lalueza-Fox, Catriona J. McKenzie, Charleen Gaunitz, Concepción Blasco, Corina Liesau, Cristina Martinez-Labarga, Dmitri V. Pozdnyakov, David Cuenca-Solana, David O. Lordkipanidze, Dmitri En'shin, Domingo C. Salazar-García, T. Douglas Price, Dušan Borić, Elena Kostyleva, Elizaveta V. Veselovskaya, Emma R. Usmanova, Enrico Cappellini, Erik Brinch Petersen, Esben Kannegaard, Francesca Radina, Fulya Eylem Yediay, Henri Duday, Igor Gutiérrez-Zugasti, Inna Potekhina, Irina Shevnina, Isin Altinkaya, Jean Guilaine, Jesper Hansen, Joan Emili Aura Tortosa, João Zilhão, Jorge Vega, Kristoffer Buck Pedersen, Krzysztof Tunia, Lei Zhao, Liudmila N. Mylnikova, Lars Larsson, Laure Metz, Levon Yepiskoposyan, Lisbeth Pedersen, Lucia Sarti, Ludovic Orlando, Ludovic Slimak, Lutz Klassen, Malou Blank, Manuel González-Morales, Mara Silvestrini, Maria Vretemark, Marina S. Nesterova, Marina Rykun, Mario Federico Rolfo, Marzena Szmyt, Marcin Przybyła, Mauro Calattini, Mikhail Sablin, Miluše Dobisíková, Morten Meldgaard, Morten Johansen, Natalia Berezina, Nick Card, Nikolai A. Saveliev, Olga Poshekhonova, Olga Rickards, Olga V. Lozovskaya, Olivér Gábor, Otto Christian Uldum, Paola Aurino, Pavel Kosintsev, Patrice Courtaud, Patricia Ríos, Peder Mortensen, Per Lotz, Per Persson, Pernille Bangsgaard, Peter de Barros Damgaard, Peter Vang Petersen, Pilar Prieto Martinez, Piotr Włodarczak, Roman V. Smolyaninov, Rikke Maring, Roberto Menduiña, Ruben Badalyan, Rune Iversen, Ruslan Turin, Sergey Vasilyiev, Sidsel Wåhlin, Svetlana Borutskaya, Svetlana Skochina, Søren Anker Sørensen, Søren H. Andersen,

Thomas Jørgensen, Yuri B. Serikov, Vyacheslav I. Molodin, Vaclav Smrcka, Victor Merz, Vivek Appadurai, Vyacheslav Moiseyev, Yvonne Magnusson, Kurt H. Kjær, Niels Lynnerup, Daniel J. Lawson, Peter H. Sudmant, Simon Rasmussen, Thorfinn Korneliussen@, Richard Durbin@, Rasmus Nielsen@, Olivier Delaneau@, Thomas Werge@, Fernando Racimo@, Kristian Kristiansen@, Eske Willerslev@
bioRxiv 2022.05.04.490594; doi: https://doi.org/10.1101/2022.05.04.490594
In review at Nature, January 2023.
It has been modified to fit the style of a dissertation.

I performed all analyses described here, apart from the data generation described in ‘Methods: Data’. I wrote all the text except for ‘Introduction: Population Genomics Of Stone Age Eurasia (from Allentoft et al., 2022)’ and ‘Methods: Data’. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

Chapter summary

In this chapter I summarise my contribution to this work. I give a general introduction to ChromoPainter (Lawson et al., 2012) and western Eurasian demography. I then test the accuracy of ChromoPainter to assign local ancestry using simulations of Eurasian demography. I develop new methods to use Chromopainter on a biobank scale to ‘paint’ modern genomes from the UK Biobank (UKB) using ancient genomes, grouped into reference populations, as donors. Painting was done following the pipeline of Margaryan et al. (2020) based on GLOBETROTTER (Hellenthal et al., 2014), and admixture proportions were estimated using Non-Negative Least squares. I stored both genome-wide and local ancestry (i.e. per variant per individual) results. In this chapter I report genome-wide ancestry proportion gradients within Britain, and discuss the implications for using apparently ancestrally ‘homogenous’ cohorts like the UKB.

Introduction

ChromoPainter using ancient DNA reference panel

ChromoPainter (Lawson et al., 2012) uses an approach premised on the observation that markers on the same chromosome are inherited together unless separated by recombination; at the population level, this results in linkage disequilibrium (LD) between close markers that reflect a shared history of descent. The haplotype-based algorithm of ChromoPainter aims to harness this information, detecting shared haplotypes to reconstruct phased recipient genomes as chunks ‘copied’ from donors.

Considering the genealogy of a single locus, we can identify one or more closest relatives to that locus, henceforth called ‘nearest neighbours’; if viewed as a genealogy, these are the other leaves of the tree underneath the first coalescence. Therefore at each locus of each haplotype, there exists one or more nearest neighbours. ChromoPainter aims to identify these using an approximate method based on that introduced by Li and Stephens (2003): the Hidden Markov Model (HMM), which explicitly reconstructs the haplotype of a recipient/target individual as a series of chunks of genetic material donated by the other donor/reference individuals, using information on the types of the recipient and potential donor at each SNP. This approach is probabilistic, calculating the expectations of which haplotype acts as donor to a recipient as a function of position over an infinite number of paintings (Lawson et al., 2012). Each contiguous donor chunk can be viewed as a single unit of co-inheritance between the donor and recipient, in the same way that two samples might share a SNP. Although ChromoPainter was originally intended to use this information, in the form of a ‘co-ancestry matrix’ (i.e. the number of chunks donated between individuals), to ascertain fine-scale population structure and clustering (in the fineSTRUCTURE software package), the software can also be used with pre-defined donor and recipient populations, and local copying information as well as genome-wide scores for each target haplotype can be recorded.

If the donor panel is formed of ancient samples grouped into ‘donor populations’ and the recipient individual is modern, the nearest neighbour will reflect the history of that locus. Here, I use nearest neighbour as a proxy for local ancestry - i.e. which population that haplotype passed through (which may not be a single unique population from the donor panel). There are three factors which are expected to affect the accuracy of local ancestry inference: the number of individuals in a donor population, the diversity of the donor panel, and the age of donor samples. Firstly, a smaller number of haplotypes will capture less of the

genetic/haplotypic diversity of a population and therefore result in lower accuracy of inference. Secondly, donor populations which are more genetically differentiated will be easier to assign, increasing accuracy; conversely, more genetically similar populations and the algorithm will find it difficult to correctly identify the nearest neighbour(s). And thirdly there is the issue of ‘masking’, whereby haplotypes from older donor populations have travelled through more recent donor populations before arriving in the modern population; this causes a painting bias towards the more recent ancient populations, though the haplotype should still be painted as both populations. I discuss the implications of this age-related bias below.

Population Genomics Of Stone Age Eurasia (from Allentoft et al., 2022)

It is argued that genetic diversity in contemporary western Eurasian human populations was largely shaped by three major migrations in the Stone Age: hunter-gatherers occupying the area since c. 45,000 BP; the Neolithic farmers expanding from the Middle and Near East c. 11,000 BP; and Steppe pastoralists coming out of the Pontic steppe c. 5,000 BP, signalling the final stages of the Stone Age and the beginning of the Bronze Age (Allentoft et al., 2015; Haak et al., 2015; Cassidy et al., 2016; Jones et al., 2017; Martiniano et al., 2017). However, due to a paucity of genomic data from skeletons older than 8 ka, knowledge of the population structure in the Mesolithic period and how it was formed is limited, and compromises our ability to understand the subsequent demographic transitions. Also, most ancient DNA (aDNA) studies have thus far been restricted to individuals from Europe, hampering our ability to understand the wider impact of these events. The spatiotemporal mapping of population dynamics east of Europe, including Siberia, Central- and Northern Asia during the same time period is limited. In these regions the local use of the term ‘Neolithic’ typically refers to new forms of lithic material culture, and/or the presence of ceramics (Fowler, Harding and Hoffman, 2015). For instance, the Neolithic cultures of the Central Asian Steppe possessed pottery, but retained a hunter-gatherer economy alongside stone blade technology similar to the preceding Mesolithic cultures (Kislenko and Tatarintseva, 1999). The archaeological record testifies to a boundary, ranging from the eastern Baltic to the Black Sea, east of which hunter-gatherer societies persisted for much longer than in western Europe (Mittnik et al., 2018). However, the possible population genomic implications of this phenomenon is not known. Another enigma in the neolithisation debate is that of Scandinavia (Fischer and Kristiansen, 2002). The introduction of farming reached a 1,000-year standstill at the doorstep to Southern Scandinavia before finally progressing into Denmark around 6 ka. It is not known what caused this delay, and whether the transition to farming in Denmark was facilitated by the migration of people (demic diffusion), similar to the rest of Europe (Fort, 2015; Lipson et al., 2017; Mathieson et al.,

2018; Brace et al., 2019) or mostly involved cultural diffusion (Zvelebil and Rowley-Conwy, 1984; Price, 2000; Melchior et al., 2010). Lastly, although analyses of ancient genomes have uncovered large-scale migrations from the Pontic Steppe both into Europe and Asia around 5 ka, the details of this transforming demographic process has remained largely unresolved.

To investigate these formative processes of the early Eurasian gene pools, in Allentoft et al. (2022) we conducted the largest aDNA study on human Stone Age skeletal material to date. We sequenced the genomes of 317 radiocarbon-dated (AMS) primarily Mesolithic and Neolithic individuals, covering major parts of Eurasia, and combined them with published shotgun-sequenced data to impute a dataset of >1600 diploid ancient genomes. Genomic data from 100 AMS-dated individuals from Denmark supported detailed analyses of the Stone Age population dynamics in Southern Scandinavia. When combined with genetically-predicted phenotypes, proxies for diet ($\delta^{13}C/\delta^{15}N$), mobility ($^{87}Sr/^{86}Sr$) and vegetation cover (pollen) we could connect this with parallel shifts in phenotype, subsistence and landscape.

This chapter contains my contribution to this study, which is focussed on assessing the genetic legacy of these ancient populations in a modern panel, the UK Biobank.

Methods

Data

Our primary data consists of genomes from 317 ancient individuals. A total of 272 were radiocarbon dated within the project, while 39 dates were derived from literature and 15 by archaeological context. Dates were corrected for marine and freshwater reservoir effects and ranged from the Upper Palaeolithic (UP) c. 25,700 calibrated years before present (cal. BP) to the mediaeval period (c. 1200 cal. BP). However, 97% of the individuals (N=309) span 11,000 cal. BP to 3,000 cal. BP, with a heavy focus on individuals associated with various Mesolithic and Neolithic cultures. Ancient DNA was extracted from dental cementum or petrous bone and the 317 genomes were shotgun sequenced to a depth of 0.01X to 7.1X (mean = 0.75X, median = 0.26X), with >1X coverage for 81 genomes. We utilised a new method optimised for low-coverage data to impute genotypes using the 1000 Genomes phased data as a reference panel. We also applied this to >1300 previously published shotgun-sequenced genomes, resulting in a dataset of 8.5 million common SNPs (>1% Minor Allele Frequency (MAF) and imputation info score > 0.5) for 1,664 imputed diploid ancient genomes. Overall, this dataset allows us to characterise the ancient cross-continental gene pools and the demographic transitions with unprecedented resolution. More information on sampling, data generation and imputation can be found in Allentoft et al., 2022.

Msprime

In order to test the accuracy of ChromoPainter, and the assertion that the local copying probabilities can be used as a proxy for local ancestry, we ran an *msprime* (Kelleher et al., 2016) simulation (Figure 1; coding for model done by Alice Pearson), designed to represent Eurasian demographic history based on previous literature (Jones et al., 2015). ChromoPainter was then run on the modern samples in the simulation using the ancient simulated samples as donors. The simulation can output the full ancestral recombination graph (ARG), and therefore be used as a ground truth with which to compare the painting output.

The number and dates of ancient genomes sampled is approximately the same as in the ancient panel (see below). The mutation rate parameter used (1.25 × 10−8 bp-1 generation-1) is an attempted consensus from differing studies (Roach et al., 2010; Scally & Durbin, 2012; Narasimhan et al., 2017), and the HapMap recombination maps (The International HapMap Consortium, 2003) were used for the simulation.

We ran ChromoPainter to paint all ancient individuals from the simulation using the remaining individuals as donors (i.e. in automatic mode). We then used the co-ancestry matrix to perform unsupervised clustering using fineSTRUCTURE, and compared the co-ancestry matrix to that produced by the ancient dataset.

To infer the local ancestry of variants in the modern individual, ancient individuals were assigned to reference populations based on their sample provenance: WHG formed from WHG samples, Farmer formed from Neolithic and Anatolian samples, and Steppe formed from Yamnaya, EHG and CHG samples (Figure 1). Then:

1. Each individual was repainted twice leaving out themselves as a possible donor: first to infer the painting parameters Ne (effective population size) and μ (mutation rate), and then to learn a genome-wide individual-specific donor-prior. For each of the three reference populations, the average amount of genome received from each donor individual is learnt.
2. Modern simulated individuals were painted using the reference populations and the parameters and priors inferred above. Information about local ancestry as well as genome-wide ancestry was stored.

The output of this analysis (local ancestry inferred from painting) was compared with information extracted directly from the simulated tree sequences – that is, the proportions of nearest neighbours from each reference population at each locus, calculated from the local trees in the simulation.

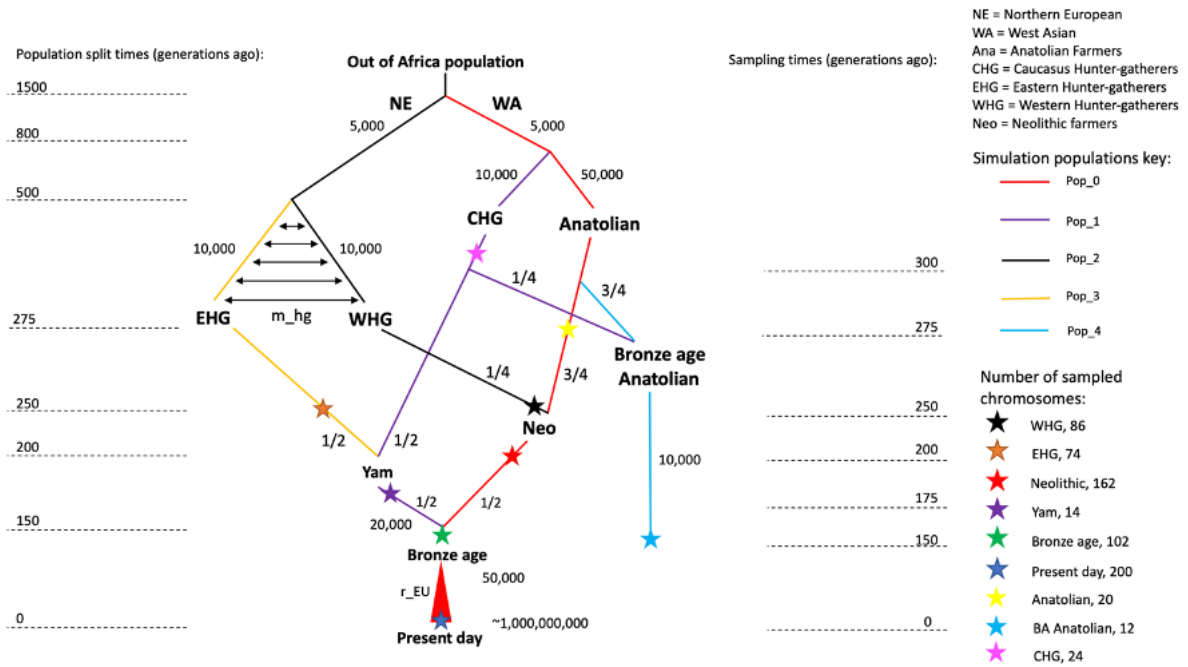


Figure 1 | Schematic of the *msprime* simulation, based on Jones et al. (2015).
Figure represents a model of European population structure as used in the *msprime* simulation. Genetic material can travel backwards from modern populations along the branches, all coalescing in the Out of Africa population. Historical populations are approximately labelled on their respective branches, with sampling times shown as stars. The numbers on each branch represent the population size, while fractions at nodes represent admixture proportions from each parent branch.There is migration between the EHG and WHG branches to reflect an admixture cline, and a population expansion from the Bronze Age to the Present day. **Figure credit: Alice Pearson.**

MOSAIC

MOSAIC (Salter-Townshend et al., 2019), a software explicitly designed to infer local ancestry segments which doesn't require any prior knowledge of the relationships between subgroups of donor haplotypes and the unseen mixing ancestral populations, was also tested as an alternative to ChromoPainter. MOSAIC implements a two-layer HMM model very similar to HapMix (Price et al., 2009) but allowing more than two admixing groups and without the requirement to have known surrogates for each ancestry. In this sense it is similar to a commonly used application of ChromoPainter, GLOBETROTTER (Hellenthal et al., 2014), where a mixture model is fitted to the output of an ancestry unaware HMM to infer the relationship between modern populations and unseen mixing populations. MOSAIC has been shown to perform similarly or better than GLOBETROTTER in the case of a two-way admixture both in terms of local ancestry and estimating admixture parameters (the generations since admixture and the proportions of each ancestry) (Salter-Townshend et al., 2019).

However, when MOSAIC was run on the samples from the *msprime* simulation the unseen admixing populations could not be meaningfully interpreted - they did not appear to represent any known ancestral groups. While there were three ancestry groups contributing to the modern population, models based on 'symmetric recent admixture' such as ADMIXTURE (Pritchard et al., 2000) and MOSAIC can be misled by the different time depths involved (Lawson et al., 2018): they try to create ancestral populations that explain the most modern variation, without having an explicit model of how the ancestral populations relate to one another. Thus, it is not clear how the results could be interpreted. Furthermore, the computational cost is considerably higher, in the order of ~10X longer compute time when compared with ChromoPainter, making the computational cost of applying this method to the entire UK Biobank unfeasible. It was therefore decided to use pre-specified donor groups with ChromoPainter as the main method of analysis.

Painting pipeline introduction

The process of painting consists of forming a reference/donor panel consisting of ancient individuals of as homogeneous ancestry as possible, having undergone QC and clustering using fineSTRUCTURE. The target/recipient panel and reference/donor panel are then filtered for variants, merged, and the target panel is painted using the reference panel as donors.

Reference/donor panel formation

We used imputed best guess haplotypes filtered for imputation information score (FORMAT/INFO) above 0.5. Samples were selected based on IBD-sharing, a visual inspection of PCA (PLINK v1.90b4.4, Chang et al., 2015) (Figure 3, Figure 4 for final selection), and fineSTRUCTURE analysis (unsupervised clustering based on the coancestry matrix output of ChromoPainter; Figure 2); low coverage, contaminated, and related individuals were excluded. The aim was to group samples into 'source' populations, defined as a group of samples which copies more from itself than other populations, while maintaining reasonable numbers in each population. We do not expect the filters for white British/non-british individuals (see Chapter Two) to be perfect; furthermore, modelling modern Eurasians as a mixture of hunter-gatherer/Steppe/farmer is overly simplistic. Therefore, we also include ancient African and EastAsian reference populations to account for possible non-European ancestry.

Ultimately, 318 ancient samples were split into ten reference populations (Table 1): western hunter-gatherer (WHG), eastern hunter-gatherer (EHG), Caucasus hunter-gatherer (CHG),

FarmerAnatolian, FarmerEarly, FarmerMiddle, FarmerLate, Yamnaya, African and EastAsian. Populations are characterised by preferentially copying from individuals within the population, as well as being biologically and historically meaningful.

The farmers were split into four separate populations due to their differing behaviour as donors (columns) in the fineSTRUCTURE analysis (Figure 2). There is a cline in their degree of WHG admixture that roughly correlates with age, while some samples also show Steppe admixture. Given the nature of the splits, the differences between these groups should be interpreted with caution, and for most downstream analyses these groups were merged into a generic 'Farmer' ancestry.

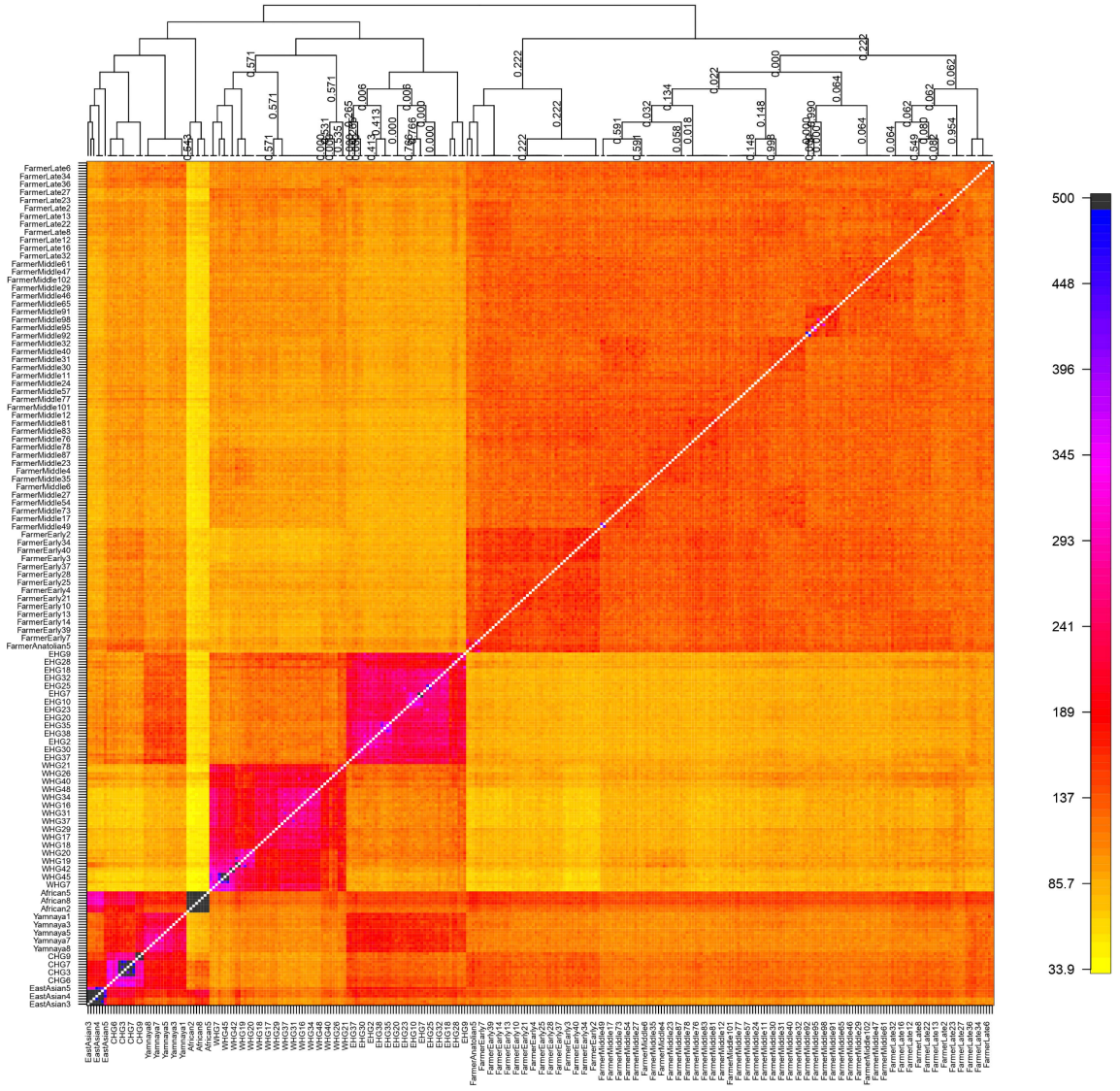


Figure 2 | Co-ancestry heatmap of selected ancient samples.
The output of fineSTRUCTURE analysis of the ancient reference panel, showing copying proportions between ancient populations (columns=donors, rows=recipients). There is a cline in Hunter-Gatherer

admixture in the Farmers, roughly correlating with age. For most downstream analyses, the Farmer populations were merged.

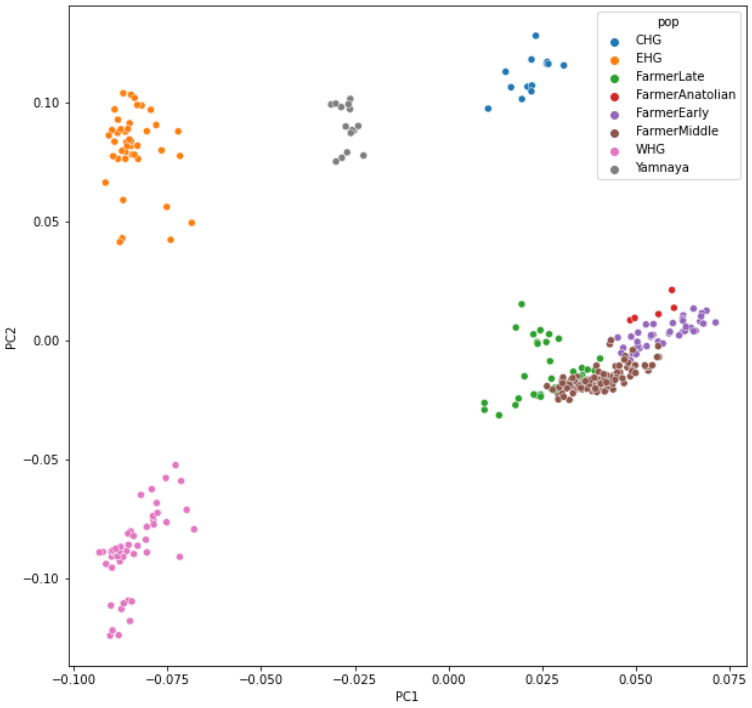


Figure 3 | PC1 vs PC2 of ancient reference samples, coloured by assigned population.
PC1 vs PC2 of a PCA of the ancient Eurasian samples (excluding African and EastAsian), coloured by their assigned population used in the painting. As can be seen, populations are fairly distinct, with intermediate admixed individuals having been excluded. Some Farmers are admixed with Steppe and Hunter-Gatherer populations to differing degrees, but particularly among later individuals.

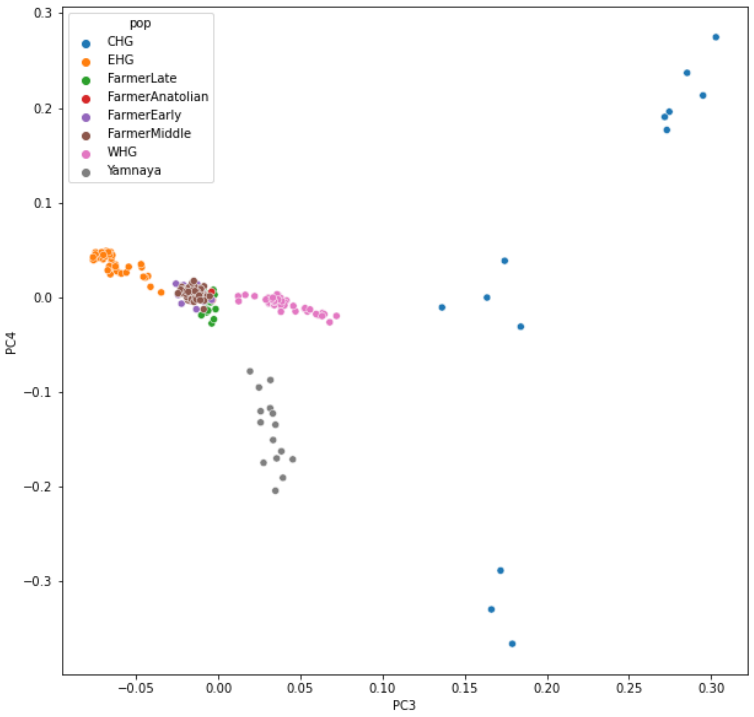


Figure 4 | PC3 vs PC4 of ancient reference samples, coloured by assigned population.
PC3 vs PC4 of a PCA of the ancient Eurasian samples (excluding African and EastAsian), coloured by their assigned population used in the painting.

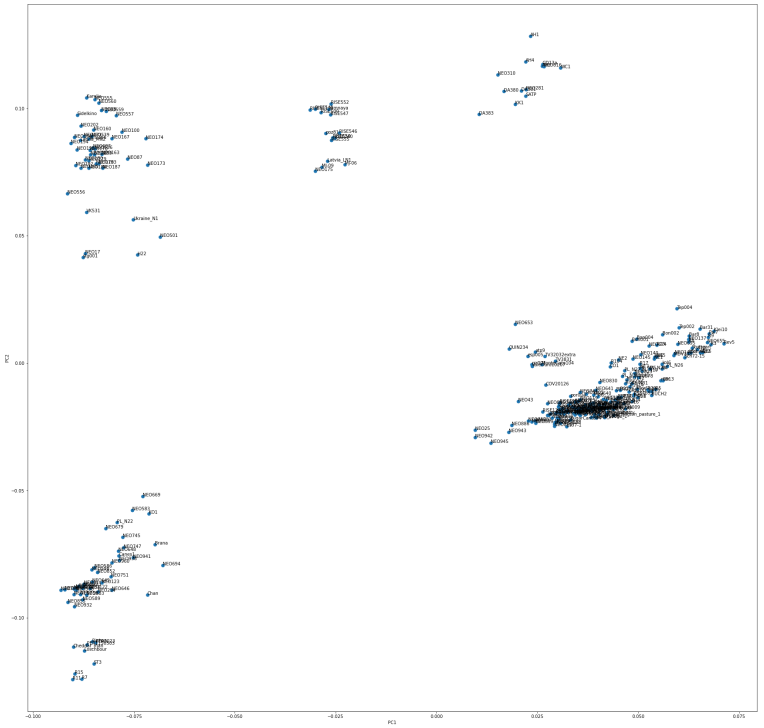


Figure 5 | PCA of ancient reference samples, labelled with sample name.
PC1 vs PC2 of a PCA of the ancient Eurasian samples (excluding African and EastAsian), labelled with sample name.

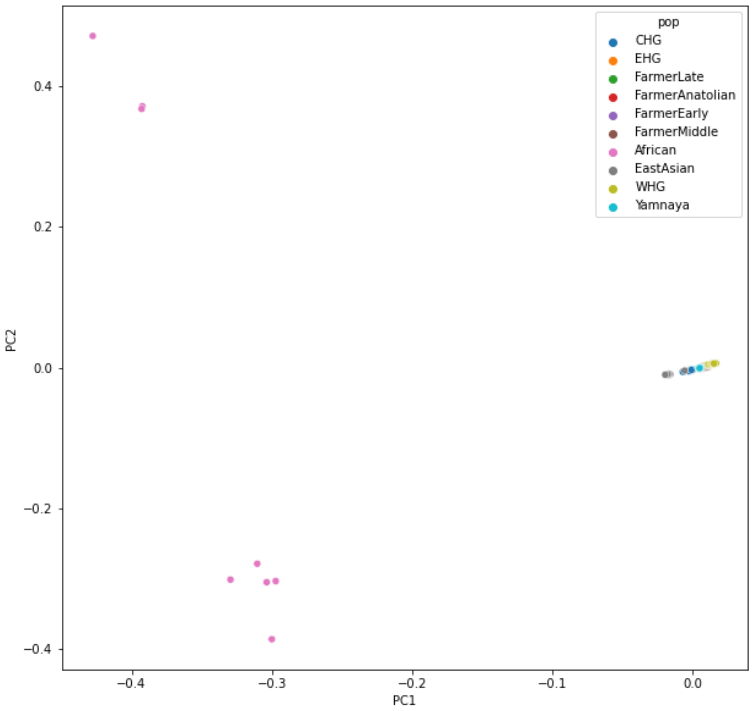


Figure 6 | PC1 vs PC2 of ancient reference samples, coloured by assigned population, including African and EastAsian populations.

When African and East Asian populations are included in the PCA, the PCs explaining the highest variance in the sample are dominated by the African (and to a lesser extent the East Asian) components

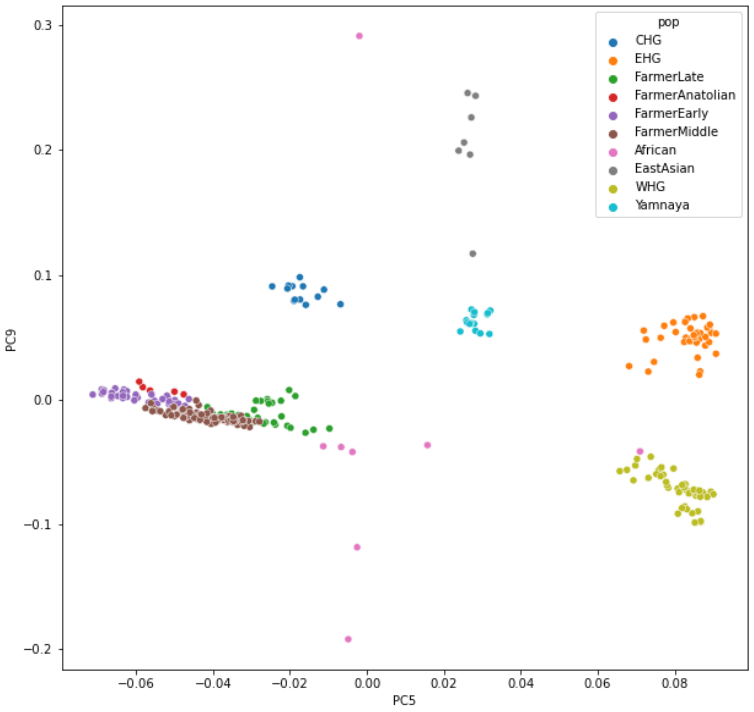


Figure 7 | PC5 vs PC9 of ancient reference samples, coloured by assigned population, including African and EastAsian populations.

The first informative PCs for splitting the Ancient Eurasian samples are PC5 and PC9. These PCs also separate out modern British population structure (Sarmanova, Morris and Lawson, 2020).

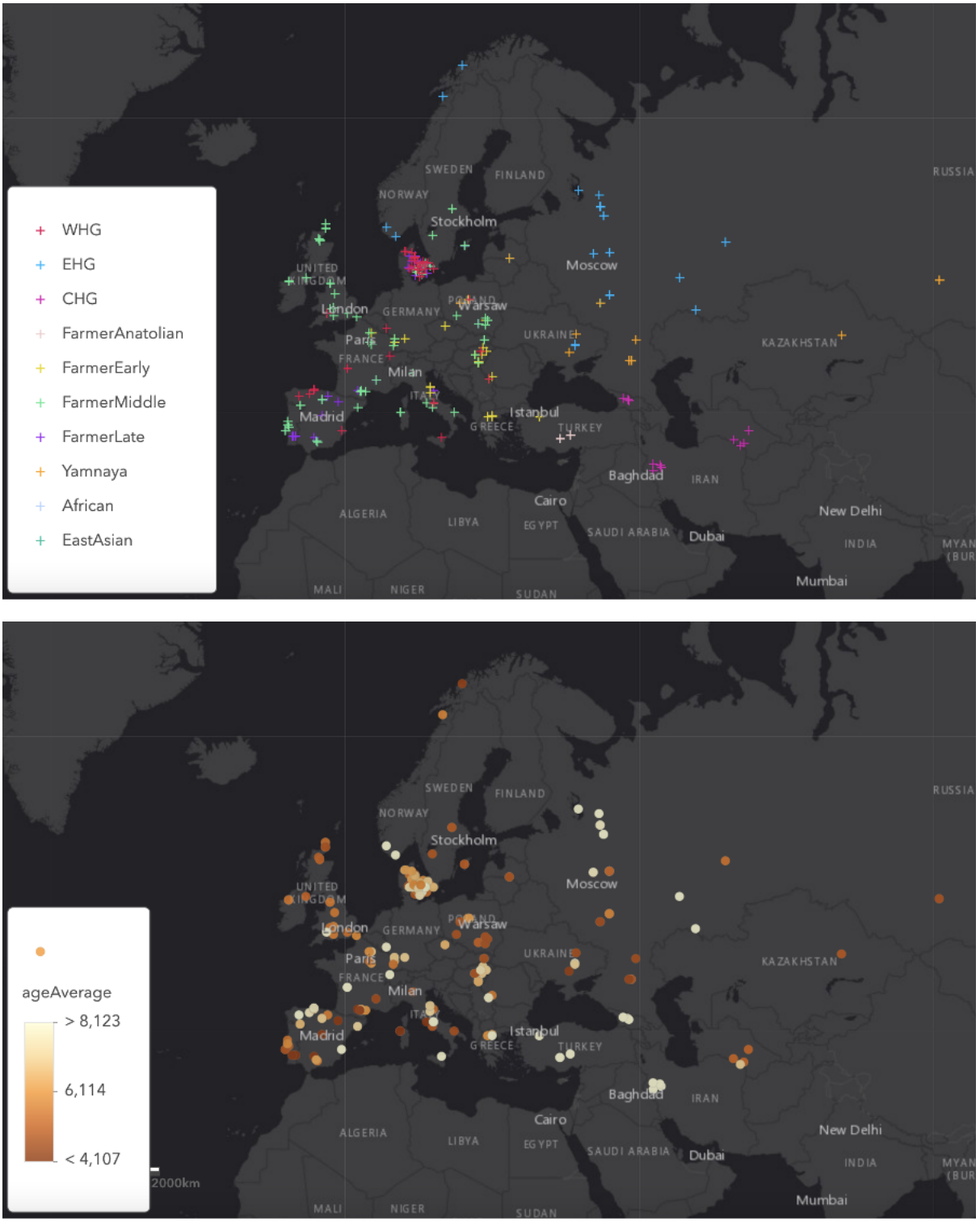


Figure 8 | Maps showing ancient Eurasian sample locations coloured by assigned reference population (above) and age (below).

Not showing African and East Asian samples.

Target/recipient panel formation

We used white British individuals from the UKB as reported in Bycroft et al. (2018); these are individuals who self-reported as white British and have British-like ancestry according to PCA. We also used individuals from the UKB of a typical ancestral background selected by country of origin (Chapter Two). We used phased haplotype data, downloaded from https://www.ukbiobank.ac.uk. This totalled 408,884 white British individuals, and 24,511 non-British individuals.

SNP selection and merging of the panels

Due to computational considerations, the number of SNPs used in the painting was limited to those in the UKB Axiom Array; these SNPs were chosen to capture genome-wide variation, rare and coding variants, and variants relevant to specific phenotypes or regions of interest (Bycroft et al., 2018). The aDNA dataset and UKB datasets were merged and filtered for these variants using QCTOOL v2 (https://www.well.ox.ac.uk/~gav/qctool_v2/), and then filtered to exclude variants with a minor allele frequency below 1% using bcftools (http://samtools.github.io/bcftools/), leaving a total of 549,323 SNPs across chromosomes 1-22.

Painting process

Chromosome painting cannot include the target of painting. Therefore, painting was done (following the pipeline of Margaryan et al. (2020) based on GLOBETROTTER (Hellenthal et al., 2014) by leaving out one individual at random (chosen independently for each chromosome) from each other donor population for all donor individuals. Target individuals from the UKB were painted by similarly removing one individual at random from all donor populations. This ensures that painted individuals from the reference and UKB are exchangeable.

Once I had a well-chosen set of ancient populations from the aDNA panel, each individual was repainted twice leaving out themselves as a possible donor: first to learn the painting parameters Ne (effective population size) and μ (mutation rate), and then to learn a genome-wide individual-specific donor-prior: for each of the reference populations, the average amount of genome received from each donor individual was learnt. I then painted the modern individuals in the UKB panel using the reference populations and the learnt parameters and priors.

The probability that each recipient copied each donor population at every SNP was recorded. The genome-wide information for each recipient was also stored, in the form of (i) chunkcounts, the number of chunks copied from each donor population and (ii) chunk lengths, the sum of the lengths of the chunks copied from each population, weighted by their copying probability. Admixture proportions were then estimated using Non-Negative Least Squares (NNLS).

Painting at biobank scale

The standard pipeline for using a reference panel to paint a set of target samples, as published in Margaryan et al. (2020), was too slow to perform on a biobank scale. The main limiting step for this was the reading of one of the input files ('phasefiles'), containing SNP data for each individual (target and donor) as a text file, with rows as phased haplotypes (i.e. two rows per sample; more detail available at https://people.maths.bris.ac.uk/~madjl/finestructure/manualse4.html#x7-90004.1). Reading this very large text file into memory increased the memory usage and compute time; in comparison, the actual painting algorithm was relatively fast.

I made two major adjustments to the pipeline to allow faster painting without changing the core algorithm. The first was to split the target panel into batches of 24,000 individuals and paint each of these batches on separate nodes; within a batch, individuals were painted on separate cores in temporary directories containing individual phasefiles, minimising the input files being read into memory. Because each target sample is painted independently against the reference panel, no batch effects result from this process.

The second change was to the format of the output file containing the per-locus copying probabilities. In the original ChromoPainter code, the format of this is a text file containing, for each target haplotype, the probability of copying from each donor population at each SNP to six decimal places. When running ChromoPainter on hundreds of thousands of individuals at hundreds of thousands of SNPs this file becomes very large. The solution was to use a new file format and add to a master file on-the-fly as each target individual was painted. To convert to the new format, I binned the copying probabilities (between 0 and 1) into 10 bins, represented by the integers 0-9 (0 >= x < 0.1 bins to 0, 0.1 <= x < 0.2 bins to 1 etc). These are stored as a zipped text file, one per chromosome. An example for one sample (two haplotypes) painted using two donor populations at seven SNPs:

```
SampleId,1,0000888
SampleId,1,9999111
```

SampleId,2,5557772
SampleId,2,4442227

This file format results in long strings of identical copying probabilities (haplotypes), which makes zipping very efficient.

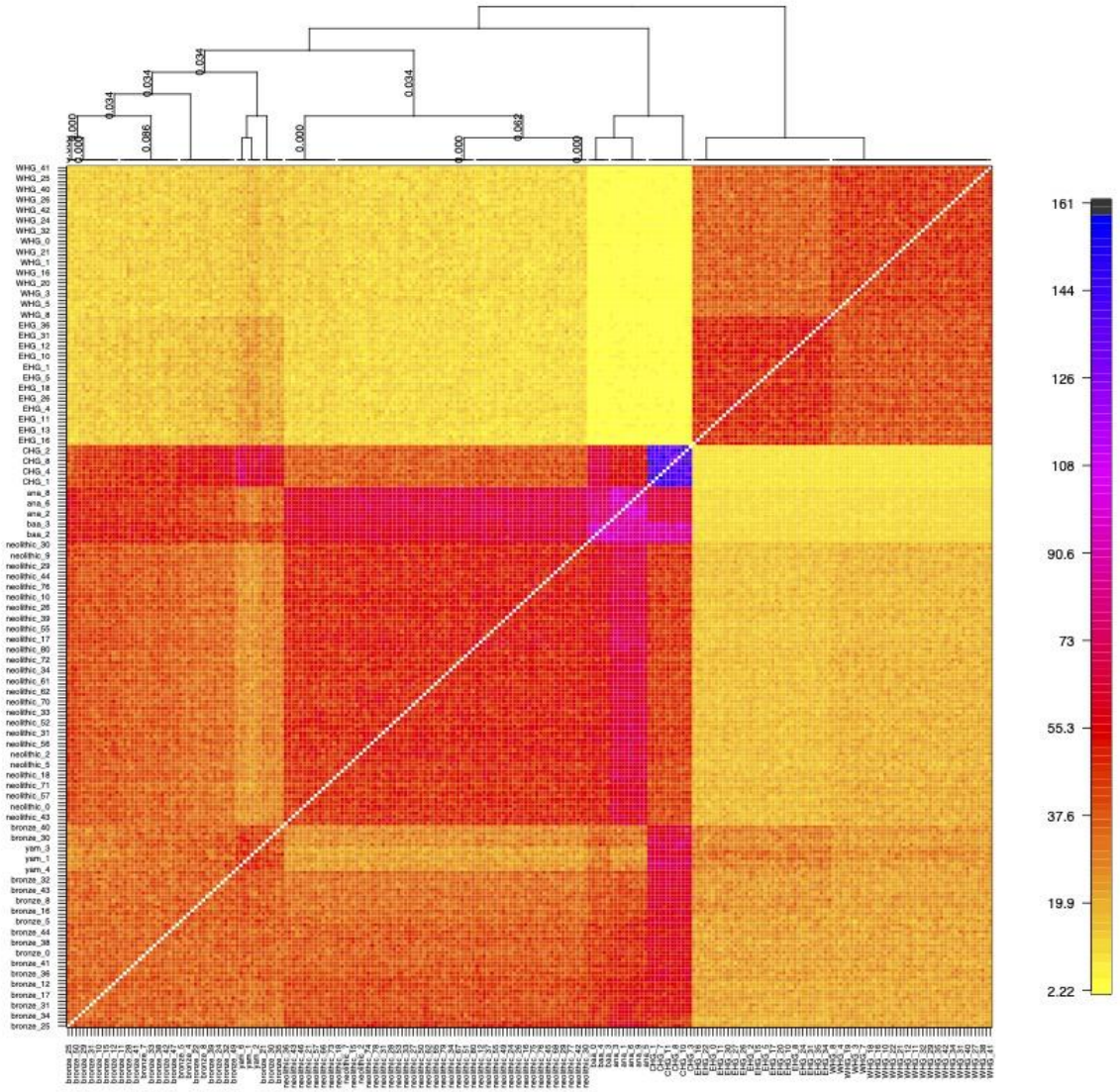
These changes enabled painting for large numbers of recipients in parallel across multiple nodes, reducing per-thread memory usage and storing of local copying probabilities in a memory-efficient format in real time (all scripts available at https://github.com/will-camb/Nero/tree/master/scripts/cp_panel_scripts). The total CPU time for painting the UKB panel was approximately 550,000 CPU hours.

A ‘painting manual’ can be found in Chapter One Appendix.

Results

msprime

Running fineSTRUCTURE on the simulated ancient individuals yielded two conclusions: firstly, and unsurprisingly, fineSTRUCTURE was better at assigning individuals to their expected clusters when more of the genome was simulated (chromosomes 1, 2 and 3 as opposed to just chromosome 1). Secondly, the co-ancestry matrix produced by the real ancient dataset was visually similar to that produced by the simulated data (Figure 9), indicating that the simulation approximated well the patterns of shared haplotypes in the real data, and therefore may be a good representation of the real demographic history.



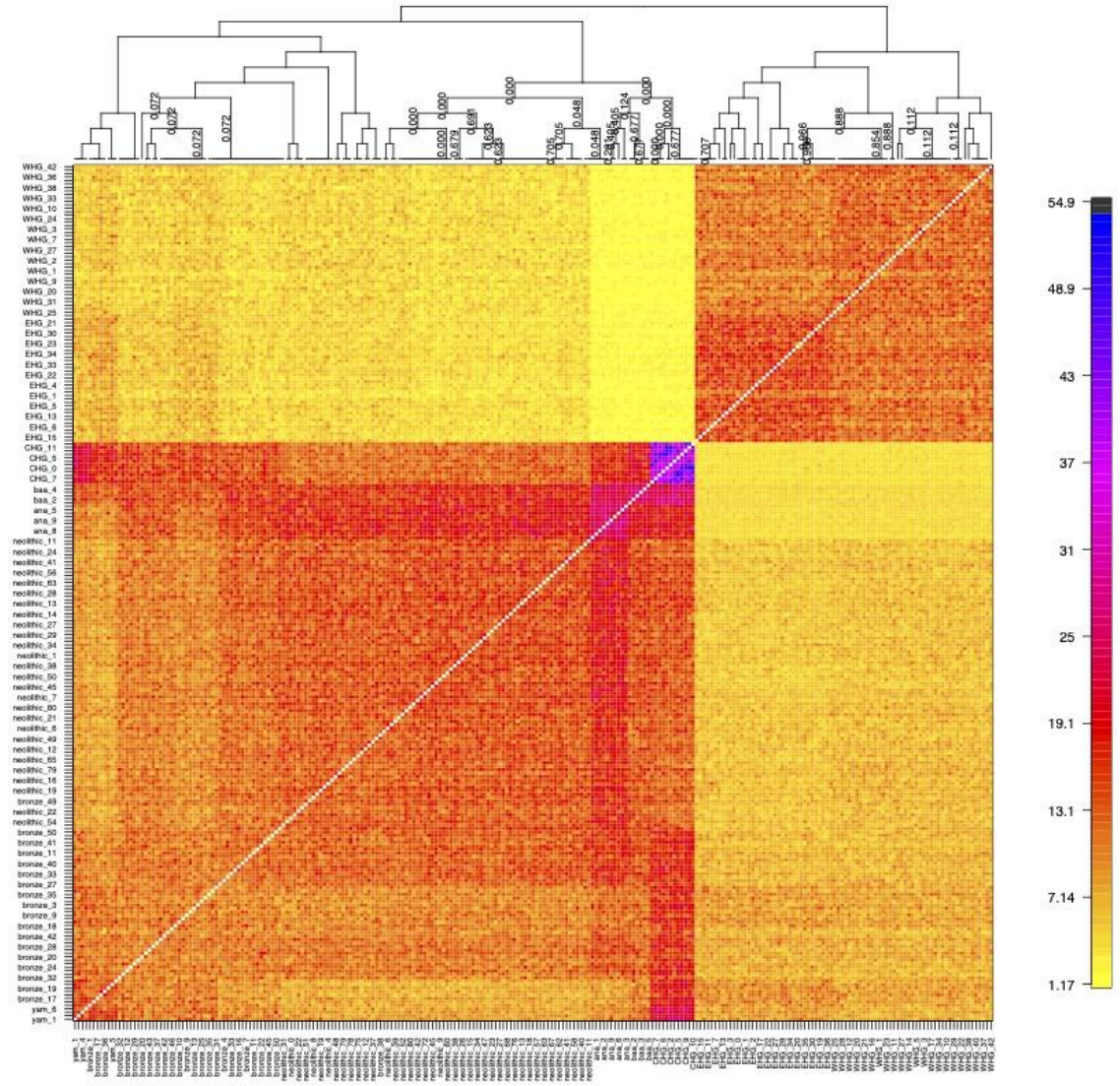


Figure 9 | Coancestry heatmap of simulated (above) and real (below) reference panels.
The output of fineSTRUCTURE analysis of the simulated and real reference panels. The similarity of the heatmaps suggests that the model is a good approximation of European genomic history. In both, the Caucasus Hunter-Gatherers copy significantly more from themselves than other populations.

From visual inspection, I found that ChromoPainter was slightly over-confident in its assessment of painting probability when nearest neighbours were from multiple donor populations (Figure 10). When looking at results over the entire simulated chromosome, when ChromoPainter was 95% confident in its painting, the probability that the majority of the nearest neighbours were from that population ranged from 77% (WHG) to 87% (Steppe) to 92% (Farmer). The accuracy seemed to depend on the number of samples in each reference population, and the age of the samples.

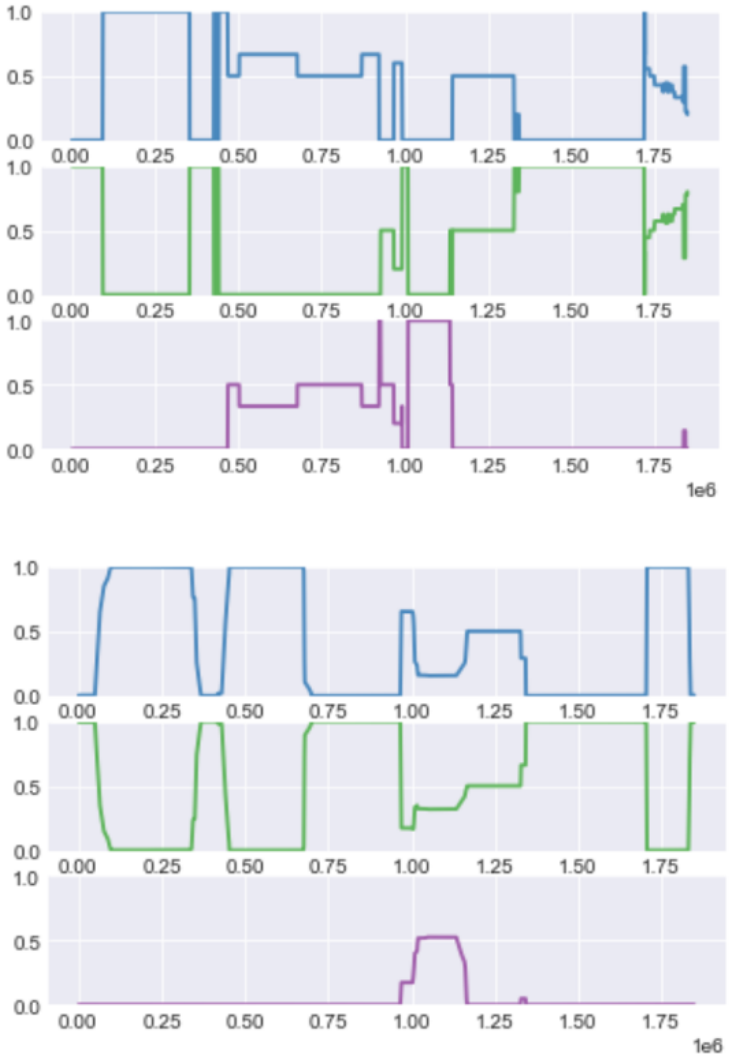


Figure 10 | Truth (above) and ChromoPainter output (below) for the first 10,000 sites of a simulated haplotype.

A plot of the proportion of nearest neighbours from each ancestry group ascertained directly from the tree sequence (above) and from ChromoPainter (below). Blue = Steppe, Green = Farmer, Purple = WHG. ChromoPainter is generally correct in its assignments, although it appears to be over-confident and to switch between ancestries less than it should.

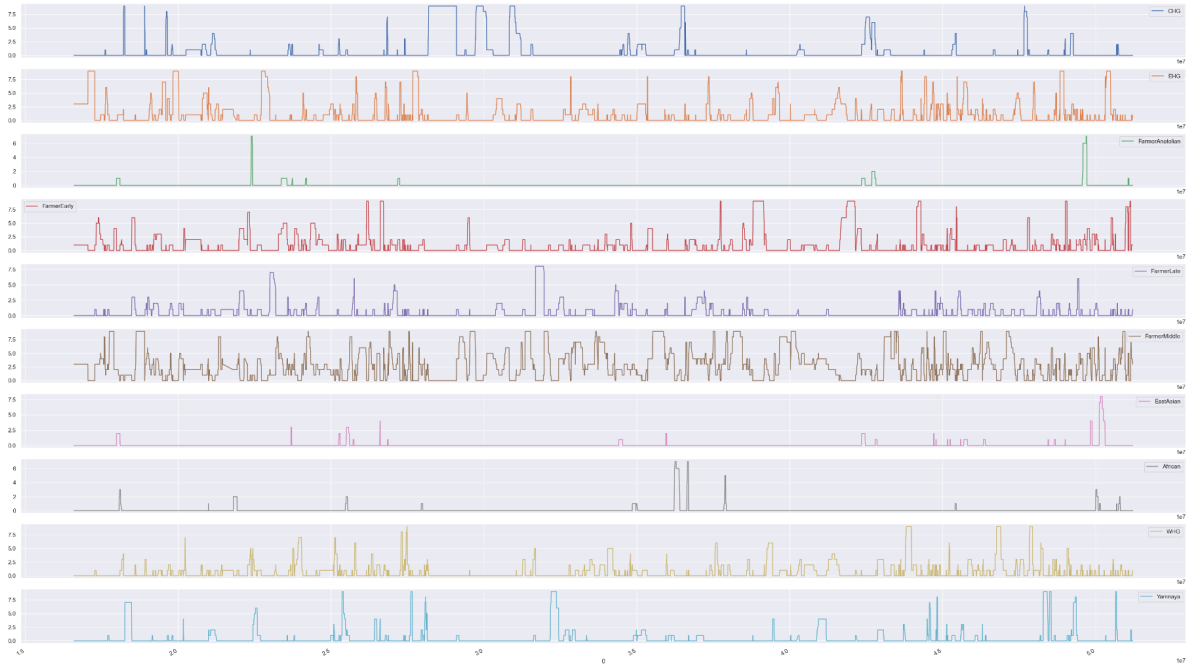


Figure 11 | Painting results for chromosome 22 for one haplotype

Painting probabilities at each position of chromosome 22, decomposed by ancestry, now using 10 reference populations.

Ancestry-geographic variation

Within Great Britain, all individuals were painted with similar proportions from each reference population (Figure 12), as expected when measuring coalescence tracts rather than direct admixture tracts and after a long time since admixture events; but, the differences in copying proportions showed significant geographic heterogeneity. I ran multivariate linear regressions, using longitude and latitude of place of birth ("Place of birth in UK – east co-ordinate" and "Place of birth in UK – north co-ordinate") to predict log-transformed NNLS ancestry fractions. I found significant correlations ($p<0.05$) for Yamnaya ancestry (R-squared=0.081), Farmer ancestry (R-squared=0.066), CHG ancestry (R-squared=0.015), WHG ancestry (R-squared=0.007), African ancestry (R-squared=0.011) and EHG ancestry (R-squared=0.002, longitude only). To visualise this, I assigned individuals to a county based on their UKB place of birth data, and plotted the average admixture proportion per county for each ancestry, binned in ten equal interval quantiles using ArcGIS Online (www.arcgis.com; Figure 5).

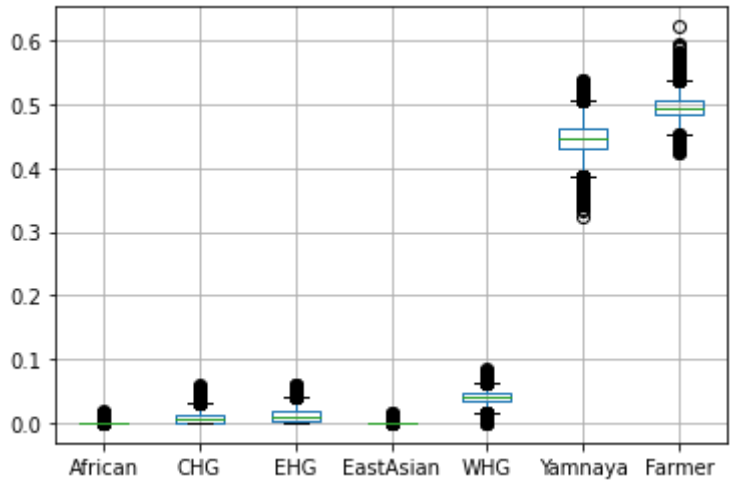


Figure 12 | Boxplot of admixture proportions for 'white British' individuals in the UKB, decomposed by ancestry. All individuals are inferred to have similar proportions of each ancestry, reflecting old admixture events.

I found that Neolithic farmer ancestry was highest in southern and eastern England today and lower in populations in Scotland, Wales and Cornwall (Figure 13, insets). I found that Steppe-related ancestry (Yamnaya) is inversely distributed, which has previously been shown to be higher in Scotland but not Wales (Galinsky et al., 2016): this was highest in the Outer Hebrides and Ireland. This regional pattern was already evident in the Pre-Roman Iron Age and persists to the present day even though immigrating Anglo-Saxons had relatively less Neolithic farmer ancestry than the Iron-Age population of southwest Briton (Allentoft et al., 2022, Extended Data Fig. 4). Although this Neolithic farmer/steppe-related dichotomy mirrors the traditional (but outdated) 'Anglo-Saxon'/'Celtic' ethnic divide, its origins are older, resulting from continuous migration from a continental population relatively enhanced in Neolithic farmer ancestry, starting as early as the Late Bronze Age, into England and Wales but not Scotland (Patterson et al., 2021). By measuring haplotypes from these ancestries in modern individuals, I was able to show that these patterns differentiate Wales and Cornwall as well as Scotland from England. I also found higher levels of WHG-related ancestry in central and Northern England. These results demonstrate clear ancestry differences within an 'ethnic group' (white British) traditionally considered relatively homogenous, which highlights the need to account for subtle population structure when using resources such as the UK Biobank genomes.

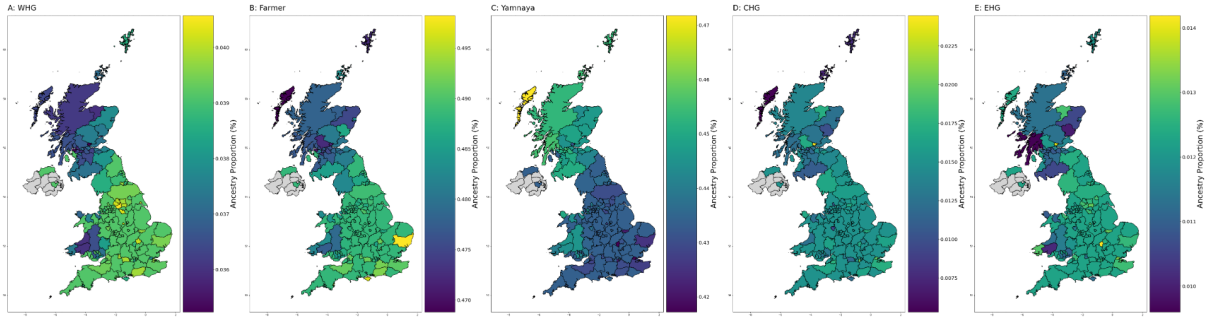


Figure 13 | The genetic legacy of Stone Age ancestry in modern populations.
Panels show average admixture proportion in modern individuals per county within the UK estimated using NNLS. A: Western hunter-gatherer, B: Neolithic Farmer, C: Yamnaya, D: Caucasus hunter-gatherer, E: Eastern hunter-gatherer.

Known ancestry-specific variants: LCT/MCM6

One way of testing the local ancestry painting is to examine variants which are known to have been under recent strong selection and have ancestry-specific origins. The most famous of these is the LCT/MCM6 locus, regulating the lactase persistence phenotype. This is the most strongly selected monogenic trait to have evolved in Europe (Evershed et al., 2022). A variety of explanations have been given for its highly structured geographical distribution in Europe, including differing strengths of selection (Cubas et al. 2020) and demographic processes (Allentoft et al., 2015).

When the average painting probabilities for each ancestry are plotted across chromosome 2, the LCT/MCM6 locus stands out as having excess Yamnaya and EHG ancestry, in line with previous studies in which Bronze Age steppe populations showed the highest derived allele frequency among ancient groups (Allentoft et al., 2015); and a relative lack of Farmer ancestry (Figure 14). This tells us that the haplotype must have been selected after the date of admixture between the Steppe populations and the ‘indigenous’ Bronze Age people of western Europe (though doesn’t preclude selection before this time too), and likely arrived in the Steppe population via their EHG ancestry component. Given the ancestry gradients within Britain (Figure 13), this may explain the observed higher frequency of lactase non-persistence (*rs4988235*: C allele) in south-eastern Britain (Smith et al., 2008), where Farmer ancestry predominates over Yamnaya/Steppe ancestry.

Figure 14 | Average painting probabilities for chromosome 2 across all 'white British' individuals in the UKB for Yamnaya (top), EHG (middle) and Farmer (bottom) ancestry. Even when looking at all individuals, not just those who are lactase persistent, we see clear evidence of excess Yamnaya and EHG ancestry at the LCT/MCM6 locus, indicating strong selection. Surrounding regions also show an excess, indicative of a classic selective sweep.

Discussion

In this chapter I have developed methods to use ChromoPainter on a Biobank scale, and used simulations to examine the accuracy of this method to infer local ancestry. Although ChromoPainter accuracy is varied, it is not expected to introduce bias in downstream analyses (see Chapter Three) because all traits and alleles are equally affected with a tendency to paint from larger and more recent reference populations. At a genome-wide level, the relative difference in the amount copied from each reference panel is still expected to reflect actual admixture proportion differences, though masking will mean that admixture values for older populations are slightly under-estimated.

The large ancient DNA panel combined with the UKB allowed me to trace for the first time the fine-scale distribution of Mesolithic/Neolithic/Bronze Age ancestry components in modern British individuals, using DNA directly from ancient individuals. Doing this provides unprecedented insight into the fine-scale ancestry gradients that still exist across the UK: even in ancestries that have long since admixed and even within as well as between the constituent nations of the United Kingdom for which we have county-level resolution - England, Scotland and Wales. This work has several important implications.

The first implication is in terms of thinking about the native population of Great Britain (here defined by self-identification as ‘white British’ and followed with PCA outlier removal). In population genetics it is common to talk about populations, in which, theoretically, individuals cannot be distinguished from each other based on the data available; central to this idea is that individuals within a population exhibit random mating with one another (Lawson, 2015). Populations can be identified statistically based on the sharing of SNPs: for example, McVean (2009) described how projecting samples based on Principal Components (PCs) of genetic variation (SNPs), ordered by variance explained, can identify samples with shared SNP frequencies and therefore can identify ‘populations’. STRUCTURE (Pritchard et al., 2000) and ADMIXTURE (Alexander et al., 2009) are commonly used approaches for identifying populations based on SNP frequencies, under the assumption that SNPs are independent (i.e. not in LD). In humans, this population concept often identifies geographically and historically meaningful groupings (e.g. Rosenberg et al., 2002) and is commonly applied. However, as methodology and data collection have become more sophisticated, population substructure has become more apparent. A more sensitive approach to identifying populations and ‘substructure’ is based on the sharing of haplotypes, a method employed by fineSTRUCTURE based on the ChromoPainter co-ancestry matrix (Lawson et al., 2012). When LD between SNPs is ignored (i.e. they are ‘unlinked’),

ChromoPainter is equivalent to STRUCTURE (Lawson et al., 2012). Theoretically, with enough data every individual would form its own ‘population’.

The results presented here may be viewed as an extension of this trend, identifying substructure within the ‘white British’ population. Where previously structure has been identified (most notably in Britain in Leslie et al., 2015), a lack of ancient reference genomes then made the interpretation of the population clusters difficult, although modern reference populations were used to decompose the approximate ancestry of different clusters, and ‘best-fit’ historical explanations for these ancestries were attempted. Here however, we can clearly explain population substructure as resulting at least partly from differences in genome-wide ancient ancestry proportions (see Chapter 2 for a discussion of the relationship between ancestry components and PCs). Dating how and when these ancestry gradients arose is difficult, and it would clearly be informative to perform a similar analysis using more recent ancient samples. We can however conclude that the gradients are relatively old, as they are continuous throughout the entire population examined: more recent admixture would result in pockets of differing ancestry rather than gradients (e.g. Gretzinger et al., 2022).

The second implication of this work is that by demonstrating the ancestry differences within an ‘ethnic group’ (white British) sometimes regarded as being relatively homogenous, this work highlights the need for care over ancestry considerations when using resources like the UKB in downstream analyses. This resource is widely used in GWAS studies, which are sensitive to the correlation between ancestry and other covariates (both genetic and non-genetic, e.g. socioeconomic or climate-related): removing this confounding allows the detection of true genetic functions. The fact that ancestry is so strongly geographically structured within Britain emphasises the importance of this control, as ancestry will correlate strongly with any measure that is also geographically structured.

Tables

referencePopulation	sampleId	country	groupLabel	latitude	longitude	ageAverage	coverage	sex	clusterIBDFine
African	mfo01	South Africa	SouthAfrica_IronAge	-28.73	30.81	378.0	7.061854822000000	XX	6.1_SouthAfrica_400BP
African	bab01	South Africa	SouthAfrica_Neolithic	-29.54	31.22	2040.5	1.29865429	XY	6.2_SouthAfrica_2000BP_1000BP
African	ela01	South Africa	SouthAfrica_IronAge	-28.92	29.13	493.0	13.46552376	XX	6.1_SouthAfrica_400BP
African	I10871	Cameroon	Cameroon_Neolithic	5.86	10.08	7885.0	15.21262736	XY	6.3_Cameroon_8000BP_3000BP
African	I10873	Cameroon	Cameroon_Neolithic	5.86	10.08	3065.0	3.276739876	XY	6.3_Cameroon_8000BP_3000BP
African	I9133	South Africa	SouthAfrica_Neolithic	-31.98	18.52	1970.0	2.0755013510000000	XY	6.2_SouthAfrica_2000BP_1000BP
African	baa01	South Africa	SouthAfrica_Neolithic	-29.54	31.22	1908.5	13.50021278	XY	6.2_SouthAfrica_2000BP_1000BP
African	new01	South Africa	SouthAfrica_IronAge	-27.76	29.92	417.5	10.89613659	XX	6.1_SouthAfrica_400BP
CHG	WC1	Iran	Iran_Neolithic	34.61	47.11	9218.5	10.4300754	XY	2.1_Iran_10000BP_8500BP
CHG	AH4	Iran	Iran_Neolithic	34.19	48.37	9929.5	0.8674508960000000	XX	2.1_Iran_10000BP_8500BP
CHG	AH2	Iran	Iran_Neolithic	34.19	48.37	9930.5	0.6492785520000000	XY	2.1_Iran_10000BP_8500BP
CHG	AH1	Iran	Iran_Neolithic	34.19	48.37	9900.0	1.161365185	XX	2.1_Iran_10000BP_8500BP
CHG	DA380	Turkmenistan	Turkmenistan_Neolithic_Namazga	37.6	59.33	5180.5	0.495449798	XX	2.1_Turkmenistan_7000BP_5000BP
CHG	DA381	Turkmenistan	Turkmenistan_Neolithic_Namazga	37.19	61.03	5181.0	0.83822953	XY	2.1_Turkmenistan_7000BP_5000BP
CHG	NEO816	Iran	Iran_Neolithic	33.76	47.1	8700.0	0.940243388	XY	2.1_Iran_10000BP_8500BP
CHG	NEO281	Georgia	Georgia_Mesolithic	42.22	43.32	9724.0	3.607878115	XY	2.1_Caucasus_13000BP_10000BP
CHG	KK1	Georgia	Georgia_Mesolithic	42.28	43.28	9720.0	11.83484526	XY	2.1_Caucasus_13000BP_10000BP

CHG	SATP	Georgia	Georgia_Mesolithic	42.38	42.59	13255.0	1.18417508	XY	2.1_Caucasus_13000BP_10000BP
CHG	GD13a	Iran	Iran_Neolithic	34.45	48.12	9846.0	1.41728065	XX	2.1_Iran_10000BP_8500BP
CHG	DA383	Turkmenistan	Turkmenistan_Neolithic_Namazga	38.72	61.69	5150.0	0.775109579 0000000	XX	2.1_Turkmenistan_7000BP_5000BP
CHG	NEO310	Turkmenistan	Turkmenistan_Neolithic	36.85	60.42	7150.0	1.278435989	XY	2.1_Turkmenistan_7000BP_5000BP
EHG	Karelia	Russia	Russia_Mesolithic	61.65	35.65	8279.5	1.692885466	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO166	Russia	Russia_Neolithic	53.0	40.4	5668.0	2.153362943 0000000	XY	4.1_DonRiver_5800BP_5300BP
EHG	NEO167	Russia	Russia_Neolithic	53.0	40.4	5657.0	0.588921338	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO170	Russia	Russia_Neolithic	53.0	40.4	5562.0	0.418896877 00000000	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO171	Russia	Russia_Neolithic	53.0	40.4	5835.0	0.818232704	XX	4.1_DonRiver_5800BP_5300BP
EHG	VK531	Norway	Norway_Neolithic	69.47	18	4350.0	1.459563752 0000000	XY	4.1_Norway_9300BP_4300BP
EHG	NEO173	Russia	Russia_Neolithic_Sredny	52.28	38.96	6345.0	0.689663925	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO100	Russia	Russia_Mesolithic	51.57	53.68	9929.0	0.107938722 00000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO88	Russia	Russia_Mesolithic	56.67	38.02	7871.0	2.621733788 0000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	H22	Norway	Norway_Mesolithic	58.83	6.33	9363.5	0.719978067	XX	4.1_Norway_9300BP_4300BP
EHG	stg001	Norway	Norway_Neolithic	67.76	14.85	5857.0	1.291071015	XY	4.1_Norway_9300BP_4300BP
EHG	NEO87	Russia	Russia_Mesolithic	56.67	38.02	8259.0	0.183898205	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO17	Norway	Norway_Mesolithic	58.06	7.74	9146.0	1.099119499 0000000	XY	4.1_Norway_9300BP_4300BP
EHG	Ukraine_N1	Ukraine	Ukraine_Neolithic	48.13	35.08	7250.0	0.167926522	XY	4.1_Ukraine_10000BP_4000BP
EHG	Latvia_MN2	Latvia	Latvia_Neolithic_CCC	56.28	25.13	5965.0	1.147611647	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO160	Russia	Russia_Neolithic	53.0	40.4	5269.0	1.326576858	XX	4.1_DonRiver_5800BP_5300BP

EHG	NEO178	Russia	Russia_Neolithic	56.78	40.45	5322.0	0.32433320600000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO163	Russia	Russia_Neolithic	53.0	40.4	5603.0	0.226384639	XY	4.1_DonRiver_5800BP_5300BP
EHG	NEO180	Russia	Russia_Neolithic	56.78	40.45	5947.0	0.38518779200000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO687	Russia	Russia_Neolithic	57.58	58.2	5446.0	0.44089811100000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO560	Russia	Russia_Neolithic	60.41	38.93	7919.0	1.548541653	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO559	Russia	Russia_Neolithic	60.41	38.93	8268.0	1.228507174	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO557	Russia	Russia_Neolithic	60.41	38.93	7917.0	0.777654091000000	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO556	Russia	Russia_Neolithic	60.41	38.93	7036.0	1.122201651	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO555	Russia	Russia_Neolithic	60.41	38.93	8280.0	2.097005144	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO539	Russia	Russia_Mesolithic	59.7	39.5	10060.0	0.291758629	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO179	Russia	Russia_Neolithic	56.78	40.45	5467.0	0.639359547000000	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO536	Russia	Russia_Mesolithic	59.7	39.5	9541.0	0.190177354	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO501	Ukraine	Ukraine_Mesolithic	48.2	35.22	10623.0	0.125455369	XY	4.1_Ukraine_10000BP_4000BP
EHG	NEO202	Russia	Russia_Mesolithic	61.27	38.91	10884.0	2.217735963	XY	4.1_RussiaNW_11000BP_8000BP
EHG	NEO197	Russia	Russia_Neolithic	56.78	40.45	5245.0	0.610890023	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO195	Russia	Russia_Neolithic	56.78	40.45	5749.0	0.59759596	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO174	Russia	Russia_Neolithic_Sredny	52.28	38.96	5306.0	1.152059756000000	XX	4.1_DonRiver_5800BP_5300BP
EHG	NEO193	Russia	Russia_Neolithic	56.78	40.45	5453.0	0.232823778	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO184	Russia	Russia_Neolithic	56.78	40.44	5458.0	0.695336472	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO185	Russia	Russia_Neolithic	56.78	40.45	7034.0	1.213238119000000	XY	4.1_RussiaNW_7000BP_5000BP

EHG	NEO194	Russia	Russia_Neolithic	56.78	40.45	5575.0	1.626304895	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO187	Russia	Russia_Neolithic	56.78	40.45	4940.0	0.16188839	XX	4.1_RussiaNW_7000BP_5000BP
EHG	Sidelkino	Russia	Russia_Mesolithic	54.53	51.11	11258.5	2.9967359070000000	XX	4.1_RussiaNW_11000BP_8000BP
EHG	NEO186	Russia	Russia_Neolithic	56.78	40.45	6922.0	0.3684240630000000	XY	4.1_RussiaNW_7000BP_5000BP
EHG	NEO192	Russia	Russia_Neolithic	56.78	40.45	6841.0	0.269995424	XX	4.1_RussiaNW_7000BP_5000BP
EHG	NEO189	Russia	Russia_Neolithic	56.78	40.45	5648.0	0.614470507	XY	4.1_RussiaNW_7000BP_5000BP
EastAsian	IK002	Japan	Japan_Jomon	34.65	137.14	2569.0	1.89109494	XX	3.1_Japan_3700BP_2600BP
EastAsian	DA45	Mongolia	Mongolia_IronAge_XiongNu	42.53	105.18	2095.0	9.039659899	XY	3.1_SEAsia_4000BP_150BP
EastAsian	DA43	Mongolia	Mongolia_IronAge_XiongNu	42.53	105.18	2095.0	1.677926438	XY	3.1_SEAsia_4000BP_150BP
EastAsian	DA39	Mongolia	Mongolia_IronAge_XiongNu	48.02	101.35	1948.0	2.087185526	XY	3.1_SteppeCE_2000BP_700BP
EastAsian	DA38	Mongolia	Mongolia_IronAge_XiongNu	49.27	101.72	2124.5	2.85446773	XX	3.1_SteppeC_TianShan_2700BP_800BP
EastAsian	Funadomari_23	Japan	Japan_Jomon	45.38	141.04	3755.0	39.44124624	XX	3.1_Japan_3700BP_2600BP
EastAsian	Funadomari_5	Japan	Japan_Jomon	45.38	141.04	3755.0	3.8214367780000000	XY	3.1_Japan_3700BP_2600BP
FarmerAnatolian	Bon001	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10032.0	0.16310050200000000	XY	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Bon002	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10078.0	6.692061812	XX	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Bon004	Turkey	Anatolia_Neolithic_Aceramic	37.75	32.86	10076.0	0.242113723	XY	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Tep002	Turkey	Anatolia_Neolithic	38.17	34.49	8585.0	0.7072354540000000	XX	2.3_Anatolia_10000BP_8000BP
FarmerAnatolian	Tep004	Turkey	Anatolia_Neolithic	38.17	34.49	8295.0	0.467990841	XX	2.3_Anatolia_10000BP_8000BP
FarmerEarly	R3	Italy	Italy_Neolithic	41.96	13.54	7729.5	4.059641042	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R17	Italy	Italy_Neolithic	43.72	13.03	7223.5	0.56582164	XY	2.3_EuropeS_8000BP_6000BP

FarmerEarly	R19	Italy	Italy_Neolithic	43.72	13.03	7233.0	0.52086802	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R18	Italy	Italy_Neolithic	43.72	13.03	7298.5	0.6271842	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R16	Italy	Italy_Neolithic	43.72	13.03	7207.5	0.565514455	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R10	Italy	Italy_Neolithic	41.96	13.54	7629.0	1.321819486 0000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R2	Italy	Italy_Neolithic	41.96	13.54	7984.0	3.704063854 0000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	R9	Italy	Italy_Neolithic	41.96	13.54	7496.0	4.04251971	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	MDV248	France	France_Neolithic_LBK	49.42	4.01	7015.5	0.121146489 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	ROS45	France	France_Neolithic_Grossgartach	48.5	7.47	6642.5	0.269292799 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	ROS78	France	France_Neolithic_Grossgartach	48.5	7.47	6550.0	0.435416491	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	Sch72-15	France	France_Neolithic_LBK	48.76	7.6	7036.5	0.235222248	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	Schw432	France	France_Neolithic_LBK	48.76	7.6	7100.0	0.148141996	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO137	Hungary	Hungary_Neolithic_Koros	46.42	20.33	7591.0	0.198402858 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO140	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6718.0	0.145142912 00000000	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO145	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6744.0	0.218956151 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO147	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6724.0	0.284571059 00000000	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	NEO695	Italy	Italy_Neolithic	43.08	13.06	7299.0	0.431411103 00000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	NEO674	Romania	Romania_Neolithic	44.9	22.43	5570.0	0.237000547	XY	2.3_Europe_8500BP_5500BP
FarmerEarly	R8	Italy	Italy_Neolithic	41.96	13.54	7723.5	0.53180897	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	kol6	CzechRepublic	Czech_Neolithic_Megalithic	50.03	15.2	6690.0	1.543500893	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NEO655	Serbia	Serbia_Mesolithic	44.54	22.04	8668.0	0.225492674 00000000	XX	2.3_Europe_8500BP_5500BP

FarmerEarly	PL_N36	Poland	Poland_Neolithic_BKG	50.67	21.38	6250.0	1.669980973 0000000	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NE5	Hungary	Hungary_Neolithic_ALP	47.17	20.83	7050.0	0.784933195	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N31	Poland	Poland_Neolithic_BKG	52.61	18.8	6137.5	3.003824151 0000000	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NE1	Hungary	Hungary_Neolithic_ALP	47.85	21.15	7138.5	18.42090683	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Stuttgart	Germany	Germany_Neolithic	48.78	9.18	7140.0	16.19254244	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Bar31	Turkey	Anatolia_Neolithic	40.3	29.61	8278.5	3.649090271 0000000	XY	2.3_Anatolia_10000BP_8000BP
FarmerEarly	Bar8	Turkey	Anatolia_Neolithic	40.3	29.61	8071.0	7.171025264	XX	2.3_Anatolia_10000BP_8000BP
FarmerEarly	NE6	Hungary	Hungary_Neolithic_LBK	47.17	19.83	7051.5	0.937998868	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N25	Poland	Poland_Neolithic_BKG	52.61	18.8	6250.0	2.304346802	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N26	Poland	Poland_Neolithic_BKG	50.67	21.38	6151.0	2.146927132	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N19	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.0	1.684160473 0000000	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Pal7	Greece	Greece_Neolithic	40.51	22.5	6351.0	1.265496659	XX	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	PL_N28	Poland	Poland_Neolithic_BKG	52.61	18.8	6073.5	1.75499685	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	Rev5	Greece	Greece_Neolithic	40.33	22.56	8301.0	1.133693538	XX	2.3_EuropeS_8000BP_6000BP
FarmerEarly	Klei10	Greece	Greece_Neolithic	40.26	21.74	6062.5	2.047213233	XY	2.3_EuropeS_8000BP_6000BP
FarmerEarly	NE2	Hungary	Hungary_Neolithic_ALP	47.52	21.59	7123.5	0.148202309	XX	2.3_Europe_8500BP_5500BP
FarmerEarly	PL_N27	Poland	Poland_Neolithic_BKG	52.61	18.8	6250.0	1.790218266	XY	2.3_EuropeCE_7000BP_5500BP
FarmerEarly	NE7	Hungary	Hungary_Neolithic_Lengyel	47.17	19.83	6374.0	0.909572155	XY	2.3_EuropeCE_7000BP_5500BP
FarmerLate	NEO119	France	France_Neolithic	44.47	4.77	4382.0	0.10885148	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO886	Denmark	Denmark_Neolithic	54.99	12.42	5457.0	0.271693685	XY	2.4_EuropeNE_5600BP_4600BP

FarmerLate	NEO925	Denmark	Denmark_Neolithic	54.77	10.68	4947.0	0.294332844	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO653	Spain	Iberia_BronzeAge	43.4	-4.71	3423.0	0.283734421	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	TV3831	Portugal	Iberia_BronzeAge	37.94	-7.6	3550.0	0.994305724 0000000	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	COV20126	Spain	Iberia_Neolithic	37.41	-4.42	5588.0	0.303266123	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO896	Denmark	Denmark_Neolithic	54.97	12.49	5446.0	0.121712208	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO43	Denmark	Denmark_Neolithic	55.99	10.25	5067.0	0.108400524	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO39	Sweden	Sweden_Neolithic	55.57	13.04	5074.0	0.174678265	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	TV32032extra	Portugal	Iberia_BronzeAge	37.94	-7.6	3550.0	0.865992675	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO744	Denmark	Denmark_Neolithic	55.86	11.59	5333.0	0.220878469	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	MonteGato104	Portugal	Iberia_BronzeAge	38.02	-7.86	3535.0	1.236961027	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO830	Italy	Italy_Neolithic	43.38	13.55	5393.0	0.104506976	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO757	Denmark	Denmark_Neolithic	55.9	11.12	5452.0	0.129920601	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	atp9	Spain	Iberia_BronzeAge	42.35	-3.52	3634.0	0.419041583	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	QUIN234	France	France_BronzeAge	43.3	1.96	3600.0	0.120622268	XX	2.2_EuropeSW_6000BP_3500BP
FarmerLate	ROS102	France	France_Neolithic_Grossgartach	48.5	7.47	6550.0	0.151456852	XY	2.3_Europe_8500BP_5500BP
FarmerLate	NEO640	Poland	Poland_Neolithic_FBC	50.27	20.45	4902.0	0.20055724	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO609	Portugal	Iberia_Neolithic	38.68	-9.16	4333.0	0.134584406	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	NEO599	Denmark	Denmark_Neolithic	55.55	11.68	5134.0	0.19019233	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO597	Denmark	Denmark_Neolithic	55.59	11.57	5210.0	0.177413333	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO595	Denmark	Denmark_Neolithic	55.79	11.29	5452.0	0.218556781	XY	2.4_EuropeNE_5600BP_4600BP

FarmerLate	ans016	Sweden	Sweden_Neolithic	57.34	18.26	4646.0	0.34085699700000000	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO945	Denmark	Denmark_Neolithic	56.49	9.83	5445.0	1.384502704	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO25	Denmark	Denmark_Neolithic	56.43	10.79	4956.0	0.35636724800000000	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	ans005	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5265.0	0.129055433	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO721	Spain	Iberia_Neolithic	40.44	-3.5	4170.0	0.35821067700000000	XX	2.2_Iberia_7300BP_3500BP
FarmerLate	NEO943	Denmark	Denmark_Neolithic	55.46	9.69	4614.0	1.754137916	XY	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO942	Denmark	Denmark_Neolithic	55.58	11.29	5491.0	0.890575166	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	NEO753	Denmark	Denmark_Neolithic	55.77	12.21	5531.0	0.163364465	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	esp005	Spain	Iberia_BronzeAge_Cogotas	42.0	-1	3370.0	2.457932356	XY	2.2_Iberia_7300BP_3500BP
FarmerLate	pir001	Spain	Iberia_BronzeAge_Argar	37.89	-4.78	3725.0	0.21560743100000000	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	por004	Spain	Iberia_Neolithic	42.35	-3.52	4955.0	0.129932409	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	san216	Spain	Iberia_Neolithic	42.69	-2.73	5665.5	0.200001027	XY	2.2_EuropeSW_6000BP_3500BP
FarmerLate	ans003	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5250.0	0.135654055	XX	2.4_EuropeNE_5600BP_4600BP
FarmerLate	ValeOuro10207	Portugal	Iberia_BronzeAge	38.06	-8.11	3550.0	0.248412062	XX	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO121	France	France_Neolithic	44.47	4.77	4531.0	0.530892193	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO28	Denmark	Denmark_Neolithic	55.91	12.31	5459.0	0.922300627	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO36	Sweden	Sweden_Neolithic	55.57	13.04	5097.0	2.384739487	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO29	Denmark	Denmark_Neolithic	55.13	10.9	5489.0	0.530600133	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO23	Denmark	Denmark_Neolithic	55.6	11.31	5533.0	3.33772768	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	WET370	France	France_Neolithic	48.06	7.3	5521.0	0.172865405	XY	2.2_EuropeAtlantic_7000BP_5000BP

FarmerMiddle	NEO866	Denmark	Denmark_Neolithic	54.87	11.84	5456.0	1.521286722	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO891	Denmark	Denmark_Neolithic	55.73	12.1	5661.0	0.595471133	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	ans017	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5080.0	7.08351096	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CO1	Hungary	Hungary_Neolithic_Baden	47.17	19.83	4750.0	0.8595327420000000	XX	2.3_EuropeS_8000BP_6000BP
FarmerMiddle	RISE489	Italy	Italy_Neolithic	45.26	10.38	4693.0	0.50758911	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO641	Poland	Poland_Neolithic_FBC	50.27	20.45	5132.0	0.26184454	XX	2.3_EuropeS_8000BP_6000BP
FarmerMiddle	NEO630	UK	Britain_Neolithic	58.73	-2.94	4898.0	0.211902615	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO627	UK	Britain_Neolithic	58.73	-2.94	5132.0	0.44952267100000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO626	UK	Britain_Neolithic	58.73	-2.94	5082.0	0.37193905200000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO624	UK	Britain_Neolithic	58.73	-2.94	4897.0	2.099933966	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO717	UK	Britain_Neolithic	58.73	-2.94	4893.0	0.481287489	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO935	Denmark	Denmark_Neolithic	55.56	12.02	5187.0	5.027509563	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	NEO933	Denmark	Denmark_Neolithic	55.25	10.75	5337.0	0.522015892	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	Aveline_1	UK	Britain_Neolithic	51.32	-2.75	5489.5	0.782428624	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO142	Hungary	Hungary_Neolithic_Tisza	46.37	20.42	6641.0	0.895629723	XX	2.3_Europe_8500BP_5500BP
FarmerMiddle	BurnGround	UK	Britain_Neolithic	51.84	-1.85	5770.0	0.410850116	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	CaveHa3_1	UK	Britain_Neolithic	54.07	-2.29	5264.0	0.112689685	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Coldrum_1	UK	Britain_Neolithic	51.32	0.37	5430.0	0.547611201	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Embo_1	UK	Britain_Neolithic	57.91	-4	5050.0	0.127757081	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Fussels_Lodge_1	UK	Britain_Neolithic	51.09	-1.73	5657.5	0.73024687	XX	2.2_EuropeAtlantic_7000BP_5000BP

FarmerMiddle	Jubilee_cave	UK	Britain_Neolithic	54.08	-2.27	5462.5	0.116653425	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Kelco_cave	UK	Britain_Neolithic	54.07	-2.29	5536.0	0.115486462	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	Ballynahatty	Ireland	Ireland_Neolithic	54.54	-5.96	5131.5	10.0157214	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N18	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.5	1.909978283	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	PSS4693	France	France_Neolithic_Noyen	48.52	3.6	5438.5	0.740638747	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N38	Poland	Poland_Neolithic_GAC	52.61	18.9	5033.0	1.772002823 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	Carsington_pasture_1	UK	Britain_Neolithic	53.08	-1.64	5538.5	9.748563794 000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	PL_N20	Poland	Poland_Neolithic_FBC	52.62	18.96	5462.0	0.802395212 0000000	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	Mor6	France	France_Neolithic_LBK	48.82	7.63	7036.0	0.161199361 00000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	ros3	Sweden	Sweden_Neolithic_FBC	60.26	16.41	4955.0	0.380939671 00000000	XX	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	RISE1161	Poland	Poland_Neolithic_GAC	48.7	21.2	4757.0	1.366034772	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1165	Poland	Poland_Neolithic_GAC	48.7	21.2	4742.5	2.189830631	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1166	Poland	Poland_Neolithic_GAC	48.7	21.2	4907.5	3.058742732	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1249	Poland	Poland_Neolithic_GAC	50.2	21.4	4736.5	1.010443069 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1248	Poland	Poland_Neolithic_GAC	50.2	21.4	4725.0	0.799515376 0000000	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1246	Poland	Poland_Neolithic_GAC	50.2	21.4	4715.0	0.549865829	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1252	Poland	Poland_Neolithic_GAC	50.8	21.5	4725.0	0.430947658 00000000	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1250	Poland	Poland_Neolithic_GAC	50.6	21.7	4725.0	0.493666575	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1254	Poland	Poland_Neolithic_GAC	51.1	17.1	4725.0	0.437402634	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	Gok2	Sweden	Sweden_Neolithic_FBC	58.18	13.41	4850.0	1.220935916	XX	2.4_EuropeNE_5600BP_4600BP

FarmerMiddle	NEO847	UK	Britain_Neolithic	51.7	-2.3	5463.0	1.777915755	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	ros005	Sweden	Sweden_Neolithic_FBC	60.26	16.41	4740.0	0.886088897	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	PEI 2.00	France	France_Neolithic_Campaniforme	43.14	2.25	4385.5	0.302517672	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	R1014	Italy	Italy_Neolithic	41.37	13.29	4950.0	0.615266206	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R104	Italy	Italy_LateAntiquity	41.89	12.48	1450.0	0.879014223	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	RISE1159	Poland	Poland_Neolithic_GAC	48.7	21.2	4730.0	27.46258284	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1170	Poland	Poland_Neolithic_GAC	48.7	21.2	4748.5	3.79009955	XX	2.4_Poland_5000BP_4700BP
FarmerMiddle	RISE1168	Poland	Poland_Neolithic_GAC	48.7	21.2	4676.0	18.93418868	XY	N/A
FarmerMiddle	mur	Spain	Iberia_Neolithic_Almagra	42.35	-3.52	7136.0	3.467561449 0000000	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	lai001	UK	Britain_Neolithic	59.13	-3.05	5180.0	0.225980982	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	mid001	UK	Britain_Neolithic_Megalithic	59.13	-3.05	5450.0	0.282625993	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	mid002	UK	Britain_Neolithic_Megalithic	57.75	-3.92	5180.0	0.255787154 00000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs002	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5675.0	5.870842597	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs003	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5600.0	0.223787208 00000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs006	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5330.0	0.263524442 00000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs009	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5310.0	7.571866381	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO823	Italy	Italy_BronzeAge	40.88	16.73	4665.0	0.404577629	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	prs010	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5530.0	0.232057468 00000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs013	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5320.0	4.952996849	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	prs016	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5560.0	8.754512586	XY	2.2_EuropeAtlantic_7000BP_5000BP

FarmerMiddle	ans008	Sweden	Sweden_Neolithic_Megalithic	57.34	18.26	5135.0	2.027487661	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CB13	Spain	Iberia_Neolithic_Cardial	41.37	1.89	7348.0	0.931947851	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	atp016	Spain	Iberia_Neolithic	42.35	-3.52	5039.5	13.20832827	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	atp12-1420	Spain	Iberia_Neolithic	42.35	-3.52	4895.5	2.528221016	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	c40331	Spain	Iberia_Neolithic	37.37	-4.25	5649.5	0.293459982	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	prs012	Ireland	Ireland_Neolithic_Megalithic	54.25	-8.56	5660.0	0.25153112700000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	LugarCanto44	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	2.016550504	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	RISE1241	Poland	Poland_Neolithic_GAC	50.6	21.7	4752.5	0.859098485	XY	2.4_Poland_5000BP_4700BP
FarmerMiddle	R22	Italy	Italy_Neolithic	40.81	8.44	3895.5	0.776104455	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	BERG157-2	France	France_Neolithic_BORSMichelsberg	43.22	2.41	6050.0	0.346274491	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	BERG157-7	France	France_Neolithic_BORSMichelsberg	43.22	2.41	6131.5	0.267013925	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	CabecoArruda117B	Portugal	Iberia_Neolithic	39.11	-8.66	5050.0	0.376607974	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	BLP10	France	France_Neolithic_Michelsberg	49.39	3.74	6052.0	0.18285663600000000	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	BUCH2	France	France_Neolithic_Cemy	48.24	4.11	6250.0	0.36476039200000000	XY	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	NEO812	France	France_Neolithic_Cardial	43.32	2.42	6545.0	6.542360034	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	CRE20D	France	France_Neolithic_ChasseenAncien	43.21	3.13	6151.0	0.256045807	XX	2.2_EuropeSW_6000BP_3500BP
FarmerMiddle	LugarCanto42	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	3.006333862	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LU339	Portugal	Iberia_Neolithic	41.71	-6.93	6797.5	4.60334026	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LD270	Portugal	Iberia_Neolithic	41.71	-6.93	6336.0	4.064587193	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	LD1174	Spain	Iberia_Neolithic	37.41	-4.42	6415.0	3.558801721	XX	2.2_Iberia_7300BP_3500BP

FarmerMiddle	CabecoArruda122A	Portugal	Iberia_Neolithic	39.11	-8.66	5050.0	1.782958508	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	CovaMoura364	Portugal	Iberia_Neolithic	38.75	-9.22	4900.0	0.794100402	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	Es97-1	France	France_Neolithic_Michelsberg	50.92	1.71	6004.5	0.294790665	XY	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	NEO790	Denmark	Denmark_Neolithic	55.71	12.27	5663.0	0.685227719	XY	2.4_EuropeNE_5600BP_4600BP
FarmerMiddle	CovaMoura9B	Portugal	Iberia_Neolithic	38.75	-9.22	4900.0	2.611737333	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	R6	Italy	Italy_Neolithic	41.96	13.54	7159.5	0.604714196	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	bal004	UK	Britain_Neolithic_Megalithic	57.77	-3.9	5190.0	1.5679103270000000	XX	2.2_EuropeAtlantic_7000BP_5000BP
FarmerMiddle	R24	Italy	Italy_Neolithic	40.81	8.44	5450.0	0.549967737	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	R25	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.53976087	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	R26	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.525953541	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R27	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.70739547	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R29	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.559161215	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R28	Italy	Italy_Neolithic	40.81	8.44	4150.0	0.728827575	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	Dolmen Ansião 96B	Portugal	Iberia_Neolithic	39.75	-8.81	5450.0	1.962153759	XY	2.2_Iberia_7300BP_3500BP
FarmerMiddle	R4	Italy	Italy_Neolithic	41.96	13.54	4865.0	3.6761982310000000	XY	2.2_Italy_7000BP_4000BP
FarmerMiddle	R5	Italy	Italy_Neolithic	41.96	13.54	4839.5	1.5029056810000000	XX	2.2_Italy_7000BP_4000BP
FarmerMiddle	LugarCanto41	Portugal	Iberia_Neolithic	39.42	-8.82	5950.0	1.06714512	XX	2.2_Iberia_7300BP_3500BP
FarmerMiddle	BERG02-2	France	France_Neolithic_BORSMichelsberg	43.22	2.41	5870.0	0.344146565	XX	2.2_EuropeSW_6000BP_3500BP
WHG	NEO855	Denmark	Denmark_Mesolithic	56.4	10.72	6302.0	1.3829770000000000	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO856	Denmark	Denmark_Mesolithic	56.37	10.64	6777.0	0.56205807	XX	4.2_Denmark_10500BP_6000BP

WHG	NEO679	Sweden	Sweden_Mesolithic	55.39	13.48	6834.0	0.164673359	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO683	Denmark	Denmark_Mesolithic	55.4	9.83	7529.0	1.81852777	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO938	Spain	Iberia_Mesolithic	43.4	-4.71	7878.0	0.47515145700000000	XX	4.2_Iberia_9000BP_7000BP
WHG	NEO694	Spain	Iberia_Mesolithic	38.73	-0.46	9217.0	0.28475883400000000	XY	4.2_Iberia_9000BP_7000BP
WHG	NEO853	Denmark	Denmark_Mesolithic	55.55	10.62	6047.0	1.964862968	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO852	Denmark	Denmark_Mesolithic	56.03	10.26	6308.0	0.189591791	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO733	Denmark	Denmark_Mesolithic	55.77	11.39	6824.0	1.316681951 0000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO791	Denmark	Denmark_Mesolithic	55.33	11.15	7048.0	2.492448215	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO941	Denmark	Denmark_Mesolithic	56.71	10.17	6372.0	0.135296816	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO751	Denmark	Denmark_Mesolithic	56.87	9.22	6343.0	0.297822065	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO932	Denmark	Denmark_Mesolithic	55.25	11.23	7499.0	2.760162200 0000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO749	Denmark	Denmark_Mesolithic	55.85	12.56	7070.0	1.905133435	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO747	Denmark	Denmark_Mesolithic	55.85	12.56	6729.0	0.24942735800000000	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO745	Denmark	Denmark_Mesolithic	55.85	12.56	6790.0	0.447895875	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO960	Denmark	Denmark_Mesolithic	55.58	11.58	5926.0	0.150141897	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO759	Denmark	Denmark_Mesolithic	55.4	12.37	9028.0	2.948103712	XY	4.2_Denmark_10500BP_6000BP
WHG	syltholm	Denmark	Denmark_Mesolithic	54.65	11.35	7709.5	2.291787968	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO648	Spain	Iberia_Mesolithic	43.4	-4.71	7539.0	1.844283075	XY	4.2_Iberia_9000BP_7000BP
WHG	Cheddar_man	UK	Britain_Mesolithic	51.28	-2.77	10300.0	2.054202123	XY	4.2_EuropeW_13500BP_8000BP
WHG	PL_N22	Poland	Poland_Neolithic_BKG	52.61	18.9	6291.0	1.491551035	XX	4.2_EuropeE_8600BP_6000BP
WHG	KO1	Hungary	Hungary_Neolithic_Koros	47.56	20.72	7660.0	1.014600016	XY	4.2_EuropeE_8600BP_6000BP

WHG	Canes1	Spain	Iberia_Mesolithic	43.36	-4.72	7115.0	1.6461215340000000	XX	4.2_Iberia_9000BP_7000BP
WHG	Chan	Spain	Iberia_Mesolithic	42.73	-7.03	9131.0	5.008215765	XX	4.2_Iberia_9000BP_7000BP
WHG	Bichon	Switzerland	Switzerland_Mesolithic	47.1	6.87	13665.0	7.692393112	XY	4.2_EuropeW_13500BP_8000BP
WHG	Loschbour	Luxembourg	Luxembourg_Mesolithic	49.81	6.4	8050.0	18.23029647	XY	4.2_EuropeW_13500BP_8000BP
WHG	Brana	Spain	Iberia_Mesolithic	42.91	-5.38	7815.0	3.019525774	XY	4.2_Iberia_9000BP_7000BP
WHG	R11	Italy	Italy_Mesolithic	41.96	13.54	11908.0	0.957641603	XY	4.2_Italy_15000BP_9000BP
WHG	NEO669	Serbia	Serbia_Mesolithic	44.56	22.03	7950.0	0.23932044	XX	4.2_EuropeE_8600BP_6000BP
WHG	R7	Italy	Italy_Mesolithic	41.96	13.54	10681.5	3.153769086	XY	4.2_Italy_15000BP_9000BP
WHG	ST3	Italy	Italy_Mesolithic	37.85	14.7	14800.0	0.475034886	XY	4.2_Italy_15000BP_9000BP
WHG	PER1150503	France	France_Mesolithic	45.77	0.33	9067.0	0.315139903	XX	4.2_EuropeW_13500BP_8000BP
WHG	PER3023	France	France_Mesolithic	45.77	0.33	9067.0	0.161333797	XX	4.2_EuropeW_13500BP_8000BP
WHG	R15	Italy	Italy_Mesolithic	41.96	13.54	9124.5	3.070164483	XY	4.2_Italy_15000BP_9000BP
WHG	NEO91	Denmark	Denmark_Mesolithic	55.39	12.31	9122.0	1.176549838	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO646	Spain	Iberia_Mesolithic	43.4	-4.71	8273.0	1.590267827	XX	4.2_Iberia_9000BP_7000BP
WHG	NEO645	Denmark	Denmark_Mesolithic	55.91	11.09	5870.0	0.211986752	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO598	Denmark	Denmark_Mesolithic	55.95	11.9	6075.0	0.7272044320000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO19	Denmark	Denmark_Mesolithic	56.27	10.47	8163.0	3.262849417	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO586	Denmark	Denmark_Mesolithic	56.37	10.57	7031.0	0.201188562	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO583	Denmark	Denmark_Mesolithic	56.37	10.57	6981.0	0.17639908900000000	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO570	Denmark	Denmark_Mesolithic	56.4	10.72	6369.0	2.861753026	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO589	Denmark	Denmark_Mesolithic	55.33	11.15	7478.0	7.410700578000000	XY	4.2_Denmark_10500BP_6000BP

WHG	NEO254	Denmark	Denmark_Mesolithic	55.46	10.13	10463.0	0.41962998	XY	4.2_Denmark_10500BP_6000BP
WHG	NEO123	Denmark	Denmark_Mesolithic	54.96	11.85	8182.0	0.286386228	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO122	Denmark	Denmark_Mesolithic	54.96	11.85	8146.0	0.564966744	XX	4.2_Denmark_10500BP_6000BP
WHG	NEO568	Denmark	Denmark_Mesolithic	56.81	9.18	6586.0	1.9814869470000000	XY	4.2_Denmark_10500BP_6000BP
Yamnaya	poz81	Poland	Poland_Neolithic_CWC	52.29	17.55	4705.0	1.92879788	XY	1.2_EuropeNE_4800BP_3000BP
Yamnaya	RISE509	Russia	Siberia_BronzeAge_Afanasievo	54.36	90.92	4732.0	4.52834127	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE511	Russia	Siberia_BronzeAge_Afanasievo	54.36	90.92	4744.0	5.20403929	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE546	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4850.0	0.125905828	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE547	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4710.5	0.6864666010000000	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE548	Russia	Russia_BronzeAge_Yamnaya	46.54	43.7	4850.0	0.9108783580000000	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE550	Russia	Russia_BronzeAge_Yamnaya	46.56	43.68	4934.5	0.440260727	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE555	Russia	Russia_BronzeAge	48.72	44.5	4627.0	0.237337432	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	Yamnaya	Kazakhstan	Kazakhstan_BronzeAge_Yamnaya	49.13	75.85	4902.5	26.39165529	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	MJ-09	Ukraine	Ukraine_BronzeAge_Catacomb	47.43	34.27	4285.5	0.19947548700000000	XX	1.2_Steppe_5000BP_4300BP
Yamnaya	MJ-06	Ukraine	Ukraine_BronzeAge_Yamnaya	49.32	35.37	4629.5	0.161999877	XX	1.2_EuropeNE_4800BP_3000BP
Yamnaya	NEO175	Russia	Russia_Neolithic_Sredny	52.28	38.96	4607.0	0.416698274	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	Latvia_LN1	Latvia	Latvia_Neolithic_CWC	56.28	25.13	4833.0	0.197755635	XX	1.2_EuropeNE_4800BP_3000BP
Yamnaya	RISE552	Russia	Russia_BronzeAge_Yamnaya	46.62	43.33	4446.0	2.458824579	XY	1.2_Steppe_5000BP_4300BP
Yamnaya	RISE240	Russia	Russia_BronzeAge_Yamnaya	46.58	43.68	4706.0	0.173195772	XX	1.2_Steppe_5000BP_4300BP

Table 1 | Metadata and grouping of ancient individuals into reference populations.

Appendix

Painting manual

This document outlines the steps necessary to paint a modern dataset using a pre-defined set of reference samples grouped into reference populations. It does not detail how to group the reference samples in the first place.

Starting point: imputed, phased data in VCF format. It is important that the phasing of the target data is to a similar standard as the ancient data, as poor phasing will mess up the results.

We can split the process into two main steps:

1. Data preparation
    a. Merging of datasets
    b. Subsetting and filtering SNPs
    c. File conversions
2. Painting
    a. Running painting scripts to record local and global ancestry estimates

Data preparation:

1. **Preparation of modern 'target' data:**

This will be different depending on the data format you use. For the UKB data I used hard genotype calls which were extracted from .bgen files. The main thing to be aware of is to use only software that keeps the phasing of the data (e.g. not certain plink commands…).

2. **Merge the ancient and modern datasets, subset for selected SNPs**

I did this in one step using qctool (https://www.well.ox.ac.uk/~gav/qctool_v2/), merging the ancient and modern data (both in VCF format), and subsetting the positions. For each chromosome, something like:

```
qctool -g
/willerslev/ukbiobank/nonbritish/merged_vcfs/ukbb.haplotype_chrN.GT.chrinfo.sampl
es.vcf.gz -g
```

/willerslev/users-shared/science-snm-willerslev-wl4sn3/step3_postprocessing/step5_release/15062020/chrN.haplotypes.filtered.N1664.D15062020.GLIMPSE.vcf.gz -og merged.N.vcf -os merged.N.samples -compare-variants-by position -flip-to-match-cohort1 -incl-samples keep_samples -incl-positions /willerslev/ukbiobank/SNP_selection/qctool_format/N_SNPs_formatted.txt

where keep_samples lists all ancient and modern IDs that you want to include, and N_SNPs_formatted.txt is a file containing SNPs in the UKB array. This is about ~820k SNPs chosen for markers of specific interest, rare coding variants, and genome-wide coverage. When combined with the imputed ancient data we end up with about 550K SNPs. Chromopainter will not accept missing data.

Ideally we want good coverage of the whole genome and to include markers of interest, but obviously the more you include the higher the computational cost.

3. **Filter for MAF > 0.01 and INFO**

I filtered for MAF > 0.01 using bcftools. Something like:

bcftools view -q 0.01:minor merged.N.vcf.gz -o merged.N.filtered.vcf.gz

4. **Convert VCF to phase format**

Phase format is the chromopainter input format (more info here https://people.maths.bris.ac.uk/~madjl/finestructure/manual.html). There are a few ways to convert a vcf to phasefile format, but they're all slow (i.e. takes many days/weeks) with large numbers of samples and can't be parallelised. I found that the best way to do this was to split the vcf by sample, then convert these individual vcf files to phase format and concatenate the results. It's a little hacky but speeds it up massively. To split the vcf for each chromosome N:

bcftools plugin split merged.N.filtered.vcf.gz --samples-file samples0 -Oz -o haps.N

Then using script vcf_to_phase.sh (which uses pbwt [https://github.com/richarddurbin/pbwt] and the script impute2chromopainter.pl which is included as part of fineSTRUCTURE), I wrote separate commands to convert each vcf to a phasefile. Using python (pandas) to

generate these commands, where file samples contains all sample IDs (modern and ancient):

```
>>> for chr in range(1,23):
...   df=pd.read_csv("samples", header=None)
...       df_temp=df
...       df_temp[1]="bash vcf_to_phase.sh " + df_temp[0].astype(str) + " " + str(chr)
...       df_temp=df_temp[1]
...       df_temp.to_csv(str(chr)+".commands", header=False, index=False)
```

Which creates a separate command file for each chromosome which can then be run in parallel. To then concatenate the files in the correct order for chromosome N:

```
while read p; do cat phase.N/$p.reduced.phase >> merged.N.phase; done <samples
```

To get the top three lines for the phase format, we run the same commands as in vcf_to_phase.sh just for one sample but keep the top three lines instead. For example, for chromosome N and sample X:

```
pbwt -readVcfGT haps.N/X.vcf.gz -writeImputeHapsG haps.N/X.haps
impute2chromopainter.pl haps.N/X.haps temp.N.X.phase
head -n 3 temp.N.X.phase > temp.N.header.phase
cat merged.N.phase >> temp.N.header.phase
mv temp.N.header.phase N.merged.phase
```

We now have a phasefile with all the samples in the same order as in file samples. The only thing we need to do is edit the number of haplotypes (i.e. 2x number of individuals) which is specified in the first line of the phasefile. Depending on the number, something like:

```
nhaps=SOME NUMBER
for chr in $chrlist ; do sed -i "1s/.*/$nhaps/" $chr.merged.phase; done
```

5. Make recombination input file

We need to make a file specifying recombination rates between our included SNPs. I used the recombination maps from the International Hapmap Consortium phase II release, and the command for each chromosome looked like:

convertrecfile.pl -M hapmap phasefiles/$chr.merged.phase
2011-01_phaseII_B37/genetic_map_GRCh37_chr$chr.txt
recombfiles/$chr.recombfile

convertrecfile.pl is a script included in fineSTRUCTURE.

6. Make idfile input file

The idfile needs to contain the individuals included in the phasefile in the same order, one per line. The format is:
<NAME> <POPULATION> <INCLUSION> <ignored extra info>

Where <NAME> and <POPULATION> are strings and <INCLUSION> is 1 to include an individual and 0 to exclude them.
EXAMPLE IDFILE:
Ind1 Pop1 1
Ind2 Pop1 1
Ind3 Pop2 0
Ind4 Pop2 1
Ind5 Pop2 1

It is best if individuals are named as e.g. UKB1, UKB2, UKB3 etc, otherwise there are some downstream scripts that won't work. So you need a mapping between the old IDs and the new IDs in this input file. The first two lines of my idfile looked like:
UKBB1 UKBB 1
UKBB2 UKBB 1

NB the idfile must be in the same order as the phasefile.

Painting
When the data preparation steps above are done, you should have:

- A phasefile for each chromosome
- A recombination file for each chromosome
- An idfile specifying all your ancient and modern individuals, and which population they are in.

PAINTING PROCESS

Make sure the idfile and phasefile have the target panel at the top and the reference panel below.

Step 1: Run fineSTRUCTURE on the reference panel only

fs ref_panel.cp -idfile ordered_all_pop_ids_mapped -phasefiles phasefiles/{1..22}.merged.phase -recombfiles ../recombfiles/{1..22}.recombfile -hpc 1 -go

This gives Ne and mu estimates, which can be found in ref_panel.cp file.

Step 2: Run stage03

Edit will_ancientvsancient.cp (including popneinf and popmuinf from fineSTRUCTURE run on ref panel above) and will_ancientvsancient.donors. Also update will_modernvsancient.cp with these numbers. You may want to change the name of the target panel from UKBB to something else, in which case this needs to be reflected in will_modernvsancient.donors and also the idfile.

Paint ancient panel vs ancient panel to get priors. This paints individuals in reference panel against the panel (but not including itself) using uniform prior first, to obtain the overall copying averages; then repaints using previous run as prior.

Commands:

bash will_03-paintpanelvspanel.sh

This runs will_paint_withinpanel.sh, which in turn runs will_paintsample_withinpanel.sh on each sample.

Step 3: Run stage04

I had a memory problem when running painting in parallel because each thread has to read in the entire phasefiles, so I decided to run in separate batches of 24k. Within each 24k batch, I ran will_3.5-paintvspanel1by1.py which writes commands for 24k individuals, then splits these into separate batchcommands in batch_files folder. To generate commands for this in python:

```
for j in range(1,410000,24000):
        print("dir=split_"+str(j)+"-"+str((j-1)+24000))
        print("phaselinenumber="+str((j+1)*2))
        print("phaselinenumber2="+str(2*(j+24000)+1))
```

print("nhaps=48636")

print("mkdir $dir")

print("mkdir $dir/phasefiles")

print("for chr in $chrlist ; do")

print(" touch $dir/phasefiles/$chr.merged.phase")

print(" head -n 3 phasefiles/$chr.merged.phase > $dir/phasefiles/$chr.merged.phase")

print(' awk "NR>=$phaselinenumber && NR<=$phaselinenumber2" phasefiles/$chr.merged.phase >> $dir/phasefiles/$chr.merged.phase')

print(" tail -n 636 phasefiles/$chr.merged.phase >> $dir/phasefiles/$chr.merged.phase")

print(' sed -i "1s/./$nhaps/" $dir/phasefiles/$chr.merged.phase')*

print(' echo "Copied lines to $dir/phasefiles/$chr.merged.phase"')

print("done")

print('awk "NR>='+str(j)+' && NR<='+str((j-1)+24000)+'" ordered_all_pop_ids_mapped > $dir/ordered_all_pop_ids_mapped')

print('tail -n 318 ordered_all_pop_ids_mapped >> $dir/ordered_all_pop_ids_mapped')

print("cp chr1..22_1000inds_test_noflag/cp_panel_scripts/ $dir")*

print("cp -r recombfiles/ $dir")

print("cd $dir")

print("python3 will_3.5-paintvspanel1by1.py -idfile ordered_all_pop_ids_mapped -cp_panel_scripts ../chr1..22_1000inds_test_noflag/cp_panel_scripts/")

print("mkdir batch_files")

print("split -l 1000 -d --additional-suffix=.txt paintvspanel1by1_commands.txt batch_commands")

print("mv batch_commands batch_files")*

print("cd ../")

Each of these command files (found in batch_commands, 1000 commands in each) was then run on a separate node. E.g. on computerome:

qsub -W group_list=geogenetics -A geogenetics -l nodes=1:ppn=40,walltime=100:00:00,mem=100gb -N paint_panel0 -F 00 ***run_batch.sh***

What this does: runs painting of one target individual per thread:

bash paintsample1by1.sh UKBB0 1 4 ../cp_panel_scripts

[NB script paintsample1by1.sh must be edited if size of ref panel changes!]

This makes a new temporary directory for each individual, including phasefile with 1 target + all reference. It then runs will_04-paintvspanel.sh, which in turn runs will_paint_withinpanel-b.sh. This paints the target individuals and stores the local painting output in a memory efficient format.

When these have all run, we cat the results together in the right order. Commands (where ukbb_samples is a file containing the 24000 IDs of the target individuals, one per line):

```
python3
import pandas as pd
df=pd.read_csv("ordered_all_pop_ids_mapped", sep=" ", header=None)
df=df[0]
df1 = df.head(24000)
df1.to_csv("ukbb_samples", index=False, header=False, sep=" ")
exit()
chrlist=`seq 1 22`
for chr in $chrlist
do while read p
do cat temp.$p/will_modernvsancient/painting/$chr.all_copyprobsperlocus.txt >>
$chr.master_all_copyprobsperlocus.txt
done < ukbb_samples
done

while read p; do
awk "NR==3"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.allchr.chunkco
unts.out >> ordered_all_pop_ids_mapped.allchr.chunkcounts.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.allchr.chunklen
gths.out >> ordered_all_pop_ids_mapped.allchr.chunklengths.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.allchr.mutation
probs.out >> ordered_all_pop_ids_mapped.allchr.mutationprobs.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.allchr.regionch
unkcounts.out >> ordered_all_pop_ids_mapped.allchr.regionchunkcounts.out
```

awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.allchr.regionsquaredchunkcounts.out >>
ordered_all_pop_ids_mapped.allchr.regionsquaredchunkcounts.out
done < ukbb_samples
mkdir perchrom_results
for chr in $chrlist
do while read p; do
awk "NR==3"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.chr$chr.chunkcounts.out >>
perchrom_results/ordered_all_pop_ids_mapped.chr$chr.chunkcounts.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.chr$chr.chunklengths.out >>
perchrom_results/ordered_all_pop_ids_mapped.chr$chr.chunklengths.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.chr$chr.mutationprobs.out >>
perchrom_results/ordered_all_pop_ids_mapped.chr$chr.mutationprobs.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.chr$chr.regionchunkcounts.out >>
perchrom_results/ordered_all_pop_ids_mapped.chr$chr.regionchunkcounts.out
awk "NR==2"
temp.$p/will_modernvsancient/painting/ordered_all_pop_ids_mapped.chr$chr.regionsquaredchunkcounts.out >>
perchrom_results/ordered_all_pop_ids_mapped.chr$chr.regionsquaredchunkcounts.out
done < ukbb_samples
done

Then run clean_and_repaint.sh which checks the results files, removes bad lines, and writes commands to repaint remaining individuals. E.g. on computerome:

qsub -W group_list=geogenetics -A geogenetics -l nodes=1:ppn=1,walltime=100:00:00,mem=100gb -N clean_and_repaint ***clean_and_repaint.sh***

Once repainted (i.e. you've run the new commands in batch_files), run clean_and_repaint.sh again. Repeat until there are no more commands to run (this may happen on the first time, or it might take a few goes).

General notes:

- Output of local painting is reversed order compared to recombfiles (i.e. SNP position is decreasing).
- Dependencies: if you have fineSTRUCTURE installed and it is running without problem then that should be the main hurdle overcome. Check that you are using fs 4.0.1. The scripts use python3 with various mainstream packages (e.g. numpy, pandas) which shouldn't cause issues.
- There are a number of steps I then took to convert the output files into more useful formats.

Chapter Two: Genetic Legacy of Stone Age Eurasians in non-British individuals in the UK Biobank

Preface

Much of the contents of this chapter was previously published as (§ denotes joint first authors, @ denotes joint last authors):

Population Genomics of Stone Age Eurasia

Morten E. Allentoft§, Martin Sikora§, Alba Refoyo-Martínez§, Evan K. Irving-Pease§, Anders Fischer§, William Barrie§, Andrés Ingason§, Jesper Stenderup, Karl-Göran Sjögren, Alice Pearson, Bárbara Sousa da Mota, Bettina Schulz Paulsson, Alma Halgren, Ruairidh Macleod, Marie Louise Schjellerup Jørkov, Fabrice Demeter, Maria Novosolov, Lasse Sørensen, Poul Otto Nielsen, Rasmus H.A. Henriksen, Tharsika Vimala, Hugh McColl, Ashot Margaryan, Melissa Ilardo, Andrew Vaughn, Morten Fischer Mortensen, Anne Birgitte Nielsen, Mikkel Ulfeldt Hede, Peter Rasmussen, Lasse Vinner, Gabriel Renaud, Aaron Stern, Theis Zetner Trolle Jensen, Niels Nørkjær Johannsen, Gabriele Scorrano, Hannes Schroeder, Per Lysdahl, Abigail Daisy Ramsøe, Andrei Skorobogatov, Andrew Joseph Schork, Anders Rosengren, Anthony Ruter, Alan Outram, Aleksey A. Timoshenko, Alexandra Buzhilova, Alfredo Coppa, Alisa Zubova, Ana Maria Silva, Anders J. Hansen, Andrey Gromov, Andrey Logvin, Anne Birgitte Gotfredsen, Bjarne Henning Nielsen, Borja González-Rabanal, Carles Lalueza-Fox, Catriona J. McKenzie, Charleen Gaunitz, Concepción Blasco, Corina Liesau, Cristina Martinez-Labarga, Dmitri V. Pozdnyakov, David Cuenca-Solana, David O. Lordkipanidze, Dmitri En'shin, Domingo C. Salazar-García, T. Douglas Price, Dušan Borić, Elena Kostyleva, Elizaveta V. Veselovskaya, Emma R. Usmanova, Enrico Cappellini, Erik Brinch Petersen, Esben Kannegaard, Francesca Radina, Fulya Eylem Yediay, Henri Duday, Igor Gutiérrez-Zugasti, Inna Potekhina, Irina Shevnina, Isin Altinkaya, Jean Guilaine, Jesper Hansen, Joan Emili Aura Tortosa, João Zilhão, Jorge Vega, Kristoffer Buck Pedersen, Krzysztof Tunia, Lei Zhao, Liudmila N. Mylnikova, Lars Larsson, Laure Metz, Levon Yepiskoposyan, Lisbeth Pedersen, Lucia Sarti, Ludovic Orlando, Ludovic Slimak, Lutz Klassen, Malou Blank, Manuel González-Morales, Mara Silvestrini, Maria Vretemark, Marina S. Nesterova, Marina Rykun, Mario Federico Rolfo, Marzena Szmyt, Marcin Przybyła, Mauro Calattini, Mikhail Sablin, Miluše Dobisíková, Morten Meldgaard, Morten Johansen, Natalia Berezina, Nick Card, Nikolai A. Saveliev, Olga Poshekhonova, Olga Rickards, Olga V. Lozovskaya, Olivér Gábor, Otto Christian Uldum, Paola Aurino, Pavel Kosintsev, Patrice Courtaud, Patricia Ríos, Peder Mortensen, Per Lotz, Per Persson, Pernille Bangsgaard, Peter de Barros Damgaard, Peter Vang Petersen, Pilar Prieto Martinez, Piotr Włodarczak, Roman V. Smolyaninov, Rikke Maring, Roberto Menduiña, Ruben Badalyan, Rune Iversen, Ruslan Turin, Sergey Vasilyiev, Sidsel Wåhlin, Svetlana Borutskaya, Svetlana Skochina, Søren Anker Sørensen, Søren H. Andersen,

Thomas Jørgensen, Yuri B. Serikov, Vyacheslav I. Molodin, Vaclav Smrcka, Victor Merz, Vivek Appadurai, Vyacheslav Moiseyev, Yvonne Magnusson, Kurt H. Kjær, Niels Lynnerup, Daniel J. Lawson, Peter H. Sudmant, Simon Rasmussen, Thorfinn Korneliussen@, Richard Durbin@, Rasmus Nielsen@, Olivier Delaneau@, Thomas Werge@, Fernando Racimo@, Kristian Kristiansen@, Eske Willerslev@

bioRxiv 2022.05.04.490594; doi: https://doi.org/10.1101/2022.05.04.490594
In review at Nature, January 2023.
It has been modified to fit the style of a dissertation.

I performed all analyses described here. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

Chapter summary

In this chapter I develop methods to select individuals in the UK Biobank born abroad and of a typical ancestral background for that country. This extends the use cases of this resource and other Biobanks by offering researchers an easy way to identify individuals who are representative of the ancestry of a country represented in the UK Biobank. I then present average country-level NNLS ancestry components based on these selections, using the painting data described in Chapter One. I also discuss the relationship between specific ancestry components and principal components (PCs).

Introduction

The UK Biobank (UKB) contains approximately 40,000 individuals not born in the UK (Figure 1). Most studies discard these and focus on ‘white British’ individuals; these are identified by self-reporting, and then a core group is selected based on Principal Component Analysis (PCA) (Bycroft et al., 2018). This has the effect of exacerbating existing health inequalities by focussing research on white, European populations. Furthermore, as wealthier developed nations, and the UK in particular, continue to invest in genomics research and data generation, this will continue to be a problem: many of the largest datasets in the future will consist of sequencing data for individuals from European countries. It is therefore crucial to develop methods to utilise the individuals who are not genetically mainstream (i.e. in the largest cluster) in these datasets.

These non-British born individuals are an untapped source of genetic variation. In this chapter, I develop methods to select individuals in the UKB who are of a typical ancestral background in each country. Because many of the individuals born abroad are admixed or British, I set up a pipeline to (1) exclude genetically British-like individuals and (2) select individuals of a typical genetic ancestral background for each country. I use these results in order to investigate the genetic contribution of each ancient ancestry to modern European, Asian and African populations.

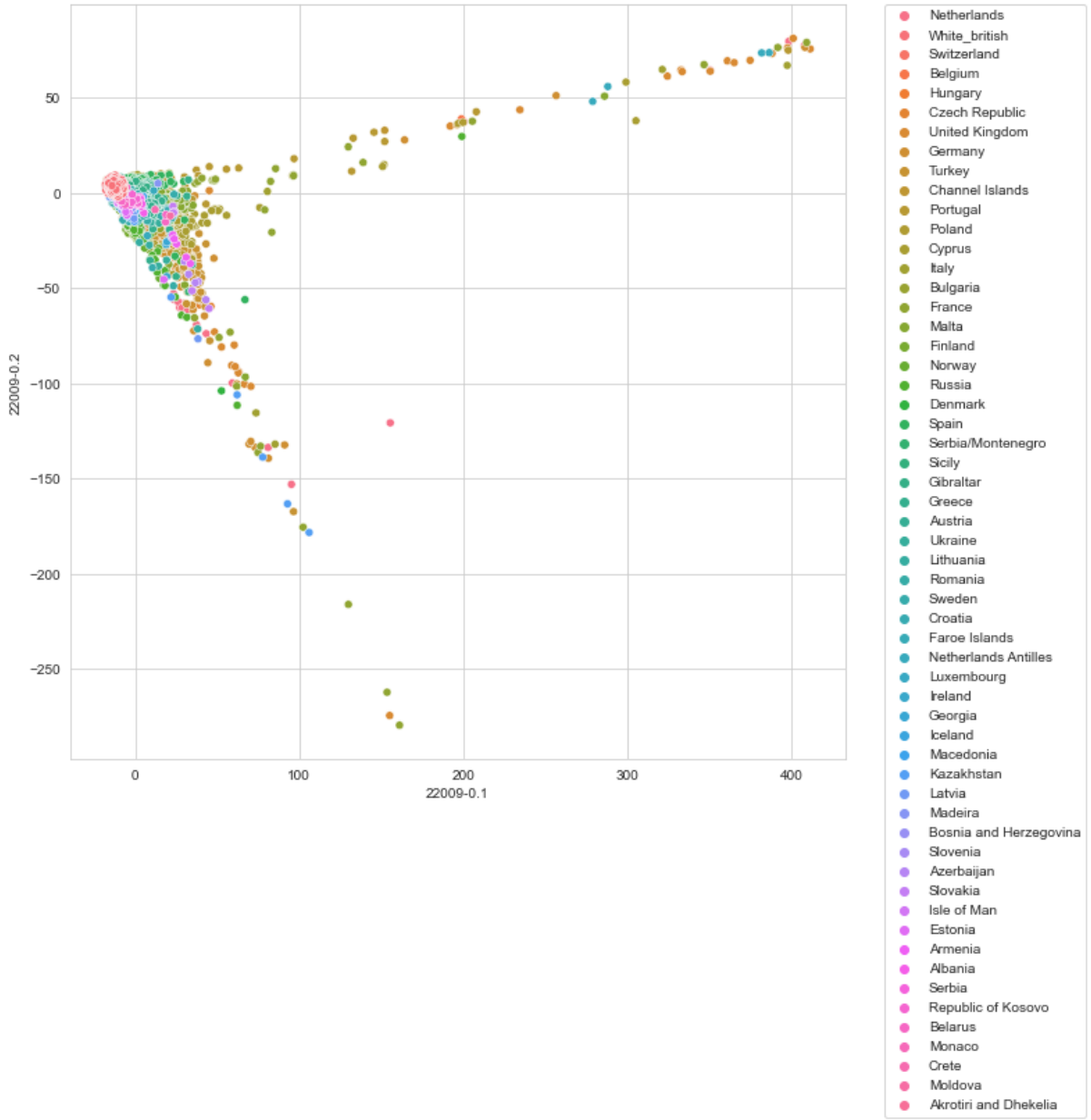


Figure 1 | PC1 vs PC2 of all non-'white British' samples in the UK Biobank born in Europe
Principal component 1 vs Principal components 2 of all non-'white British' samples (n=18,164) in the UK Biobank born in Europe, shaded by country of birth. PCs are from Bycroft et al. (2018). PCs 1 and 2 separate out samples based on their African and East Asian ancestry.

Methods

For individuals from each country in Europe, Asia and Africa but not the UK, (Data-Field 1647: Country of birth (UK/elsewhere) and Data-Field 20115: Country of Birth (non-UK origin)), I ran two density-based scans using Scikit-learn (0.21.2)'s (Pedregosa et al., 2011) DBSCAN method (Density-Based Spatial Clustering of Applications with Noise) on a distance matrix constructed using the first 18 PCs (Bycroft et al., 2018), weighted by their Eigenvalues. This algorithm finds areas ('cores') of high density within a distance matrix; samples within this area are known as 'core samples'. A cluster consists of these core samples as well as nearby non-core samples which are close to the core. Cores can be any shape. The *eps* parameter can be adjusted to determine how strict the clustering is, by defining the maximum distance between core samples and between core samples and non-core samples. Strictly speaking, a core sample is one where there are at least *min_samples* other samples which are within *eps* distance, and non-core samples are within *eps* distance of the core. The *min_samples* parameter therefore defines how many cores are found. Intuitively, core samples are found in areas of high density in the vector space, while non-core samples are on the fringes of these areas. This method is preferable to using visual PC cut-offs, for which it is difficult to include higher PCs; and to k-means clustering, which assumes clusters are convex and all points must be clustered.

In the first scan, designed to remove individuals with British-like ancestry who were born abroad, individuals from a given country were combined with 8,000 random white British individuals (i.e. significantly more than were born in the non-British country), and the clustering algorithm was run on the combined data (example in Figure 2). Any individuals born abroad who clustered with the white British were excluded (*eps*=60). For countries that are very similar to Britain in ancestry (e.g. Germany, Denmark) this is a balance between excluding individuals who are genuinely British (very common in the 'German' samples due to British military airbases there) and not biassing the samples away from British-like ancestry.

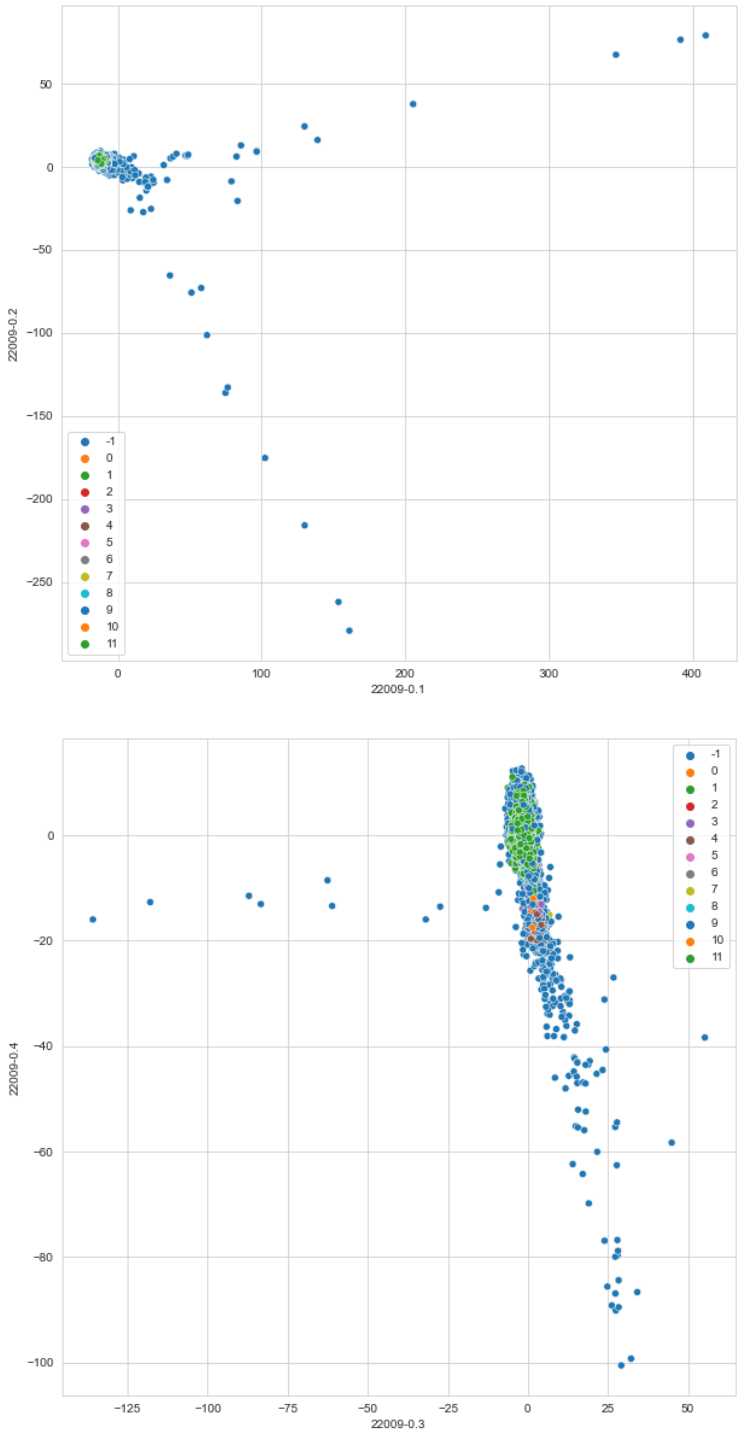


Figure 2 | Example of the first DBSCAN

PC1 vs PC2 (top) and PC3 vs PC4 (bottom) of samples born in France and 8,000 ‘white British’ samples. Samples are coloured by the cluster inferred by DBSCAN (eps=60). In this case, there are 856 samples born in France; of these, 99 fall within the largest cluster along with the ‘white British’ samples (cluster 1, green) and are excluded from further analysis.

In the second scan, the remaining individuals were clustered, and the largest cluster was chosen to represent a typical ancestry for that country (Example in Figure 3). The appropriate *eps* value (i.e. how strict the clustering should be) is a reflection of the genetic diversity of a country, and so was adjusted manually to reflect this (Table 1). In a minority of cases, the largest cluster was not the indigenous ancestral background, and so the second-largest was chosen. For example, in Kenya the largest cluster consisted of individuals of Indian origin. All selections were visually verified. Countries that had no obvious main cluster (usually due to low sample numbers) were excluded; any country with 3 or fewer individuals was also excluded.

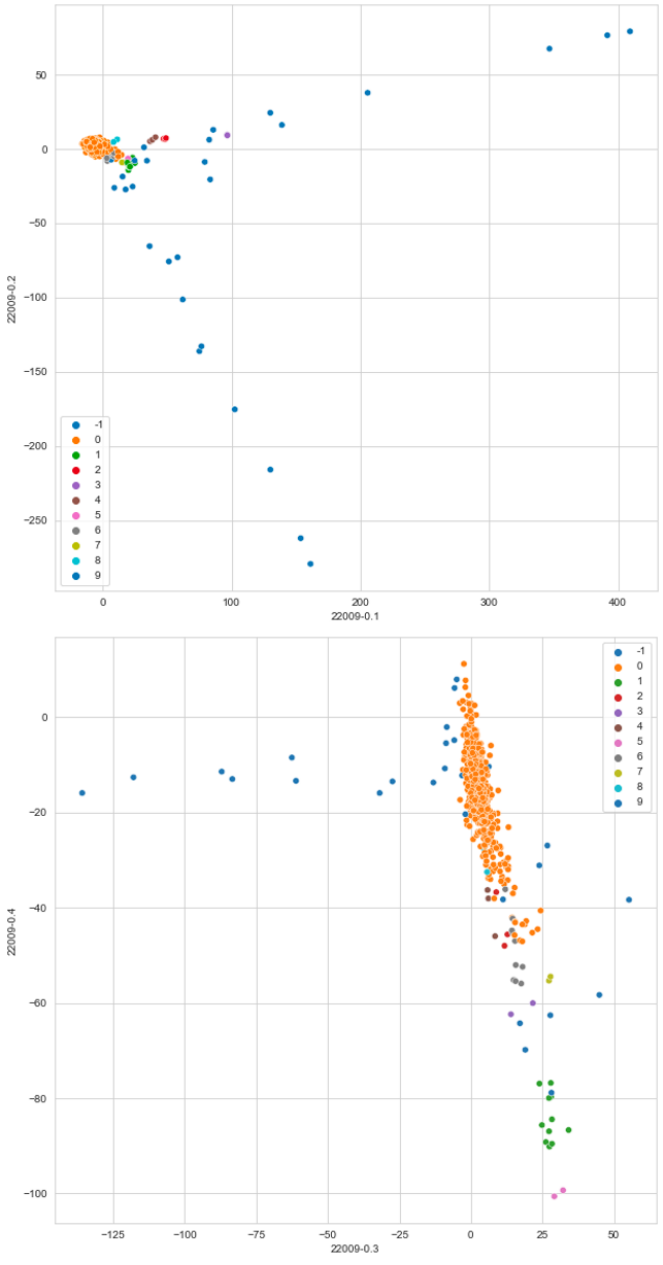


Figure 3 | Example of the second DBSCAN

PC1 vs PC2 (top) and PC3 vs PC4 (bottom) of samples born in France without those removed in the first scan. Samples are coloured by the cluster inferred by DBSCAN (eps=230). In this case, 757 samples remained, of which 694 were in the largest cluster (0, orange). These samples were used in further analyses and considered as being of a ‘typical ancestral background’ for France.

In order to select Irish individuals (Republic of Ireland and Northern Ireland), step 1 was skipped but step 2 was run with relatively tight parameters, in both cases excluding approximately 20% of individuals.

In order to test the effectiveness of the pipeline in selecting individuals of a similar ancestral background, I calculated the variance in the genome-wide painting proportions for each country. Countries with high variance would indicate recent admixture.

To better understand how countries varied in their ancestry proportions, I ran Scikit-learn’s PCA (Pedregosa et al., 2011) on the average admixture proportions. I then ran a hierarchical clustering algorithm on the first 4 PCs (explained variance=0.244), and built a dendrogram (Figure 4) (Virtanen et al., 2020). For further analysis, I excluded countries in clusters dominated by African or East Asian ancestry, leaving 80 countries.

I used Sklearn’s StandardScaler utility class to standardise each feature (zero mean and unit variance) so that none would dominate the distribution, then ran a separate PCA using the standardised average admixture proportions for all countries and the 80 remaining countries (https://github.com/erdogant/distfit), and plotted a biplot (PC1 vs PC2 with loadings for each feature plotted), which shows the correlations between ancestries (Figure 5, Figure 6). This gives a more visual representation of how countries group by the predefined ancestries, and broadly reflects actual geography.

Results

This pipeline selected 24,511 individuals from 126 countries. These selected samples were painted using a reference/donor panel of ancient individuals (see Chapter One).

The countries that had high variance in ancestry proportions among individuals and therefore for which it was likely that the DBSCAN was not effective in choosing individuals of a similar ancestral background were Kazakhstan, Yemen, Egypt, and the Seychelles. Results for these countries should be interpreted with caution.

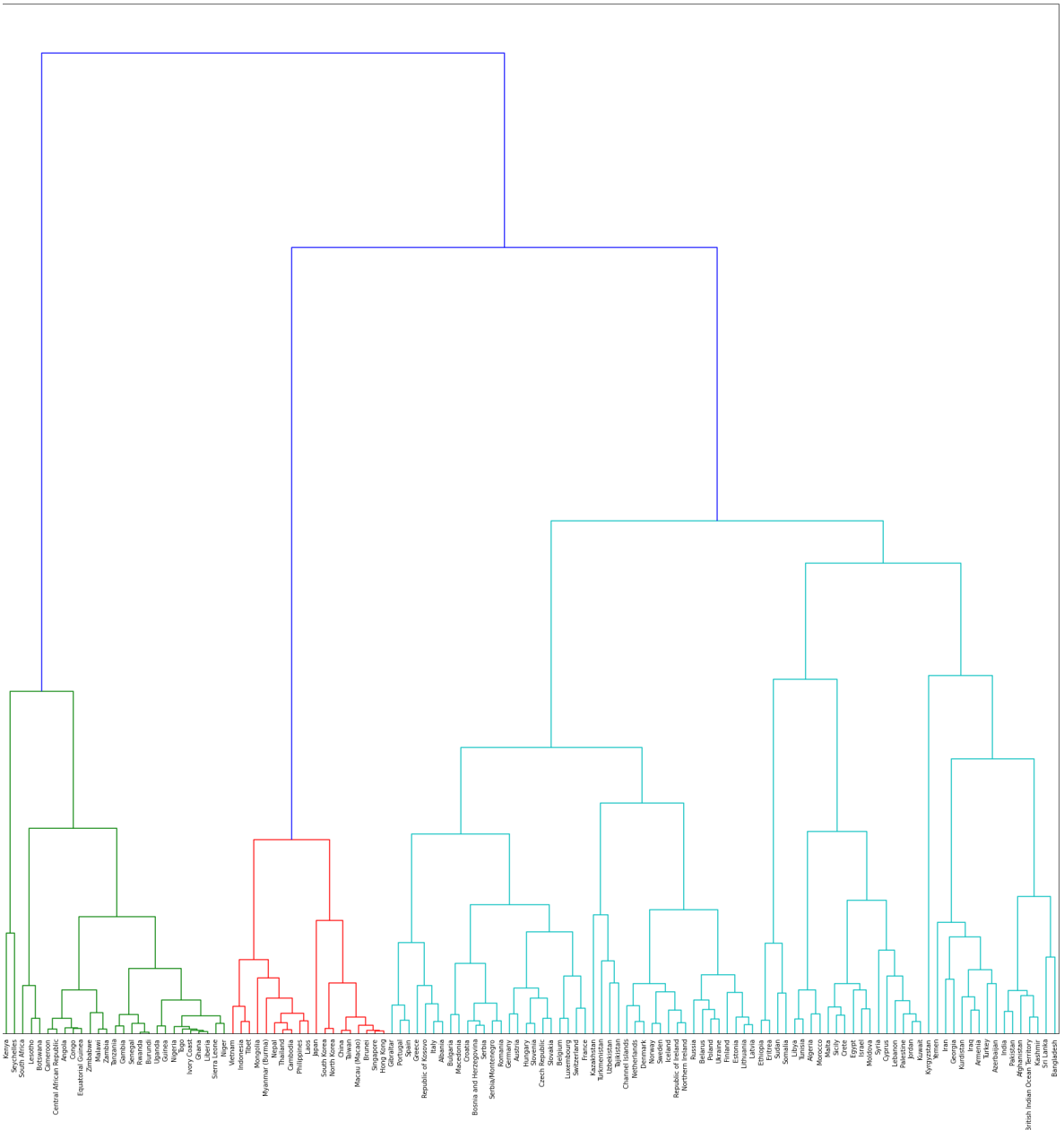


Figure 4 | Dendrogram based on hierarchical clustering of first 4 PCs of average admixture proportions per country.

The clustering algorithm was run on the first four PCs of NNLS admixture proportions per country, therefore clustering countries with similar admixture proportions. A dendrogram was then built based on the hierarchical clustering, with broad clusters shown as colours on the tree. For further analysis, countries in clusters dominated by African (green) or East Asian (red) ancestry were dropped.

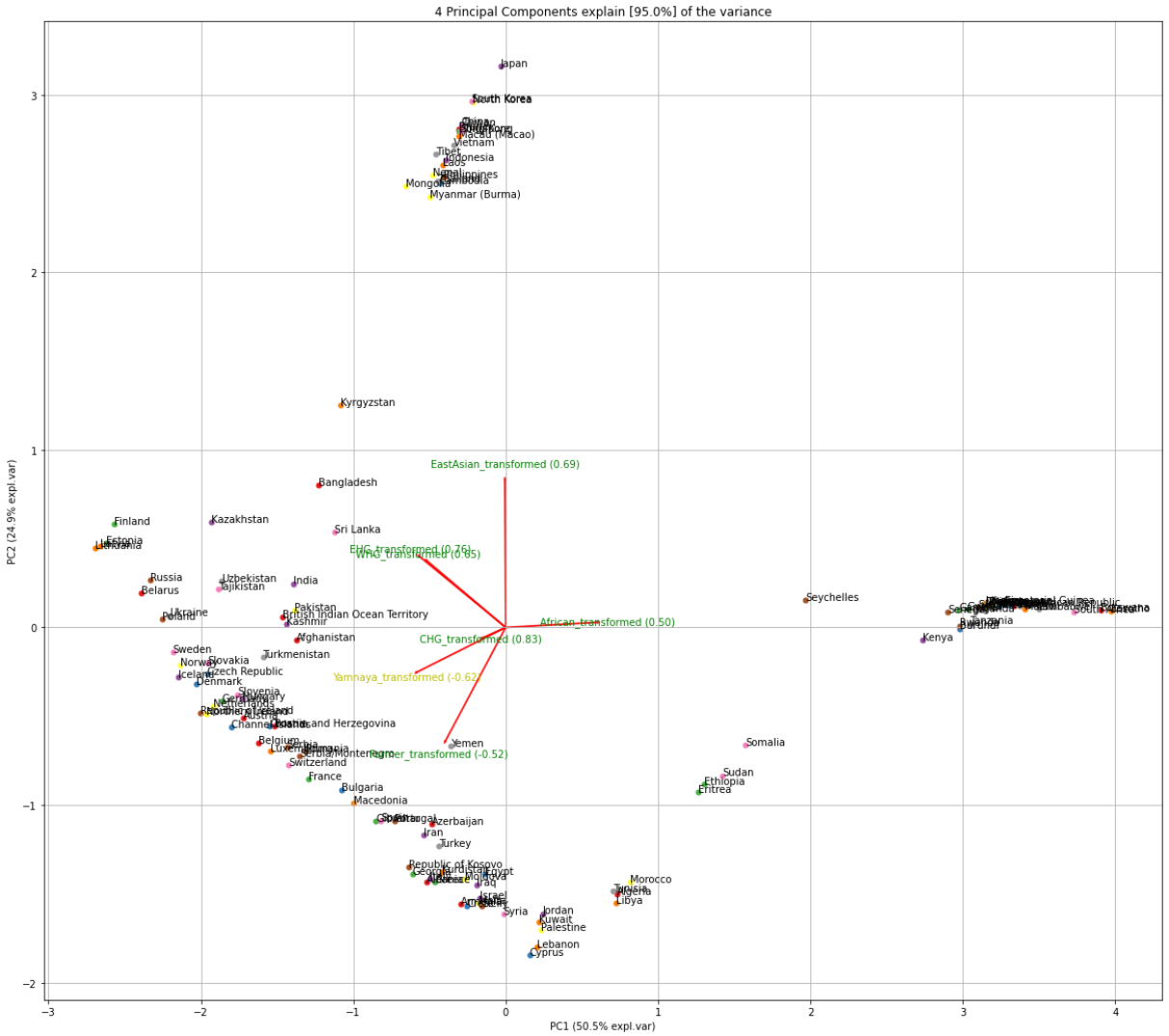


Figure 5 | PCA biplot (PC2 vs PC1) of standardised average NNLS admixture proportion per country, based on all countries.

When countries are separated by their NNLS admixture proportions, the African and East Asian components dominate, as in a regular PCA of genotypes with mixed ancestry samples. Samples are partially separated based on their other admixture proportions, particularly EHG/WHG vs Farmer (PC2).

Figure 6 | PCA biplot (PC1 vs PC2) of standardised average NNLS admixture proportion per country, based on 80 countries in Europe, West/Southern Asia, the Middle East and North Africa.

When countries are removed which are dominated by EastAsian or African ancestry (Figure 4), countries separate out by the other ancestries' admixture proportions. In this case, PC2 vs PC1 (i.e. a 90 degree rotation left) approximates geography.

Ancestry-PCs relationship

PCA is a dimensionality reduction technique that can be applied to genetic data, the results of which are useful as a means to visualise variation between individuals/groups, and are expected to reflect historical events that cause differences in ancestry due to drift, admixture etc. It is well established that PC1 vs PC2 vs PC3 generally separate African, European and East Asian populations. I ran multivariate linear regressions using ancestry components to predict UKB PCs (PCs from Bycroft et al., 2018, calculated on the entire UKB) (Table 2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where y is the NNLS ancestry proportion, β_0 is the y-intercept, x_1 is latitude, x_2 is longitude, β_1 and β_2 are the coefficients, and ε is the error.

Previous work has shown that the main UKB PCs that reflect British population structure are PCs 5 and 9, describing variation between English, Scottish and Welsh ancestry, and PCs 11 and 14 which further separate structure within Wales and England (Sarmanova, Morris and Lawson, 2020).

I found significant correlations between ancestry components and PC4 (R-squared=0.553) and PC5 (R-squared=0.344), as well as PC1 (R-squared=0.165) and PC7 (R-squared=0.130). As expected, PC1 separated individuals based mainly on their African ancestry component while PC2 separated individuals mainly on their East Asian component. I found that the high PC4 correlation with ancestry component was largely driven by a Steppe (Yamnaya/EHG) vs Farmer divide, both within Britain and internationally: high PC4 values are associated with high Steppe/low Farmer ancestry, while low PC4 values are associated with low Steppe/high Farmer ancestry.

To visualise this, I used Scikit-learn (0.21.2)'s (Pedregosa et al., 2011) k-means clustering on the first 20 PCs to group individuals into a predefined number of clusters, which neatly divides the population into approximate geographic groupings, as in Galinsky et al. (2016) (Figure 7, Figure 8). When PC4 is plotted against PC5, PC4 does not separate the clusters well (Figure 8), but when samples are coloured according to their Yamnaya or Farmer component, PC4 separates according to this gradient (Figure 9, Figure 10).

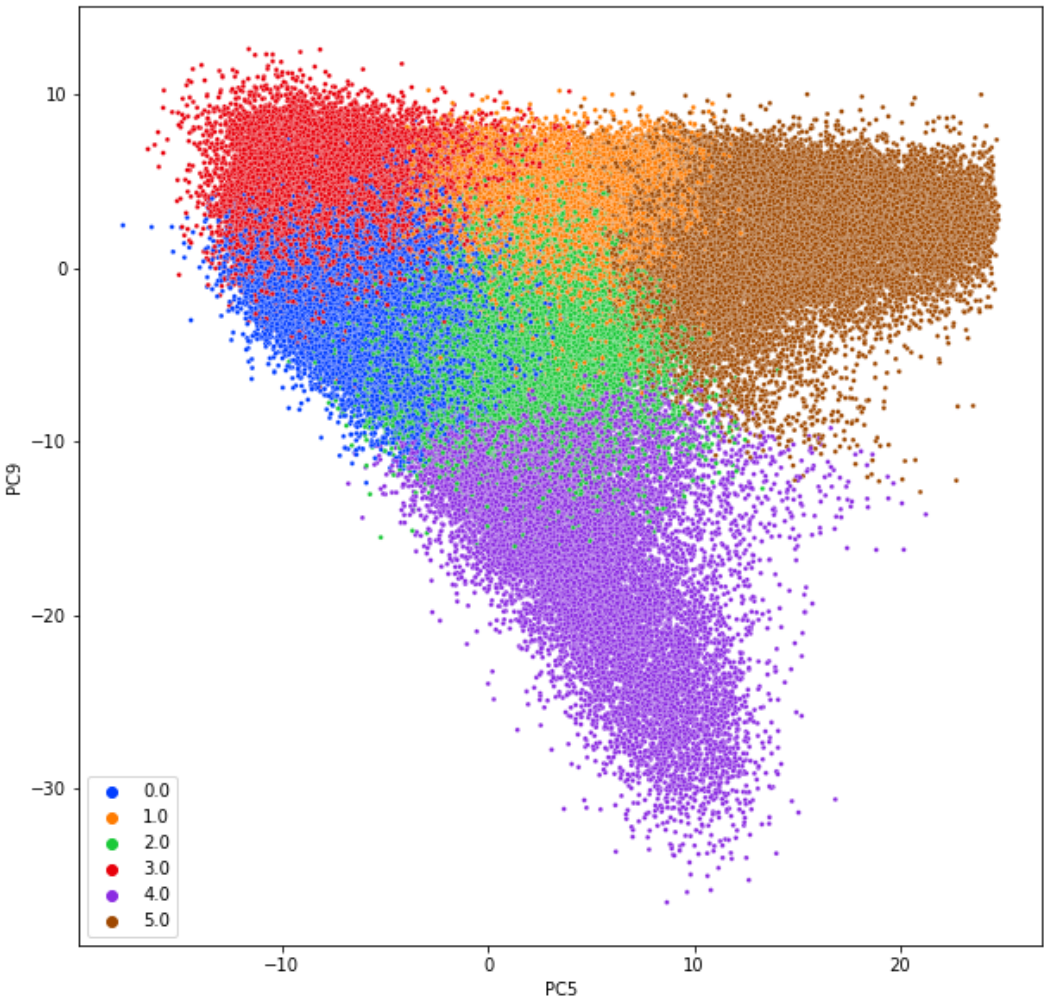


Figure 7 | PC5 vs PC9 of all 'white British' individuals in the UKB coloured by k-means clustering based on the first 20 PCs.

PCs are taken from Bycroft et al. (2018), i.e. run on all samples in the UKB, not just white british. K-means clustering was run on all 'white British' samples based on the first 20PCs, with n_clusters=6. Samples are plotted by PC5 vs PC9, which are the highest PCs which separate British population structure, as noted in Sarmanova, Morris and Lawson (2020). The k-means clusters approximately correspond to geographic regions (Galinsky et al., 2016) as follows:

Blue (cluster 0) = Southern England

Orange (cluster 1) = Scotland

Green (cluster 2) = North Wales

Red (cluster 3) = Northern England

Purple (cluster 4) = South Wales (Pembrokeshire)

Brown (cluster 5) = Northern Ireland

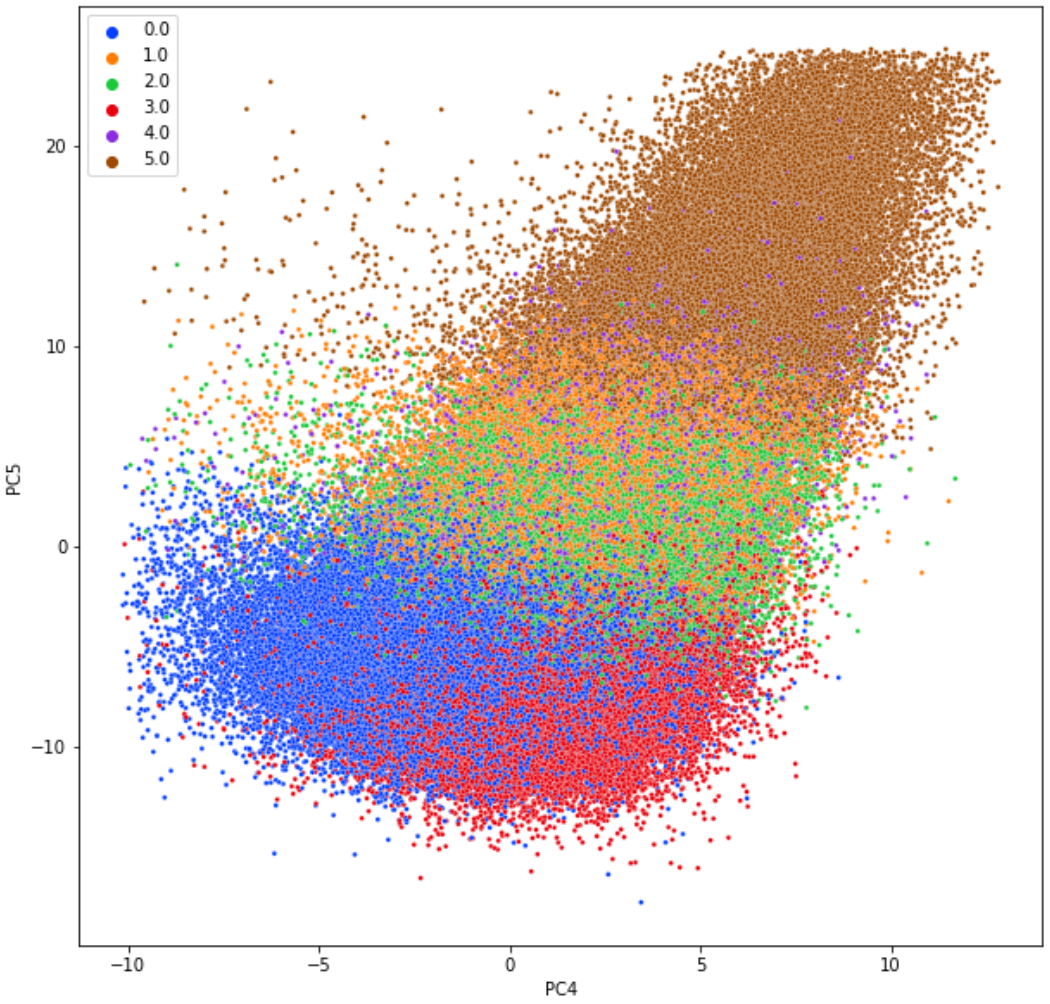


Figure 8 | PC4 vs PC5 of all 'white British' individuals in the UKB coloured by k-means clustering based on the first 20 PCs.

PCs are taken from Bycroft et al. (2018), i.e. run on all samples in the UKB, not just white british. K-means clustering was run on all 'white British' samples based on the first 20PCs, with n_clusters=6. Samples are plotted by PC4 vs PC5. PC4 does not separate the geographic clusters well, despite showing significant correlation with ancestry components. The k-means clusters approximately correspond to geographic regions (Galinsky et al., 2016) as follows:

Blue (cluster 0) = Scotland

Orange (cluster 1) = South Wales (Pembrokeshire)

Green (cluster 2) = North Wales

Red (cluster 3) = Southern England

Purple (cluster 4) = Northern Ireland

Brown (cluster 5) = Northern England

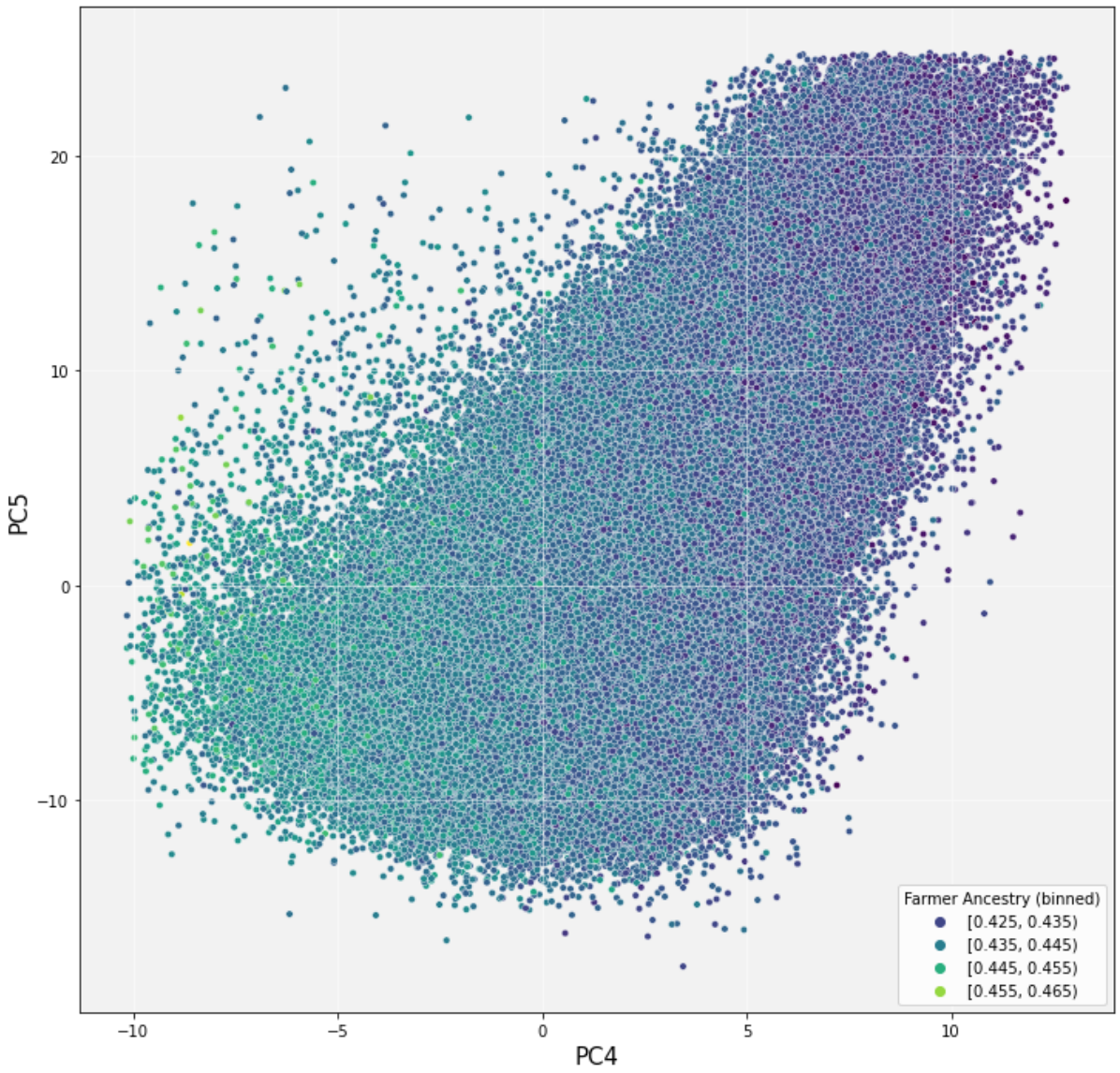


Figure 9 | PC4 vs PC5 of all white British samples, coloured by their Farmer ancestry component.

PC4 separates samples based on their Farmer vs Steppe ancestry component. Shown here is the Farmer component, binned into 20 bins. A lighter colour corresponds to a higher Farmer component.

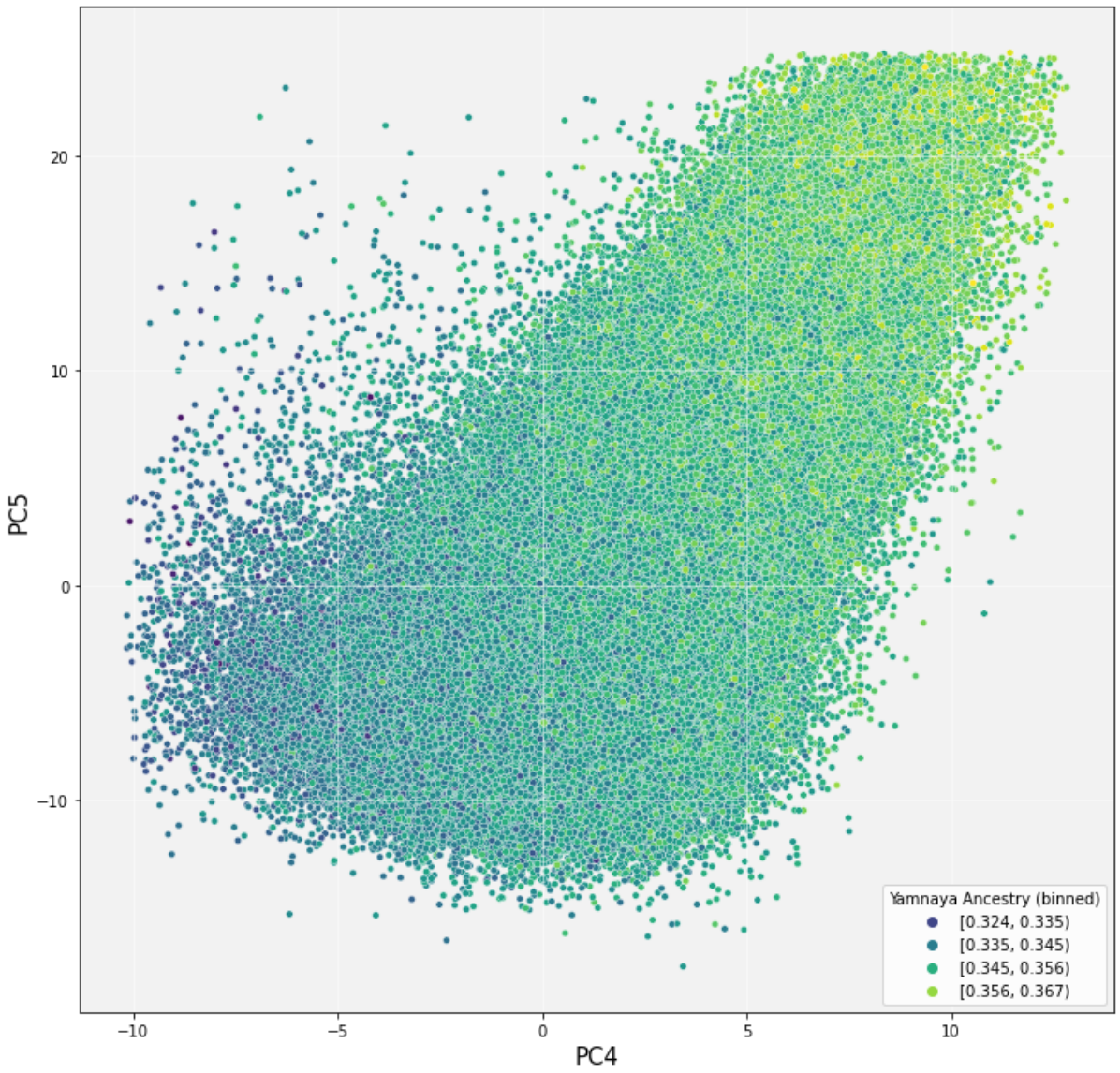


Figure 10 | PC4 vs PC5 of all white British samples, coloured by their Yamnaya ancestry component.

PC4 separates samples based on their Farmer vs Steppe ancestry component. Shown here is the Yamnaya component, binned into 20 bins. A lighter colour corresponds to a higher Yamnaya component.

Ancestry-geographic variation

Looking at a continent-wide level, the hunter-gatherer ancestries display distinct structure in modern populations (Figure 11). WHG-related ancestry is highest in present-day individuals from the Baltic States, Belarus, Poland and Russia; EHG-related ancestry is highest in Mongolia, Finland, Estonia and Central Asia; and CHG-related ancestry is maximised in countries east of the Caucasus, in Pakistan, India, Afghanistan and Iran, in accordance with previous results (Sikora et al., 2019). The CHG-related ancestry likely picks up both Caucasus hunter-gatherer and Iranian Neolithic signals, explaining the relatively high levels

in south Asia (Shinde et al., 2019). Consistent with expectations (Hofmanová et al., 2016; Feldman et al., 2019), Neolithic Anatolian-related farmer ancestry is concentrated around the Mediterranean basin, with high levels in southern Europe, the Near East and North Africa, including the Horn of Africa but less in Northern Europe. A contrasting pattern was observed in Yamnaya-related ancestry, decreasing from high levels in northern Europe and peaking in Ireland, Iceland, Norway and Sweden, but decreasing further south where Neolithic farmer ancestry still dominates. There is also evidence for its spread into southern Asia. These results provide a new level of detail on the modern distribution of ancient ancestries.

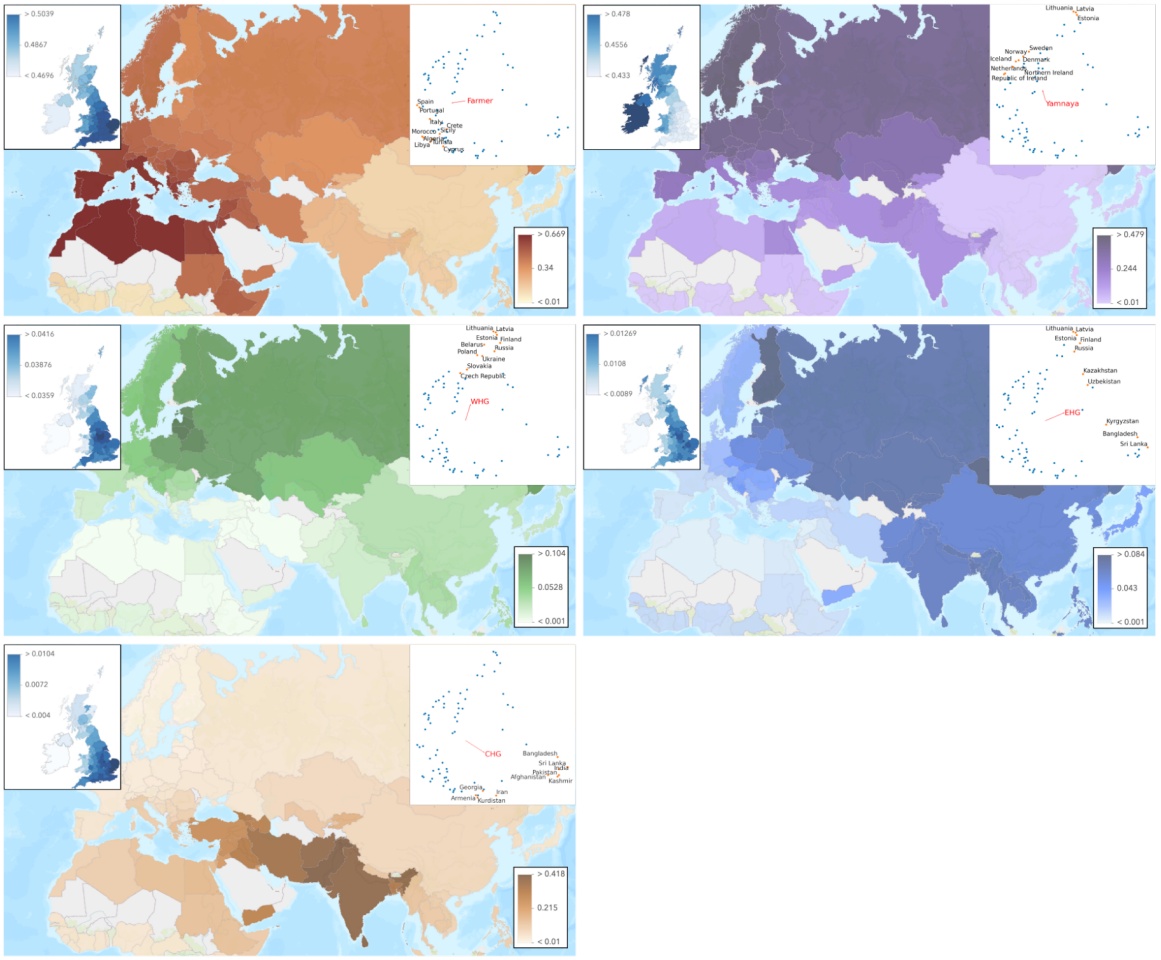


Figure 11 | The genetic legacy of Stone Age ancestry in modern populations.
From top left clockwise: Neolithic Farmer, Yamnaya, Caucasus hunter-gatherer, Eastern hunter-gatherer, Western hunter-gatherer. Panels show average admixture proportion in modern individuals per country estimated using NNLS (large maps), average per county within the UK (top left insert), and PCA (PC2 vs PC1) of admixture proportions, with the top 10 highest countries by admixture fraction labelled and PCA loadings for that ancestry.

Discussion

The UKB represents an important source of data for white British people but also for people from other countries globally. Usually, researchers restrict themselves to the white British cohort, but here I developed a method to select individuals from other countries which can be considered to be of a typical ancestral background for that country. This transforms the UKB from a resource that is informative about British ancestry to one that can be used to make inferences about populations worldwide. This method could be applied to any large genetic resource where individuals are from diverse countries but were not selected based on e.g. the birthplace of their grandparents, and so cannot be assumed to represent "ancestrally typical" individuals from a given country.

Because the ancestries used in this work were selected as sources to encapsulate western Eurasian ancestry, I urge caution in interpreting results for countries outside of this region. This is reflected in higher variances in the ancestry proportions found for individuals born in these countries.

The finding that these ancient ancestries have complex and heterogeneous modern distributions is perhaps unsurprising, given the demographic histories of the populations which distributed them. While these results are interesting from an historical and archaeological perspective, they also have profound implications for the distributions of genetic risk. As is shown in later chapters, differing allele frequencies in ancestral populations, driven by past selection and drift, has left a diverse landscape of genetic risk for diseases and phenotypes across the world. This may explain differences in disease load between populations which can be directly linked to the ancient populations which eventually formed them, informing risk prediction, our basic understanding of disease aetiology, and even treatment in some cases. Decomposing the genetic risk contributed by each ancestry to a modern population is the topic of the remainder of this thesis.

Tables

Country	Number of individuals	Number in wb cluster	eps value	Final number selected
Kenya	1684	277	800	110
Netherlands	491	300	230	153
Switzerland	175	19	230	143
India	4012	358	230	3107
Belgium	158	75	230	70
Singapore	502	262	230	86
Palestine	60	13	700	28
Nigeria	1159	90	800	1016
Hungary	105	0	230	79
Czech Republic	126	2	230	107
Ghana	929	35	700	848
Sri Lanka	744	54	230	620
Egypt	313	82	600	223
Japan	266	12	230	242
Hong Kong	648	107	230	448
Germany	2136	1044	230	1045
Turkey	182	5	400	160
Iran	540	16	230	469
South Africa	1364	488	700	57
Angola	56	2	700	22
Cameroon	54	5	230	44
Pakistan	1439	47	230	1332
Zimbabwe	750	252	700	254
Channel Islands	121	94	230	24
Bangladesh	246	4	230	225

Tanzania	425	84	1000	25
Sierra Leone	230	5	800	199
Portugal	320	18	230	275
Uganda	616	73	700	126
Poland	637	2	230	619
China	413	26	230	371
Cyprus	328	114	230	160
Italy	821	16	230	789
Bulgaria	71	2	230	65
Israel	87	10	300	56
France	856	103	230	690
Malta	365	170	230	135
Myanmar (Burma)	124	23	400	31
Philippines	333	8	230	310
Iraq	337	17	400	298
Finland	158	2	230	154
Libya	110	42	800	35
Norway	134	19	230	104
Russia	159	0	300	124
Nepal	161	0	230	119
Denmark	231	137	230	86
Spain	355	2	230	339
Serbia/Montenegro	56	0	230	55
Algeria	92	1	600	68
Sicily	3	0	230	2
Afghanistan	112	1	500	103
Gibraltar	77	34	230	32
Lebanon	77	13	500	50
Sudan	117	21	800	61

Morocco	93	3	800	66
Greece	131	5	230	116
Austria	196	20	230	171
Ukraine	62	1	400	54
Congo	167	11	800	146
Lithuania	72	0	230	66
Vietnam	74	0	230	69
Romania	68	0	230	61
Malawi	111	25	800	10
Gambia	42	0	800	38
Equatorial Guinea	4	0	800	2
Thailand	104	6	300	87
Indonesia	62	7	400	35
Central African Republic	42	6	800	14
Sweden	216	9	230	196
Jordan	15	1	600	9
Croatia	46	0	230	45
Ethiopia	81	8	800	57
Somalia	119	2	800	78
Zambia	246	101	800	56
Tunisia	27	2	900	14
Rwanda	25	1	1000	19
Yemen	108	38	2000	56
Burundi	26	3	1000	19
Eritrea	54	4	800	44
Syria	31	0	800	27
Luxembourg	7	0	230	6
Cambodia	8	1	1500	7
Macau (Macao)	8	0	500	6

Seychelles	46	1	2000	23
Liberia	22	3	800	15
Kuwait	31	16	800	5
Taiwan	26	1	230	23
Niger	5	0	800	2
Georgia	4	0	400	3
Iceland	19	4	230	15
Macedonia	18	0	230	17
Ivory Coast	32	0	800	29
Mongolia	6	0	400	6
Kazakhstan	13	0	4000	13
Brunei	19	4	400	8
Latvia	53	0	300	52
Bosnia and Herzegovina	41	0	230	41
Guinea	11	0	800	6
Slovenia	11	0	230	11
Azerbaijan	6	0	500	5
Slovakia	35	0	230	30
Kyrgyzstan	4	0	2000	3
Estonia	15	0	230	14
Senegal	12	0	800	7
South Korea	26	0	230	24
Togo	10	0	230	9
Armenia	5	0	600	5
Albania	12	0	230	12
British Indian Ocean Territory	10	2	500	4
Kurdistan	2	0	600	2
North Korea	6	0	230	5

Laos	3	0	500	3
Lesotho	2	0	400	2
Serbia	2	0	230	2
Republic of Kosovo	6	0	230	6
Botswana	7	1	800	4
Uzbekistan	3	0	5000	3
Kashmir	3	0	300	3
Turkmenistan	1	0	230	1
Belarus	6	0	230	4
Tibet	1	0	230	1
Crete	1	0	230	1
Moldova	1	0	230	1
Tajikistan	1	0	230	1

Table 1 | Parameters for selection of individuals in the UKB born in a given country of a 'typical ancestral background'.

This shows the initial number of individuals coded as being born in a country; the number removed because they clustered with the white British; the eps value for selecting the main cluster; and the number of individuals in the final selected cluster. Countries with fewer than 3 individuals in the final cluster, or no obvious main cluster, were discarded. The eps value is dependent on the genetic diversity of the population being selected, and was chosen manually and visually checked. In most but not all cases the largest cluster was chosen.

Principal Component	R-squared	Comments based on Sarmanova, Morris and Lawson (2020)
PC1	0.165	African vs East Asians vs Eurasians
PC2	0.057	African vs East Asians vs Eurasians
PC3	0.091	African vs East Asians vs Eurasians
PC4	**0.553**	
PC5	**0.344**	**Scottish vs English**
PC6	0.018	

PC7	**0.13**	
PC8	0.016	
PC9	**0.008**	**South Wales ancestry**
PC10	0.077	
PC11	**0.028**	**Northern England ancestry**
PC12	0.027	
PC13	0.01	
PC14	**0.026**	**Scottish ancestry**
PC15	0.008	
PC16	0.001	
PC17	0	
PC18	0.003	
PC19	0	
PC20	0.001	

Table 2 | R-squared values for multivariate linear regression, using ancestry components to predict PC values.

PCs with the highest R-squared values are highlighted in bold, with annotations based on Sarmanova, Morris and Lawson (2020) also shown.

Chapter Three: Ancestral contributions to complex phenotypes

Preface

Much of the contents of this chapter was previously published as (§ denotes joint first authors, @ denotes joint last authors):

The Selection Landscape and Genetic Legacy of Ancient Eurasians
Evan K. Irving-Pease§, Alba Refoyo-Martínez§, Andrés Ingason§, Alice Pearson§, Anders Fischer§, William Barrie§, Karl-Göran Sjögren, Alma S. Halgren, Ruairidh Macleod, Fabrice Demeter, Rasmus A. Henriksen, Tharsika Vimala, Hugh McColl, Andrew Vaughn, Aaron J. Stern, Leo Speidel, Gabriele Scorrano, Abigail Ramsøe, Andrew J. Schork, Anders Rosengren, Lei Zhao, Kristian Kristiansen, Peter H. Sudmant@, Daniel J. Lawson@, Richard Durbin@, Thorfinn Korneliussen@, Thomas Werge@, Morten E. Allentoft@, Martin Sikora@, Rasmus Nielsen@, Fernando Racimo@, Eske Willerslev@
bioRxiv 2022.09.22.509027; doi: https://doi.org/10.1101/2022.09.22.509027
In review at Nature, January 2023.
It has been modified to fit the style of a dissertation.

I performed all analyses described here. I wrote all the text except for the first three paragraphs of the Introduction, and the second and seventh paragraphs of the Discussion, which were written by Evan Irving-Pease. In addition to this description, I have noted in the legends of figures and tables if they were contributed by others.

Chapter summary

In this chapter I develop new methods to estimate ancestral contributions to complex, polygenic phenotypes based on the painting of the UK Biobank (Chapter One), gaining statistical power by bootstrapping over individuals and loci. I apply this to phenotypes known to be over-dispersed among ancient populations based on polygenic risk scores, re-capitulating previous results and presenting new findings. This has implications for genetic risk distribution in modern populations in light of the ancestry differences presented in Chapter Two.

Introduction

One of the central goals of human evolutionary genetics is to understand how natural selection has shaped the genomes of present-day people in response to changes in culture and environment. The transition from hunting and gathering, to farming, and subsequently pastoralism, during the Holocene in Eurasia, involved some of the most dramatic changes in diet, health and social organisation experienced during recent human evolution. The dietary changes, and the expansions into new climate zones, represent major shifts in environmental exposure, impacting the evolutionary forces acting on the human gene pool. These changes imposed a series of large-scale heterogeneous selection pressures on humans, beginning around 12,000 years ago and extending to the present-day. As human lifestyles changed, close contact with domestic animals and higher population densities are likely to have increased exposure to, and transmission of, infectious diseases; introducing new challenges to our survival (Page et al., 2016, Marciniak et al., 2022).

Our understanding of the genetic architecture of complex traits in humans has been substantially advanced by genome-wide association studies (GWAS) of present-day populations, which have identified large numbers of genetic variants associated with phenotypes of interest (MacArthur et al., 2017; Visscher et al., 2017; Bycroft et al., 2018). However, the extent to which these variants have been under directional selection during recent human evolution remains unclear, and the highly polygenic nature of most complex traits makes identifying selection difficult. While signatures of selection can be identified from patterns of genetic diversity in extant populations (Nielsen, 2005; Vitti, Grossman and Sabeti, 2013), this can be challenging in species such as humans, which show very wide geographic distributions and have thus been exposed to highly diverse and dynamic local environments through time and space. In the complex mosaic of ancestries that constitute a modern human genome, any putative signatures of selection may therefore misrepresent the timing and magnitude of the selective process. Similarly, episodes of admixture between ancestral populations can result in present-day haplotypes which contain no evidence of selective processes occurring further back in time. Ancient DNA (aDNA) provides the potential to resolve these issues, by directly observing changes in trait associated allele frequencies over time.

Whilst numerous prior studies have used ancient DNA to infer patterns of selection in Eurasia during the Holocene (e.g., Wilde et al., 2014; Mathieson et al., 2015; Ju and Mathieson, 2021), many key questions remain unanswered. To what extent are present-day differences in human phenotypes due to natural selection or to differing proportions of

ancient ancestry? What are the genetic legacies of Stone Age hunter-gatherer groups in present-day complex traits? How has the complicated admixture history of Holocene Eurasia affected our ability to detect natural selection in genetic data? To investigate these questions, and the selective landscape of Eurasian prehistory, we conducted the largest ancient DNA study to date of human Stone Age skeletal material, generating a phased and imputed dataset of >1,600 ancient genomes (Allentoft et al., 2022).

Most studies that look at polygenic risk scores in ancient populations use genotypes of ancient individuals, combined with effect sizes from modern GWAS studies, to reconstruct risk scores for ancient individuals (Irving-Pease et al., 2021). This involves exporting effect sizes across space and time, which is known to dramatically reduce the accuracy of the estimates (Duncan et al., 2019). Additionally, these scores are usually impossible to verify (except with specific phenotypes such as height where calibration is possible (Cox et al., 2019, 2022), and don't necessarily measure what an ancient population contributed to phenotypic diversity in a modern population(s), especially when there has been selection or bottleneck events in between.

Here, I estimate the contribution from different ancestral populations (EHG, CHG, WHG, Yamnaya and Anatolian farmer) to variation in polygenic phenotypes in present-day individuals, leveraging the exceptional resolution offered by the UK Biobank genomes (Bycroft et al., 2018) to investigate this. I calculate ancestry-specific polygenic risk scores based on chromosome painting of the >400,000 UKB genomes, using ChromoPainter (Chapter One). This allowed me to identify if any of the ancient ancestry components were over-represented in modern UK populations at loci significantly associated with a given trait, and also avoids exporting risk scores over space and time.

This is a well-powered approach due to the large modern sample size, and is a more direct measure of the variants that a given ancestry contributed to the "white British" genetic landscape. Thus we can draw conclusions about the differing contributions of each ancestry to modern genetic risk, whether due to drift or selection. I use bootstrapping to test whether some ancestries are significantly and systematically over-represented for a phenotype, indicating selection. Additionally, I look at the ancestral haplotypic background of a high effect variant, ApoE4, which is implicated in Alzheimer's Disease (Corder et al., 1993; Strittmatter et al., 1993).

Methods

I used effect size estimates from the UK Biobank Neale lab GWAS (Bycroft et al., 2018), and used 1,703 non-overlapping and approximately independent linkage disequilibrium (LD) blocks (Berisa and Pickrell, 2016). For each block, I restricted the SNPs to those with a p-value less than the genome-wide significance threshold (5e-8), and from these chose the SNP with the lowest p-value. I then used these SNPs to calculate polygenic risk scores for each ancestry, using ancestry-specific 'effect allele frequencies' derived from the painting.

In order to calculate the effect allele painting frequency for a given ancestry f_{anc,i} for SNP *i* I used the formula:

$$f_{\{anc,i\}} = \frac{\sum_{j}^{M_{effect}} Painting\ certainty_{\{j,i,anc\}}}{\sum_{j}^{M_{alt}} Painting\ certainty_{\{j,i,anc\}} + \sum_{j}^{M_{effect}} Painting\ certainty_{\{j,i,anc\}}}$$

Where there are M_{effect} individuals homozygous for the effect allele, M_{alt} individuals homozygous for the other allele, and $\sum_{j}^{M_{efefct}} Painting\ certainty_{\{j,i,anc\}}$ is the sum of the painting probabilities for that ancestry *anc* in individuals homozygous for the effect allele at SNP i. This calculates an estimate of an ancestral contribution to effect allele frequency in a modern population. One benefit of this approach is that because it only matters how effect alleles are painted relative to alternate alleles for an ancestry group, differences in genome-wide painting averages between ancestries will not cause bias.

To calculate an ancestry-specific PRS I summed over all I pruned SNPs in an additive model:

$$ARS_{anc} = \sum_{i}^{I} f_{\{anc,i\}} * beta_{i}.$$

I then ran a transformation as in Berg & Coop (2014), centering results around the ancestral mean (i.e. all ancestries) and reporting as a Z-score. I derived standard deviations for each score by running a block bootstrap (1000 iterations) on (1) loci and (2) individuals. I calculated polygenic risk scores for 39 traits shown to be significantly over-dispersed across ancient populations beyond what would be expected under a null model of genetic drift (Irving-Pease et al., 2022 Supplementary Note S2c). For computational reasons, I used a

random batch of 48,000 painted individuals to calculate the effect allele frequencies, which is sufficiently large to approximate the frequencies even for ancestries that are painted less often.

The calculations were limited to the 549,323 SNPs used in the painting of the UKB (Chapter One). This is expected to reduce predictive power compared to using the full set of imputed SNPs in the UKB, but only slightly (Choi and O'Reilly, 2019). There was a ~15% decrease in the number of SNPs included per phenotype in the PRS calculation compared with the imputed data.

To test the ancestral background of a single variant, APOE4, I calculated the average painting score for each ancestry at all sites on the chromosome of haplotypes containing the effect allele. This makes it clear when there is an excess of a particular ancestry at the site of interest.

Results

My results tell us about the ancestral contribution to modern phenotypes in the white British population (Figure 1), and I stress I am not making claims about the phenotypes of ancient populations.

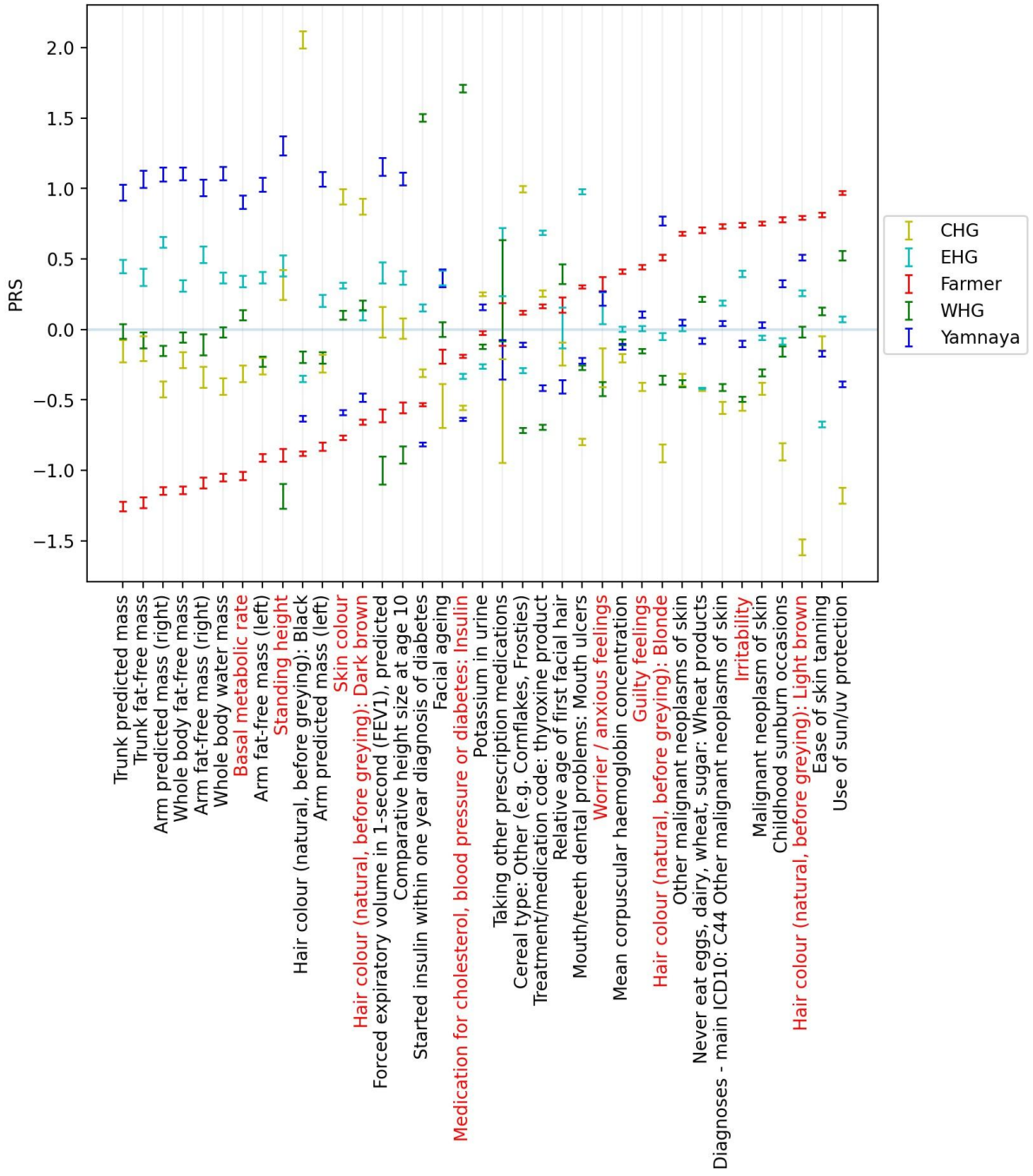


Figure 1 | Ancestry-specific polygenic scores based on chromosome painting of the UK Biobank, for traits significantly over-dispersed in ancient populations.
Confidence intervals are estimated by bootstrapping modern samples, and traits mentioned in the main text are highlighted in red. Confidence intervals were calculated by re-running PRS calculation

on random batches of 48,000 individuals, with replacement (1000 iterations), while keeping all other annotations intact. Here we show 2 x standard deviation error bars, expected to represent ~95% confidence interval under a normal distribution. Bootstrapping individuals tests the extent to which ancestry X contributed higher genetic risk for phenotype Y in a given population, either due to drift or selection.

I found that Yamnaya, CHG and EHG ancestral contributions (which together form a 'Steppe' component) have relatively high scores for height, whereas Farmers and WHG ancestral contributions have relatively low scores. This accords with most previous studies (Mathieson et al., 2015; Martiniano et al., 2017; Cox et al., 2019) but not all (Marnetto et al., 2022). EHG and Yamnaya both score highly for body mass and basal metabolic rate.

Hair and skin pigmentation show significant differences between the ancestral contributions, with risk scores for skin colour for the three hunter-gatherer ancestries being higher (i.e. darker pigmentation) than Farmer and Steppe (as in Ju and Mathieson (2021)). On the other hand, traits related to malignant neoplasms of skin show higher scores for the Farmer ancestral contribution; while Farmer and Yamnaya ancestral contributions have higher scores for blonde and light brown hair, with the hunter-gatherer ancestries showing higher scores for dark brown. CHG is the only ancestral contribution which stands out as having a high risk score for black hair.

Intriguingly, the WHG ancestral component has strikingly high scores for traits related to cholesterol, blood pressure and diabetes, both when bootstrapping individuals and loci (cf. Marnetto et al., 2022). In terms of psychiatric traits, the Farmer component scores highest for anxiety, guilty feelings, and irritability.

The two bootstrapping methods mean slightly different things. Individuals in the UKB are related through shared genealogies, and so by bootstrapping over non-independent individuals (Figure 1) I am testing the consistency of the signal within the population. From this bootstrapping exercise I can conclude whether a difference in allele frequencies in ancient populations contributed to phenotypic variation today. Unsurprisingly, with a large enough sample size most phenotypes show differences in ancestral contributions for this, usually due to drift or founder effects. However, this goes further than just reporting risk scores for ancient populations, because I am looking directly at coalescent tracts in the British population. I can conclude that "ancestry X contributes higher genetic risk for phenotype Y in the test population". On the other hand, because I have used independent LD blocks to select SNPs to include in the PRS calculation, the requirement for

independence is met when I bootstrap with loci (Figure 2). A positive result here is therefore much stronger, showing a systematic over/under-representation of an ancestry at loci affecting a given trait, beyond what is expected given the correlation among individuals. This points towards selection as an explanation.

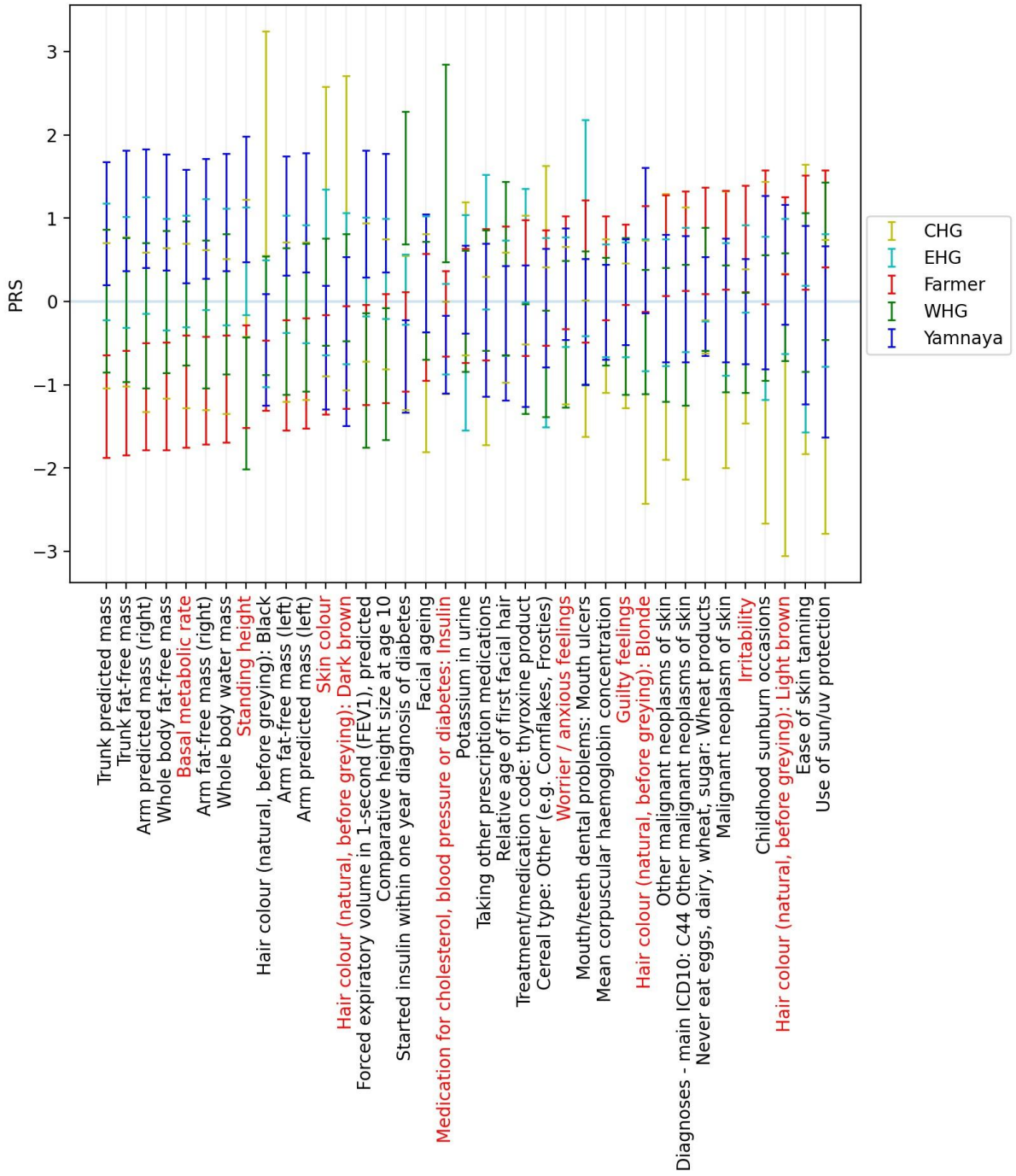


Figure 2 | Ancestry-specific polygenic risk scores with 95% confidence intervals derived from bootstrapping loci for phenotypes shown to be significantly over-dispersed between ancient populations.

Confidence intervals were calculated by bootstrapping independent loci from separate LD blocks (1000 iterations), while keeping all other annotations intact. Here we show 2 x standard deviation error bars, expected to represent ~95% confidence interval under a normal distribution. Bootstrapping loci tests whether there is a systematic bias towards an ancestry for a given phenotype across all significant SNPs, possibly indicating selection.

The effect/risk allele (rs429358, n=127,760) of ApoE4 is preferentially painted as WHG/EHG, with a clear depletion of other ancestries (especially Farmer) at this locus compared to the genome-wide average (Figure 3). This indicates that this allele was contributed at least in part by hunter-gatherer ancestry into modern (British) populations, above what we would expect by chance.

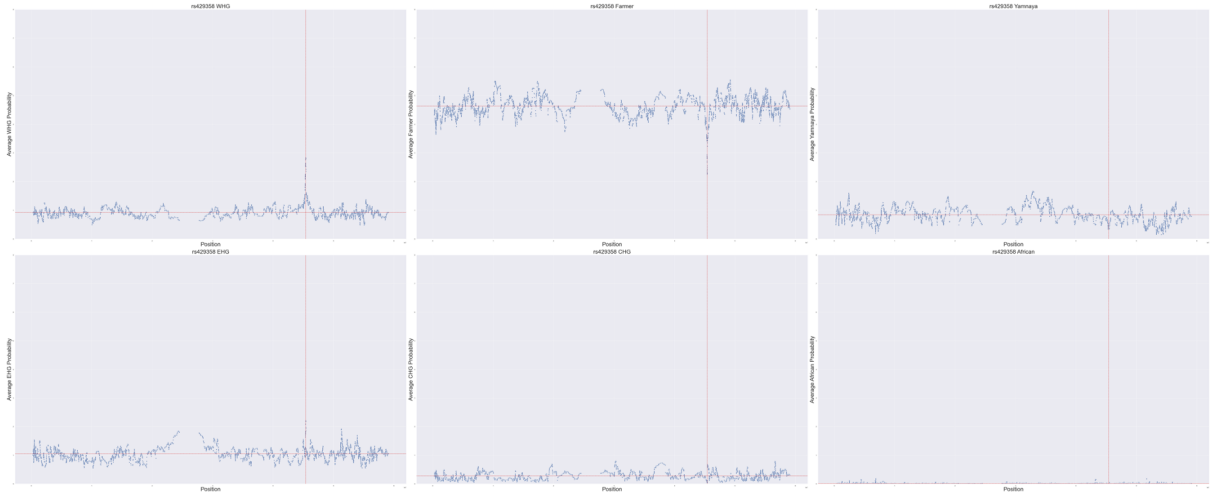


Figure 3 | Average painting score for each ancestry at all sites on chromosome 19 of haplotypes containing the effect allele for ApoE4 (rs429358, n=127,760).
Vertical red line indicates the position of the SNP of interest; horizontal red line indicates the average painting score for that ancestry for haplotypes containing the effect allele across the entirety of chromosome 19. There is a clear excess of WHG/EHG ancestry and a depletion of Farmer ancestry at this locus.

Discussion

The methods here directly link genetic contributions from pre-defined ancestries to complex phenotypes in modern people. For most traits, each ancestry contributed differently to the modern genetic landscape, with some conveying enhanced or reduced risk either due to drift (including population bottlenecks/founder events) or selection. Because gradients exist in these ancestries across Britain and further afield (Chapter One, Chapter Two), these differing risk scores indicate how geographically heterogeneous ancestry distributions may contribute to differing genetic risk profiles, in addition to other factors such as geography, socio-economic status etc.

It is important to emphasise that these analyses do not represent an explicit test for selection. This can be useful, as they provide a less sensitive test for differential contributions from ancestries, sacrificing specificity for sensitivity; furthermore, it is questionable whether tests for selection are really able to detect selection on a focal trait, as they claim. For a fuller discussion, see Discussion: Theoretical Implications.

Taken together with other results in Irving-Pease et al. (2022), these analyses help to address the famous discussion of selection in Europe relating to height (Mathieson et al., 2015; Cox et al., 2019; Rosenstock et al., 2019). The finding that Steppe individuals have consistently high genetic values for height (Irving-Pease et al., 2022 Supplementary Note 2c), is mirrored by the UK Biobank results, which find that the ‘Steppe’ ancestral components (Yamnaya/EHG) contributed to increased height in present-day populations. This shows that the height differences in Europe between north and south may not be due to selection in Europe, as claimed in many previous studies, but may be a consequence of differential ancestry.

A caveat for all studies involving polygenic risk calculation is that they rely on effect size estimates from an original GWAS which may be affected by population stratification in the GWAS panel, even when it has apparently been controlled for. This seems to be less of a problem in the UKB than in previous GWAS studies (Berg et al., 2019), but should be kept in mind. One benefit of my approach is that there is no requirement to export these risk scores across time and space: I am using effect sizes estimated from the modern population to calculate ancestral contributions to the same modern population.

ApoE4 is an isoform of the APOE gene, resulting from linkage disequilibrium between two SNPs, rs429358 and rs7412 (Rall, Weisgraber and Mahley, 1982), and associated with

increased risk for metabolic, vascular and neurodegenerative diseases in adulthood (de-Almada et al., 2012). It may provide some enhanced cognitive ability in children and young adults (Tuminello and Han, 2011) and other health and immunity benefits, particularly in highly infected environments (e.g. Oriá et al., 2007). There are several lines of evidence suggesting a link between the evolution of diet and the ApoE isoforms: ε2 and ε3 alleles are associated with lower levels of blood cholesterol (Carvalho-Wells et al., 2010; Petkeviciene et al., 2012), while ε4 is associated with higher levels, leading some to speculate that the derived ε3 allele is 'meat-adaptive' (Finch and Stanford, 2004; Allen, Bruss and Damasio, 2005). In a study of South Americans, there was a fivefold increase in the ApoE4 allele in hunter-gatherers versus horticulturalists (Reales et al., 2017), potentially because the immune benefits outweighed the advantages of low blood cholesterol (Trumble et al., 2017). Generally, ε4 prevalence is higher in indigenous foraging groups such as the Pygmies, Khoi San, Papuans and some Native Amercians, while ε3 is most frequent in populations with a long-established agricultural economy (Corbo and Scacchi, 1999). Finally, ApoE4 is implicated in higher blood vitamin D levels (Huebbe et al., 2011).

The ε4 variant has been shown to be ancestral in humans (Fullerton et al., 2000). There is a linear increasing trend in ε4 prevalence from South to North in Europe, with Sardinians showing the lowest prevalence (Corbo et al., 1995; Lucotte, Loirat and Hazout, 1997; Adler et al., 2017), while there is a more than two-fold increase in Nordic versus Mediterranean countries (Trumble and Finch, 2019). Sardinians are an unusual population, having the highest level of neolithic farmer ancestry of all modern European populations (Chiang et al., 2018). In this light, differences in genome-wide ancestry proportions between northern (high WHG/EHG, low Farmer) and southern Europe (high Farmer, low WHG/EHG) (Chapter Two) may explain at least part of the differences in frequency of the ε4 variant and subsequent AD genetic risk.

In light of the ancestry gradients within Britain and Eurasia (Allentoft et al., 2022), these results support the hypothesis that ancestry-mediated geographic variation in disease risks and phenotypes is commonplace. It points to a way forward for disentangling how ancestry contributed to differences in risk of genetic disease – including metabolic and mental health disorders – between present-day populations.

The transition from hunting and gathering, to farming, and subsequently pastoralism, precipitated far-reaching consequences for the diet, and physical and mental health of Eurasian populations. These dramatic cultural changes created a heterogeneous mix of selection pressures. The analyses in Irving-Pease et al. (2022) reveal that the ability to

detect signatures of natural selection in modern human genomes is drastically limited by conflicting selection pressures in different ancestral populations masking the signals. Developing methods to trace selection in individual ancestry components allowed us to effectively double the number of significant selection peaks, which helped clarify the trajectories of a number of traits related to diet and lifestyle. Furthermore, numerous complex traits thought to have been under local selection are better explained by differing proportions of ancient ancestry in present-day populations. Overall, these results emphasise how the interplay between ancient selection and major admixture events occurring across Europe and Asia in the Stone and Bronze Ages have profoundly shaped the patterns of genetic variation observed in present-day human populations.

Chapter Four: Genetic risk for Multiple Sclerosis originated in Pastoralist Steppe populations

Preface

The contents of this chapter was previously published as (§ denotes joint first authors, @ denotes joint last authors):

Genetic risk for Multiple Sclerosis originated in Pastoralist Steppe populations
William Barrie§, Yaoling Yang§, Kathrine E. Attfield§, Evan Irving-Pease§, Gabriele Scorrano§, Lise Torp Jensen§, Angelos P. Armen, Evangelos Antonios Dimopoulos, Aaron Stern, Alba Refoyo-Martinez, Abigail Ramsøe, Charleen Gaunitz, Fabrice Demeter, Marie Louise S. Jørkov, Stig Bermann Møller, Bente Springborg, Lutz Klassen, Inger Marie Hyldgård, Niels Wickmann, Lasse Vinner, Thorfinn Sand Korneliussen, Martin Sikora, Kristian Kristiansen, Santiago Rodriguez, Rasmus Nielsen, Astrid K. N. Iversen@, Daniel J. Lawson@, Lars Fugger@, Eske Willerslev@
bioRxiv 2022.09.23.509097; doi: https://doi.org/10.1101/2022.09.23.509097
In review at Nature, January 2023.
It has been modified to fit the style of a dissertation.

Full author contributions are listed below. I wrote the main text, and performed the population painting, generated the local ancestry from the population painting, and performed the ARS analysis. I generated the SNP associations along with Angelos P. Armen. The data generation and standard population genetic analyses were performed by Gabriele Scorrano; the anomaly score analysis was performed by Dan Lawson; the cluster, weighted average prevalence, local ancestry GWAS, LDA and HTRX were performed by Yaoling Yang; the polygenic selection test was performed by Evan Irving-Pease and Evangelos Dimopoulos.

Full Supplementary Material can be found online at https://doi.org/10.1101/2022.09.23.509097.

Author Contributions

W.B., Y.Y., K.E.A., E.K.I-P, G.S., and L.T.J. contributed equally to this work.
A.I., D.J.L., L.F., and E.W. led the study.
W.B., A.R-M., L.F., R.N., and E.W. conceptualised the study.
R.N., K.K., L.F., and E.W. acquired funding for research.
A.R., C.G., F.D., M.L.S.J., S.B.M., B.S., L.K., I.M.H., N.W., L.V., and T.S.K., were involved in sample collection and processing
W.B., Y.Y., E.K.I-P, A.S., S.R., and D.J.L. were involved in developing and applying methodology.
W.B., Y.Y., E.K.I-P, G.S., A.A., A.R., E.D., M.S., S.R., A.I., and D.J.L. undertook formal analyses of data.
W.B., Y.Y., K.E.A., E.K.I-P, and L.T.J., A.I., L.F., and E.W. drafted the main text (W.B. led this).
W.B., Y.Y., E.K.I-P, G.S., L.T.J., E.D., A.S., F.D., M.L.S.J., S.B.M., B.S., L.K., I.M.H., N.W., L.V., A.I., and D.J.L. drafted supplementary notes and materials.
W.B., Y.Y., K.E.A., E.K.I-P, L.T.J., A.A., K.K., R.N., A.I., D.J.L., L.F., and E.W. were involved in reviewing drafts and editing.
All co-authors read, commented on, and agreed upon the submitted manuscript.

Chapter summary

Multiple sclerosis (MS) is a modern neuro-inflammatory and -degenerative disease, which is most prevalent in Northern Europe. Whilst it is known that inherited risk to MS is located within or within close proximity to immune genes it is unknown when, where and how this genetic risk originated. By using the largest ancient genome dataset from the Stone Age, along with new Medieval and post-Medieval genomes, we show that many of the genetic risk variants for MS rose to higher frequency among pastoralists located on the Pontic Steppe, and were brought into Europe by the Yamnaya-related migration approximately 5,000 years ago. We further show that these MS-associated immunogenetic variants underwent positive selection both within the Steppe population, and later in Europe, likely driven by pathogenic challenges coinciding with dietary and lifestyle environmental changes. This study highlights the critical importance of this period as a determinant of modern immune responses and its subsequent impact on the risk of developing MS in a changing environment.

Introduction

Multiple sclerosis (MS) is an autoimmune disease of the brain and spinal cord that currently affects more than 2.5 million people worldwide. The prevalence varies markedly with ethnicity and geographical location, with the highest prevalence observed in Europe (142.81 per 100.000 people), and Northern Europeans being particularly susceptible to developing the disease (Walton et al., 2020). The origins and reasons for the geographical variation are poorly understood, yet such biases may hold important clues as to why the prevalence of autoimmune diseases, including MS, has continued to rise during the last 50 years.

While still elusive, MS aetiology is thought to involve gene-gene and gene-environmental interactions. Accumulating evidence suggests that exogenous triggers initiate a cascade of events involving a multitude of cells and immune pathways in genetically vulnerable individuals, which may ultimately lead to MS neuropathology (Attfield et al., 2022).

Genome-wide association studies have identified 233 commonly occurring genetic variants that are associated with MS; 32 variants are located in the HLA region and 201 outside the HLA region (IMSGC, 2019). The strongest MS associations are found in the HLA region with the most prominent of these, HLA-DRB1*15:01, conferring an approximately three-fold increase in the risk of MS. Collectively, genetic factors are estimated to explain approximately 30% of the overall disease risk, while environmental and lifestyle factors are considered the major contributors to MS. Such determinants may include geographically varying exposures like infections and low sun exposure/vitamin D deficiency. For instance, while infection with Epstein-Barr virus frequently occurs in childhood and usually is symptomless, delayed infection into early adulthood, as typically observed in countries with high standards of hygiene, is associated with a 32-fold increased risk of MS (Bjornevik et al., 2022, Lanz et al., 2022). Lifestyle factors associated with increased MS risk such as smoking, obesity during adolescence, and nutrition/gut health also vary geographically (Olsson et al., 2017). Dysregulations including autoimmunity in modern immune systems could also result from the absence of ancient immunological triggers to which humans have evolutionarily adapted, for instance by disturbing the delicate balance of pro- and anti-inflammatory pathways (Benton et al., 2021).

European ancestry has been postulated to explain part of the global difference in MS prevalence globally in admixed populations (Chi et al. 2019). Specifically, cases in African Americans exhibit increased European ancestry in the HLA region compared to controls, with European haplotypes conferring more MS risk for most HLA alleles, including

HLA-DRB1*15:01. Conversely, Asian American cases have decreased European ancestry in the HLA region compared to controls. Although Ancient European ancestry and MS risk in Europe are known to be geographically structured (Figure 1a-b), the effect of ancestry variation within Europe on MS prevalence is unknown.

Modern ancestry is viewed as a mixture of genetic ancestries derived from ancient populations, who can be distinguished by their subsistence lifestyle: Western Hunter-Gatherers (WHG), Eastern Hunter-Gatherers (EHG), Caucasus Hunter-Gatherers (CHG), Anatolian Farmers, and Steppe Pastoralists (Figure 1c-d). Using the largest ancient genome dataset from the Stone Age, presented in the accompanying study 'Population Genomics of Stone Age Eurasia' (Allentoft et al., 2022), coupled with new Medieval and post-Medieval genomes, we quantified modern European ancestry with respect to these ancient ancestries to identify signals of lifestyle-specific evolution. Then we determined whether the variants associated with an increased risk for MS have undergone positive selection. We asked when selection occurred and whether the targets of selection were specific to diet and lifestyle. Finally, we examined the environmental conditions that may have caused selection for risk variants, including human subsistence practice and exposure to pathogens. An overview of the evidence provided by all methods used can be found in Supplementary Figure 9.1.

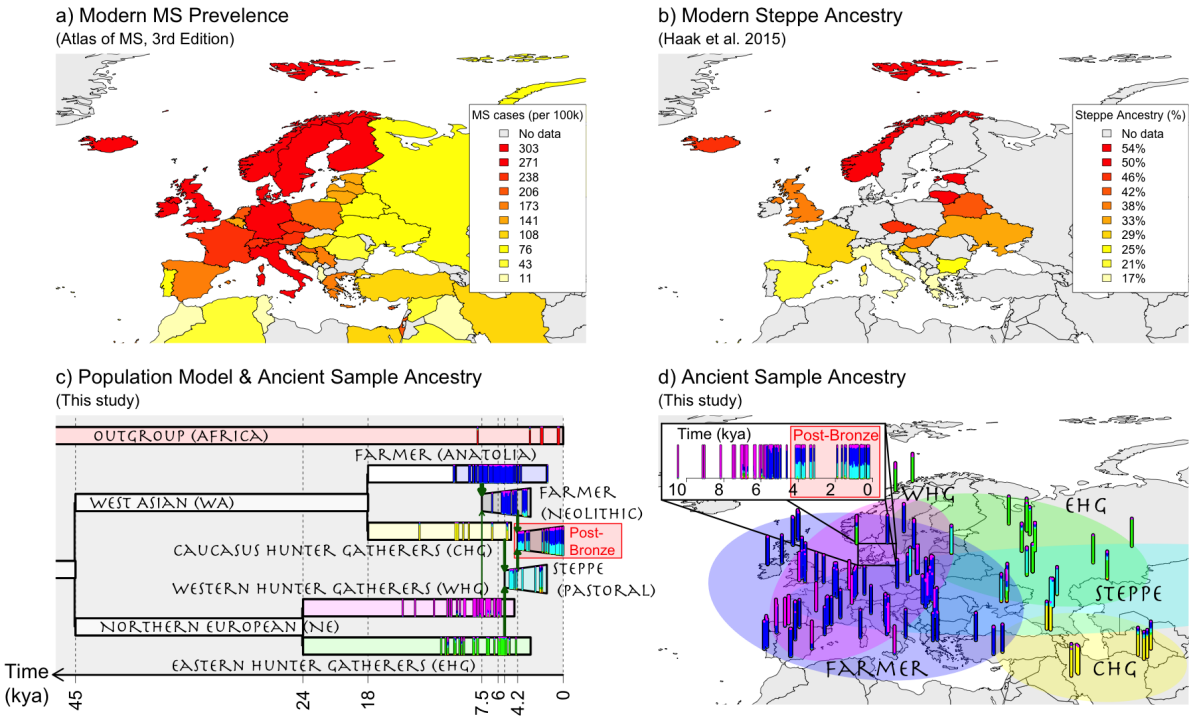


Figure 1 | Population history of Europe is associated with the modern-day distribution of MS.

a) Modern-day geographical distribution of MS in Europe. Prevalence data for MS (cases per 100,000) was obtained from the Atlas of MS - 3rd edition. b) Steppe ancestry in modern samples as estimated by Haak et al., 2015. c-d) A model of European prehistory (Jones et al., 2015) onto which our reference samples have been projected using NNLS (see Methods), and the same data represented spatially. Chronologically, Western Hunter-Gatherers (WHG) and Eastern Hunter-Gatherers (EHG) were largely replaced by Anatolian Farmers amid demographic changes during the “Neolithic transition” around 9,000 years ago. Later migrations during the Bronze Age about 5,000 years ago brought a roughly equal Steppe ancestry component from the Pontic-Caspian Steppe to Europe, an ancestry descended from the EHG from the Middle Don River region and Caucasus Hunter-Gatherers (CHG) (Allentoft et al., 2022). Steppe ancestry has been associated with the Yamnaya culture and then with the expansion westwards of the Corded Ware Complex and Bell Beaker culture, and the eastwards expansion in the form of the Afanasievo culture (Allentoft et al., 2015; Haak et al., 2015). Samples are vertical bars representing their “admixture estimate” estimated by NNLS (methods) from six ancestries: EHG (green), WHG (pink), CHG (yellow), Farmer (blue), Steppe (cyan) or an Outgroup (represented by ancient Africans, red). Important population expansions are shown as growing bars and “recent” (post-Bronze age) non-reference admixed populations are shown for the Denmark time-transect (see Supplementary Figure 1.1 for details).

Results

We obtained local ancestry (i.e. ancestry at specific loci) labels for ~410,000 self-identified “white British” individuals in the UK Biobank (Bycroft et al., 2018), using a reference panel of 318 ancient DNA (aDNA) samples (Figure 1; Supplementary Figure 1.1; Allentoft et al., 2022) from the Mesolithic and Neolithic, including Steppe pastoralists. Comparing the ancestry at each labelled single nucleotide polymorphism (SNP, n=549,323) to genome-wide ancestry in the UK Biobank provided a “local ancestry anomaly score” (Methods), for which two regions stood out as having undergone the most significant ancestry-specific evolution in this period: LCT/MCM6, regulating lactase persistence (Itan et al., 2009), and the HLA region (Figure 2, top).

To determine whether this evolution of the HLA region has subsequently impacted diseases that are strongly associated with risk alleles found within this region, we investigated the history of variants associated with two HLA-associated autoimmune diseases, multiple sclerosis (MS) and rheumatoid arthritis (RA), using the largest ancient genome dataset from the Stone Age (full description in Allentoft et al., 20229) coupled with 86 new Medieval and post-Medieval genomes from Denmark (Supplementary Figure 1.1, Supplementary Note 1, ST1). This dataset totals 1,750 imputed diploid shotgun-sequenced ancient genomes, of which 1,509 are from Eurasia; alongside modern data, with our newly published genomes we have an almost complete transect from approximately 10,000 years ago to the present.

The allele frequencies of SNPs conferring the highest risk for MS (all in the HLA class II region) in our ancient groups show striking patterns. In particular the tag SNP (rs3135388-T) for HLA-DRB1*15:01, the largest risk factor for MS, first appeared in an Italian Neolithic individual (sampleId R3 from Grotta Continenza, C14 dated to between 5,836-5,723 BCE, coverage 4.05X) and rapidly increased in frequency around the time of the emergence of the Yamnaya culture around 5,300 years ago in Steppe and Steppe-derived populations (Figure 2). From risk allele frequencies of individuals in the UK Biobank born in, and of a ‘typical ancestral background’ for, each country (Allentoft et al., 2022), we found HLA-DRB1*15:01 frequency peaks in modern populations of Finland, Sweden and Iceland, and in ancient populations with high Steppe ancestry (Figure 2, inset).

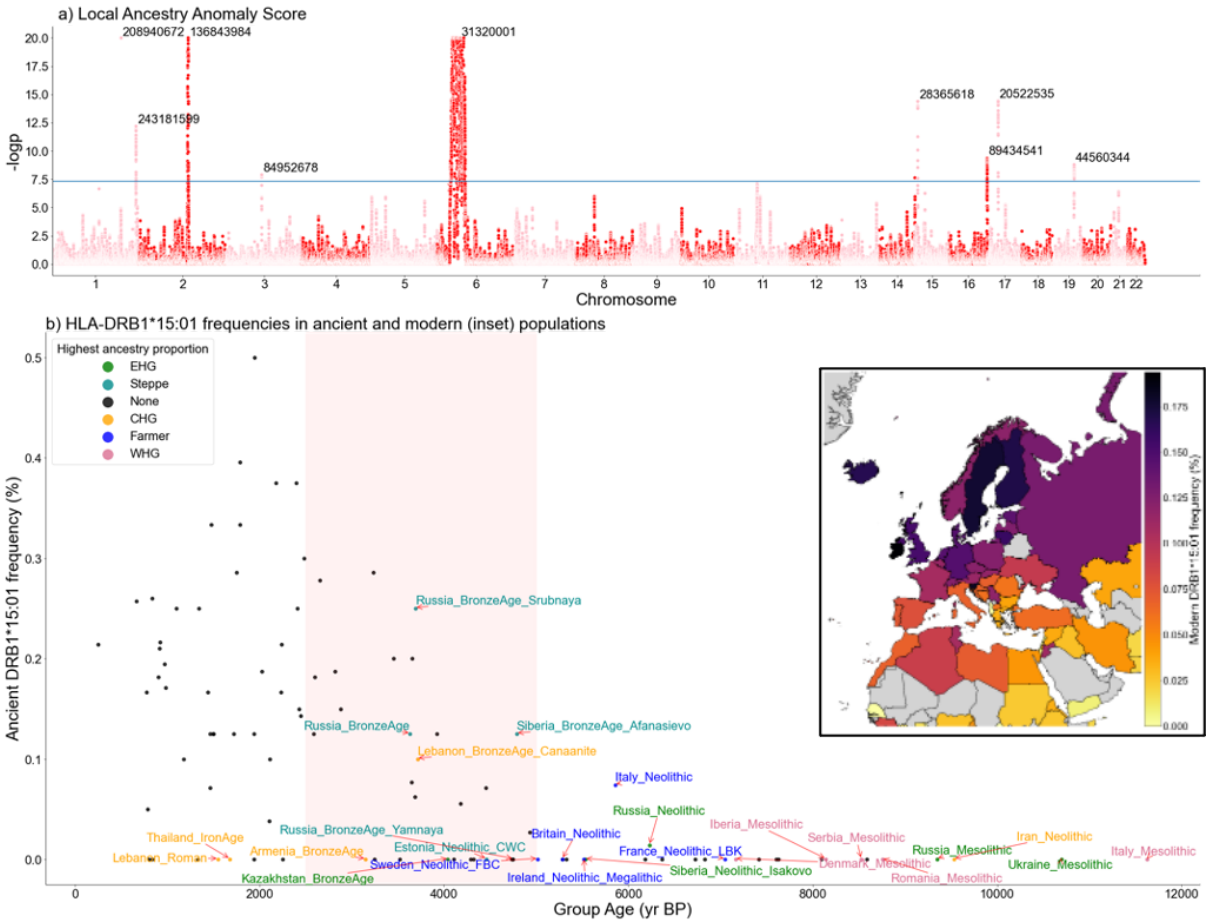


Figure 2 | Areas of unusual local ancestry in the genome, and ancient and modern frequencies of DRB1*15:01.

a) Local Ancestry Anomaly Score measuring the difference between the local ancestry and the genome-wide average (capped at -log10(p)=20; see Methods). b) HLA-DRB1*15:01 frequencies in ancient and modern (inset) populations; this is the highest effect variant for MS risk (calculated using rs3135388 tag SNP). For the ancient data, for each ancestry (CHG, EHG, WHG, Farmer, Steppe) the five populations with the highest amount of that ancestry are coloured and labelled. DRB1*15:01 was present before the Steppe expansion, but rose to high frequency during the Yamnaya formation (shaded red). The geographical distribution of DRB1*15:01 frequency in modern populations in the UK Biobank is also shown (inset).

To investigate the risk of a particular ancestry at all MS-associated fine-mapped loci present in the UK Biobank imputed dataset (n=205/233, IMSGC, 2019, see methods), we used the local ancestry dataset to calculate a risk ratio (see Methods: Weighted Average Prevalence) for each ancestry. For MS, Steppe ancestry has the highest risk ratio in nearly all HLA SNPs, while Farmer and 'Outgroup' ancestry (represented by ancient Africans) are often the most protective (Figure 3, top), meaning a Steppe-derived haplotype at these positions confers MS risk.

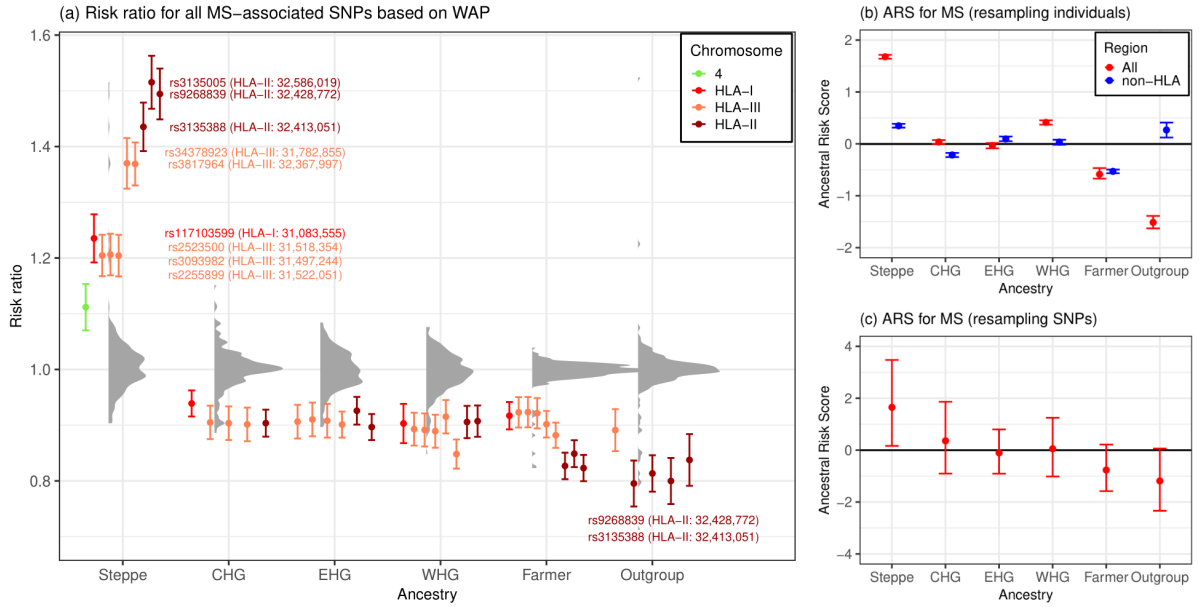


Figure 3 | Associations between local ancestry and MS in a modern population.
a) Risk ratio of SNPs for MS based on weighted average prevalence (WAP; see Methods), when decomposed by inferred ancestry. A mean and standard deviation are calculated for each ancestry based on bootstrap resampling, for each chromosome. The distribution of all SNPs' risk ratios at each ancestry are shown as a raincloud plot, while only SNPs significant at the 1% level are shown individually, coloured by chromosome or HLA region, and those with risk ratio >1.2 or <0.8 are annotated with rsID, HLA region and position (build GRCh37/hg19). b-c) Genome-wide Ancestral Risk Scores (ARS, see Methods) for MS for all associated SNPs (red) or non-HLA SNPs only (blue). Confidence intervals are estimated by either bootstrapping over individuals (b, which can be interpreted as testing power to reject a null of no association between MS and ancestry) and bootstrapping over SNPs (c, which can be interpreted as testing whether ancestry is associated with MS genome-wide).

Having shown that some ancestries carry higher risk at particular SNPs, we wanted to calculate an aggregate risk score for each ancestry. We used a statistic, the Ancestral Risk Score (ARS, introduced in Irving-Pease et al., 2022), which is equivalent to a polygenic risk score (PRS) for a modern individual consisting of entirely one ancestry. ARS offers an improvement on calculating PRS using ancient genotype calls directly, as it mitigates the effects of low aDNA sample numbers and bias (Dehasque et al., 2020), while being robust to intervening drift and selection. We used effect size estimates from previous association studies, under an additive model, with confidence intervals obtained via an accelerated bootstrap (Efron, 1987) (Supplementary Note 4). In the ARS for MS (Figure 3b), Steppe ancestry had the largest risk, followed by WHG, CHG and EHG; Farmer and Outgroup ancestry had the lowest ARS. Therefore, Steppe ancestry is contributing the most risk for

MS across all associated SNPs. We tested for a genome-wide association by resampling loci, and found that Steppe risk is much reduced but still clearly exceeds Farmer (Figure 3c). Although most of the signal is driven by SNPs in the HLA region, this pattern persists even when excluding these SNPs (Figure 3b).

The fact that all but two MS-associated HLA SNPs confer risk within Steppe ancestry implies that this risk has a common evolutionary history. We therefore investigated whether ancestry was important for prediction using three types of association study in the UK Biobank for disease-associated SNPs, controlling for age, sex and the first 18 PCs. The first of these is a regular SNP-based association as conducted in GWAS. The second uses local ancestry probabilities instead of genotype values (Supplementary Note 3). The third is based on Haplotype trend regression (HTR) which is used to detect interactions between SNPs (Zaykin et al., 2002) by treating each haplotype's probability as a feature from which to predict a trait, instead of using SNPs as in a regular GWAS. We developed a new method called Haplotype Trend Regression with eXtra flexibility (HTRX, Supplementary Note 5) that searches for haplotype patterns that include single SNPs and non-contiguous haplotypes. To prevent overfitting, we reported out-of-sample variation explained, and showed by simulation (see Supplementary Figure 4.4) that HTRX predicts the same variance as regular GWAS when interactions are absent, but explains more variance when the interaction strength increases.

Although our cohort of self-identified "white British" individuals is relatively under-powered with respect to MS (cases=1,949; controls=398,049; prevalence=0.487%), MS was associated with Steppe and Farmer ancestry (p<1e-10) in the HLA region (Supplementary Figure 4.1). In 3 out of 4 main LD blocks within the HLA (class I, two subregions of class II determined by LD blocks at 32.41-32.68Mb and 33.04-33.08Mb, and class III), local ancestry explains significantly more variation in total than SNP variation (Figure 4; measured by average out-of-sample McFadden's R^2 for logistic regression, a widely used statistic for estimating the variance explained by the logistic regression models, see Methods). While increased ancestry performance over GWAS can be explained by tagging of SNPs outside the region, increased HTRX performance over GWAS quantifies the total effect of a haplotype, including rare SNPs and epistasis. Across the entire HLA region, haplotypes explain at least 17% more out-of-sample variation than GWAS (2.90%, compared to 2.48%). Interaction signals are also observed within class I, within class II, and between class I and III.

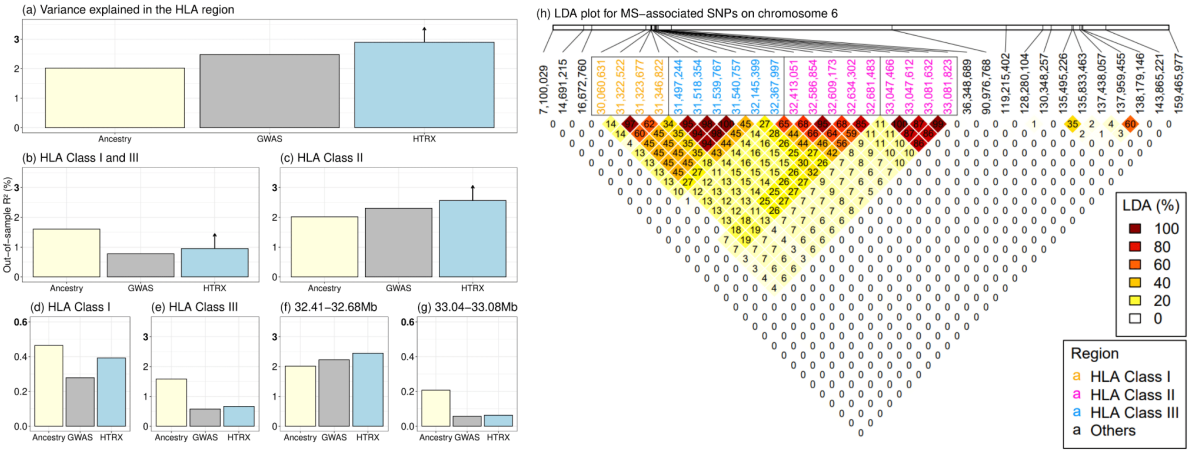


Figure 4 | MS association in the HLA.

Comparison of variance explained in MS within the UK Biobank, for all fine-mapped HLA SNPs with an independent contribution (IMSGC, 2019). The plots compare GWAS (treating SNPs as having independent effect), local ancestry at those SNPs, and HTRX (haplotypes) after accounting for covariates (Methods). a) is for fine-mapped MS-associated SNPs in the HLA. b) is HLA class I and -III, c) is HLA class II, d) is HLA class I, e) is HLA class III, f) and g) are subregions of HLA class II chosen from LD. HTRX has small “up-arrows” where these are lower bounds (Methods). h) Genetic correlations in the HLA region at our time-depth from Ancestry-based LD (LDA, see Methods) and Supplementary Figure 6.5 for LD.

Multiple SNPs at the 32.41-32.68Mb region are Steppe-associated, have high MS odds ratios, and are in LDA (Figure 4), which may explain the increased HTRX variance explained. We further tested whether co-occurring ancestries at each loci were associated with MS (Methods; Supplementary Figure 4.2), but found no evidence that risk was associated with anything other than Steppe ancestry.

Having established that Steppe ancestry contributes most of the HLA-associated risk for MS, we investigated evidence for polygenic selection on the disease-associated variants using two methods. Firstly, we used a novel chromosome painting technique based on inference of a sample’s nearest neighbours in the marginal trees of an ARG that contains labelled individuals (Irving-Pease et al., 2022). The resulting ancestral path labels, for haplotypes in both ancient and modern individuals, allowed us to infer allele frequency trajectories for risk associated variants, while controlling for changes in admixture proportions through time. These paths extend backwards from the present day to approximately 15,000 years ago, and are labelled with the unique population that a path travels through. We stress that the path labels are not representative of a continuous population, but represent a path backwards in time that encompasses that ancestry. For example, the CHG path originates in

Caucasus hunter-gatherers, before merging with EHG to form the Steppe population, and then merges with other ancestries in later European populations (Figure 1).

For our ancestry path analysis, a substantial fraction of the fine-mapped MS variants were not imputed in our ancient dataset, due to quality control filtering and the difficulty of accurately inferring HLA alleles in ancient samples (Theusen et al., 2022). To address this, we LD-pruned genome-wide significant summary statistics from the same study (IMSGC, 2019), for which we could reliably assign ancestry path labels (n = 62, see Methods). This allowed us to test for polygenic selection across disease-associated variants using CLUES (Stern et al., 2019) and PALM (Stern et al., 2021). CLUES was used to infer allele frequency trajectories and selection coefficients for the set of SNPs using a time-series of imputed aDNA genotype probabilities obtained from 1,015 ancient West Eurasian samples that passed all quality control filters. PALM was used to infer polygenic selection gradients and p-values for each trait, i.e. across all trait-associated SNPs.

For MS, we found evidence that disease risk was selectively increased when considering all ancestries collectively (p=5.06e-05; ω=0.0029), between 5,000-2,000 years ago (Figure 5). Conditioning on each of the four long-term ancestral paths (CHG, EHG, WHG and ANA), we found a statistically significant signal of selection in CHG (p=6.45e-3; ω=0.009). None of the other ancestral paths reached nominal significance, although ANA (p=0.0743; ω=0.011) and EHG (p=0.064; ω=0.0045) paths were close. Again, it is likely that the selection occurred in the pastoralist population of the Steppe, as that population consists of approximately half CHG ancestry (Jones et al., 2015, Figure 1). The SNP driving the largest change in genetic risk over time was rs3129934, in both the pan-ancestry (p=9.52e-06; s=0.017) and CHG (p=0.019; s=0.008) analyses, which tags the HLA-DRB1*15:01 haplotype (Comabella et al., 2008). We also tested three other alleles that tag the HLA-DRB1*15:01 haplotype (rs3129889, rs3135388 and rs3135391) for evidence of selection, and found that the ancestry stratified signal was consistently strongest in CHG (Figure 5). None of the four tag SNPs were detected on either the EHG or WHG backgrounds, indicating that the HLA-DRB1*15:01 haplotype likely originated in the basal population ancestral to both ANA and CHG.

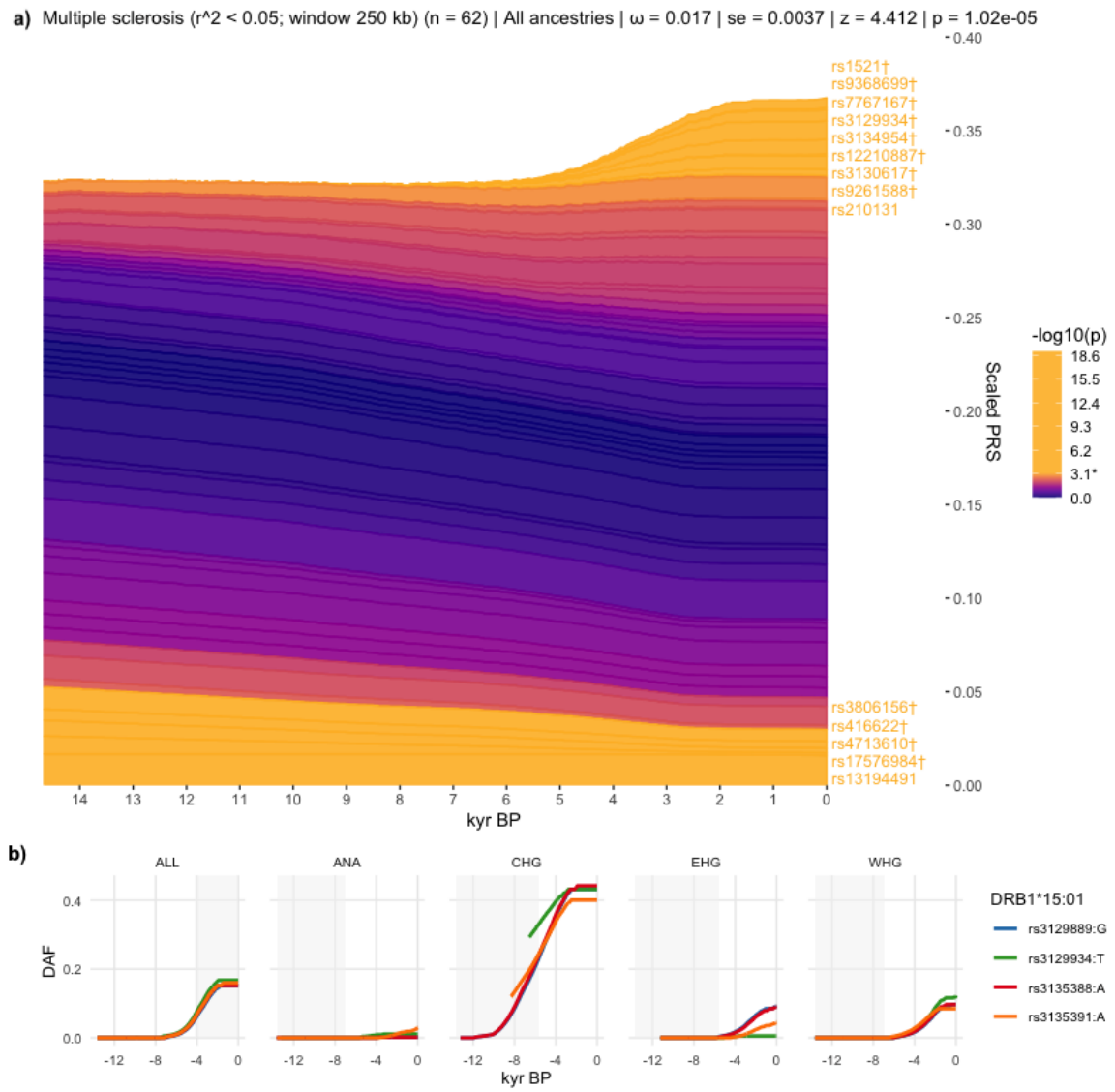


Figure 5 | Evidence for selection on MS-associated SNPs.

a) Stacked line plot of the pan-ancestry PALM analysis for MS, showing the contribution of alleles to disease risk over time. Individual SNPs are stacked, with their trajectories polarised to show the frequency of the positive risk allele and weighted by their scaled effect size: when a given SNP bar becomes wider over time the risk allele has increased in frequency, and vice versa. SNPs are sorted by their marginal p-value and direction of effect, with selected SNPs that increase risk plotted on top. SNPs are also coloured by their marginal p-values, and significant SNPs are shown in yellow. The y-axis shows the scaled polygenic risk score (PRS), which ranges from 0 to 1, representing the maximum possible additive genetic risk in a population.

b) Maximum likelihood trajectories for four SNPs tagging DRB1*15:01. The background is shaded for the approximate time period in which the ancestry path existed as the population it is named after (i.e. WHG, EHG, CHG or Anatolian Neolithic Farmer). None of the tagging alleles are present on the EHG or WHG ancestral paths.

To further examine the nature of selection, we developed a new summary statistic, Linkage Disequilibrium of Ancestry (LDA). LDA is the correlation between local ancestries at two SNPs, measuring whether recombination events between ancestries are high compared to recombination events within ancestries. We subsequently defined the “LDA score” of a SNP as the total LDA of the SNP with the rest of the genome. A high LDA score indicates that the haplotype inherited from the reference population is longer than expected, while a low score indicates that the haplotype is shorter than expected (i.e. underwent more recombination). For example, the LCT/MCM6 region exhibits a high LDA score (Extended Data Figure 7.4), as expected from a relatively recent selective sweep (Bersaglieri et al., 2004).

The HLA has significantly *lower* LDA scores than the rest of chromosome 6 (Supplementary Figure 6.4). We simulated the LDA score under selection (Supplementary Figure 6.1; Methods), which showed that when SNP frequencies are increasing in the most recent population, single locus selection cannot explain this signal (Supplementary Figure 6.2-3). Instead, different loci in LD must have independently reached high frequency in different ancestral populations that admixed, with selection favouring haplotypes of mixed ancestry over single-ancestry haplotypes. Although multi-SNP selection has been modelled (He et al., 2020), the interaction with prior population structure is less explored and is important for the HLA, justifying a new term, "recombinant favouring selection".

The HLA region contains the highest “Outgroup” ancestry anywhere on the genome (Figure 6), reflecting high nucleotide diversity. Unlike other measures of balancing selection such as Fst (Figure 6), LDA describes excess ancestry LD from specific, dated populations and therefore need not be correlated with them. For the HLA class II region, the selection measures all line up (LDA score, Fst, pi), but for class I, the LDA score has an additional non-diverse minimum at 30.8Mb, implying that here the genome is ancestrally diverse but genetically strongly constrained. The LDA score is thus informative about the type of selection being detected, and whether it has been subject to change.

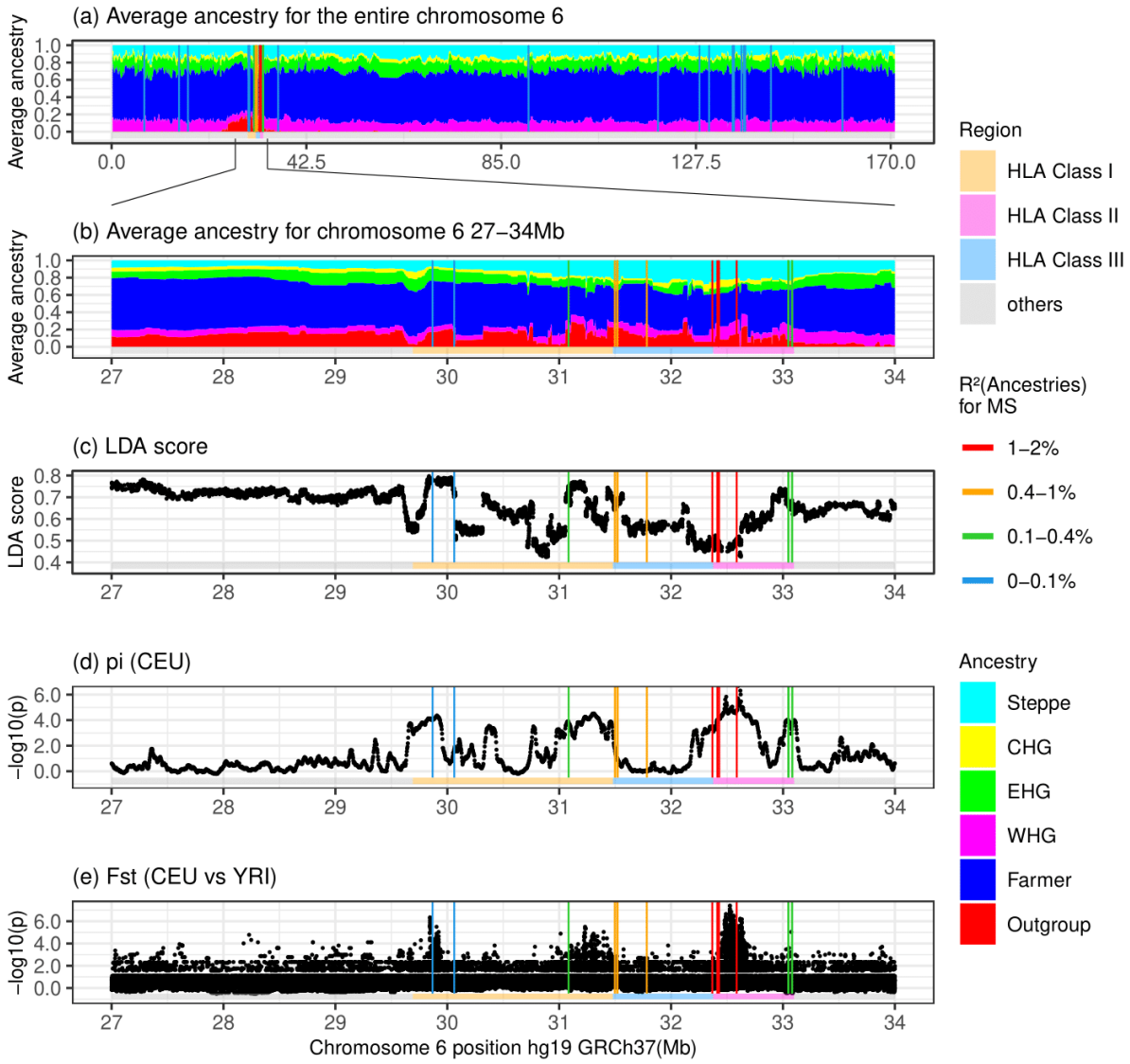


Figure 6 | Signatures of selection at the HLA locus showing different regions of the HLA (coloured bar) and locations of MS-associated SNPs (vertical lines, coloured by the variance explained by 6 ancestries).

a): Whole Chromosome 6 “local ancestry” decomposition by genetic position. b). HLA “local ancestry” decomposition. c): LDA score; low values are indicative of selection for multiple linked loci, while high values indicate positive selection. d): pi scores (nucleotide diversity) for CEU (Northern and Western European ancestry). MS-associated SNPs fall in highly diverse regions of the HLA. e): Fst scores (divergence between two populations) for CEU vs YRI(Yoruba); locally higher scores indicate regions that have undergone differential selection between the two populations.

Because MS would not have conferred a fitness advantage on ancient individuals, it is likely that this selection was driven by traits with shared genetic architecture, of which increased risk for MS in the present is a consequence. We therefore looked at LD-pruned MS-associated SNPs that showed statistically significant evidence for selection using

CLUES (n=26) and which also had a genome-wide significant trait association ($p < 5e-8$) in any of the 4,359 traits from the UK Biobank (Bycroft et al. 2018; UK Biobank Neale Lab, Round 2: http://www.nealelab.is/uk-biobank/). We found that many selected SNPs are also associated with celiac disease (n=15), white blood cell/neutrophil count (n=15/n=15), hypothyroidism (n=14) and haemoglobin concentration (n=14) (Supplementary Figure 7.1). This raised the possibility that the selection had increased risk for both MS and celiac disease, and when we tested celiac disease for polygenic selection, we found significant evidence for positive selection, increasing genetic risk ($p=9.65e-3$; $\omega=0.846$, Supplementary Note 6).

Because the UK Biobank is underpowered with respect to many traits and diseases, we also undertook a manual literature search (see methods) for all SNPs that reached genome-wide significance for association with MS in the summary stats (i.e., not LD-pruned, as independence is not required) and which showed statistically significant evidence for selection using CLUES (n=94). We found that most of the alleles under positive selection are associated with protective effects against specific pathogens (virus, bacteria, fungi and parasites) and/or infectious diseases within one or several ancestral paths (disease or pathogen associated/total selected in ancestry path: pan-ancestry 36/44; ANA 24/31; CHG 25/29; EHG 27/35; WHG 9/10, Supplementary Note 8, ST13, Supplementary Figure 8.1), although we note that GWAS data for many infectious diseases are not available. We observed that the selected alleles had protective associations with several chronic viruses (EBV, VZV, HSV, and CMV) and to viruses or diseases not associated with transmission in small hunter-gatherer groups (e.g., measles, mumps, influenza, whooping cough). Moreover, many selected alleles conferred a reduction of risk of parasites, of skin and subcutaneous tissue, gastrointestinal, respiratory, urinary tract, and sexually transmitted infections, or of pathogens associated with these or other infections (e.g., malaria, toxoplasmosis, entamoeba histolytica, clostridium difficile, tuberculosis, streptococcus pyrogenes, and chlamydia) (Supplementary Note 8, ST13, Supplementary Figure 8.1).

We contrasted these findings for MS with results for RA, a common inflammatory HLA class II-associated disease that primarily affects the joints causing pain, swelling and stiffness (Fugger et al., 2020), which shows a strikingly different ancestry risk profile. HLA-DRB1*04:01 is the largest genetic risk factor for RA; in the CLUES analysis, the tag SNP for this allele (rs660895) displayed evidence of continuous negative selection until approximately 3,000 years ago ($p=4.63e-4$, Supplementary Figure 5.1). We found that WHG and EHG ancestries often confer the most risk at SNPs associated with RA (Relative Risk ratio of RA-associated SNPs based on WAP, see Methods); and these ancestries have

contributed the greatest risk for RA on aggregate, reflected in a higher ARS for these ancestries (Supplementary Note 4), while Steppe and Outgroup ancestry have the lowest scores (Supplementary Figure 3.1). These results were recapitulated in the local ancestry GWAS (Supplementary Note 3).

We found that RA-associated SNPs have undergone negative polygenic selection ($p = 3.26e-3$, Extended Data Figure 6.1) over the last approximately 15,000 years. When decomposed by ancestry path, we found that all paths exhibited a negative selection gradient, but none achieved nominal significance; although the CHG ($p = 6.33e-2$; $\omega = -0.014$) path came close.

These results demonstrate that genetic risk for RA was higher in the distant past, in contrast to MS, with RA-associated risk variants present at higher frequencies in European hunter-gatherer populations before the arrival of agriculture. In order to understand what caused the high risk in hunter-gatherer populations and subsequent negative selection, we again undertook a manual literature search for pleiotropic effects of SNPs associated with RA. Because the number of SNPs that reached genome-wide significance in the GWAS study and also showed statistically significant evidence for directional selection was large, we only analysed LD-pruned SNPs (n=42). We found that the majority of selected SNPs were associated with protection against distinct pathogens and/or infectious diseases across all paths (disease or pathogen associated/total selected in ancestry path: pan-ancestry 9/13; ANA 10/13; CHG 8/11; EHG 10/16; WHG 10/12). We found that selected RA-risk alleles were often linked to the same pathogens or diseases as in the MS analysis, although the number of protective associations to distinct pathogens were fewer (Supplementary Note 8, ST14, Supplementary Figure 8.1).

Discussion

The last 10,000 years have seen some of the most extreme global experiments in lifestyle with the emergence of farming in some regions and a pastoral lifestyle in others. While 5,000 years ago farmer ancestry predominated across Europe, a relatively diverged ancestry arrived with the Steppe migrations around this time. We have shown that this ancestry contributes the most genetic risk for MS today, and that these variants were the result of positive selection coinciding with the emergence of a pastoralist lifestyle on the Pontic-Caspian Steppe, and continued selection in the subsequent admixed post-Stone Age populations in Europe. This ultimately created a legacy of heterogeneity in MS risk observed across Europe today. These results address the long-standing debate around the north-south gradient in MS prevalence in Europe, and suggest that the Steppe ancestry gradient in modern populations - specifically at the HLA region - across the continent causes this phenomenon in combination with environmental factors. Furthermore, while epistasis between MS-associated variants in the HLA region has been demonstrated before (Gregersen et al., 2006, Wang et al., 2011, Cotsapas et al. 2018, Slim et al. 2022), we have shown that accounting for this explains 17% more variance than independent SNPs effects alone. Many of the haplotypes carrying these risk alleles have ancestry-specific origins, which could be exploited for individual risk prediction and may offer a pathway from ancestry associations into a mechanistic understanding of MS risk. We have contrasted these findings with results for rheumatoid arthritis (RA), another HLA class II associated chronic inflammatory disease, and found that the genetic risk for RA exhibits a contrasting pattern: genetic risk was highest in Stone Age hunter-gatherer ancestry and decreased over time.

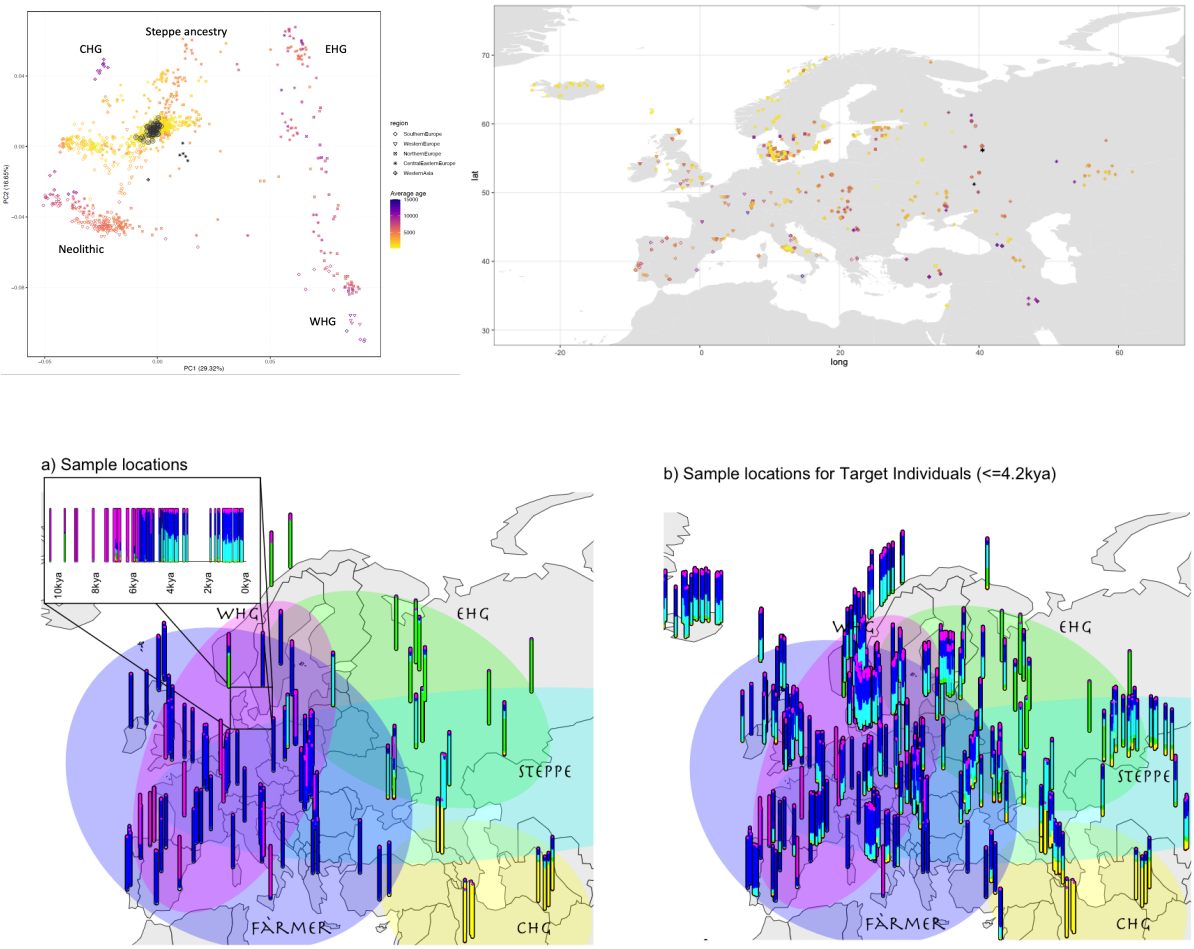
Our interpretation of this history is that co-evolution between pathogens and their human hosts has resulted in massive and divergent ancestry-specific selection on immune response genes according to lifestyle and environment, driven by a range of pathogenic drivers, and “recombinant favouring selection” after these populations merged. The Late Neolithic and Early Bronze Age was a time of massively increased infectious diseases in human populations, due to increased population density as well as contact with, and consumption of, domesticated animals. Many diseases trace their origins to this period, such as tuberculosis (TB) caused by the intracellular bacteria *Mycobacterium tuberculosis* or *Mycobacterium bovis* (Bos et al., 2014; Sabin et al., 2020), bubonic plague caused by *Yersinia pestis* (Rasmussen et al., 2015, Spyrou et al., 2018; Rascovan et al., 2019), herpes simplex virus (Guellil et al., 2022), and chickenpox caused by varicella-zoster virus (Pontremoli et al., 2019), and we have provided evidence that many of the MS- and

RA-associated variants under selection confer resistance to a range of infectious diseases and pathogens (Supplementary Note 8). For example, HLA-DRB1*15:01 is associated with protection against TB (Tian et al., 2017) and increased risk for lepromatous leprosy (Krause-Kyora et al., 2018). However, we are underpowered to detect specific associations beyond this hypothesis due to poor knowledge of the distribution and diversity of past diseases, poor preservation of endogenous pathogens in the archaeological record, and a lack of well-powered GWAS studies for many infectious diseases.

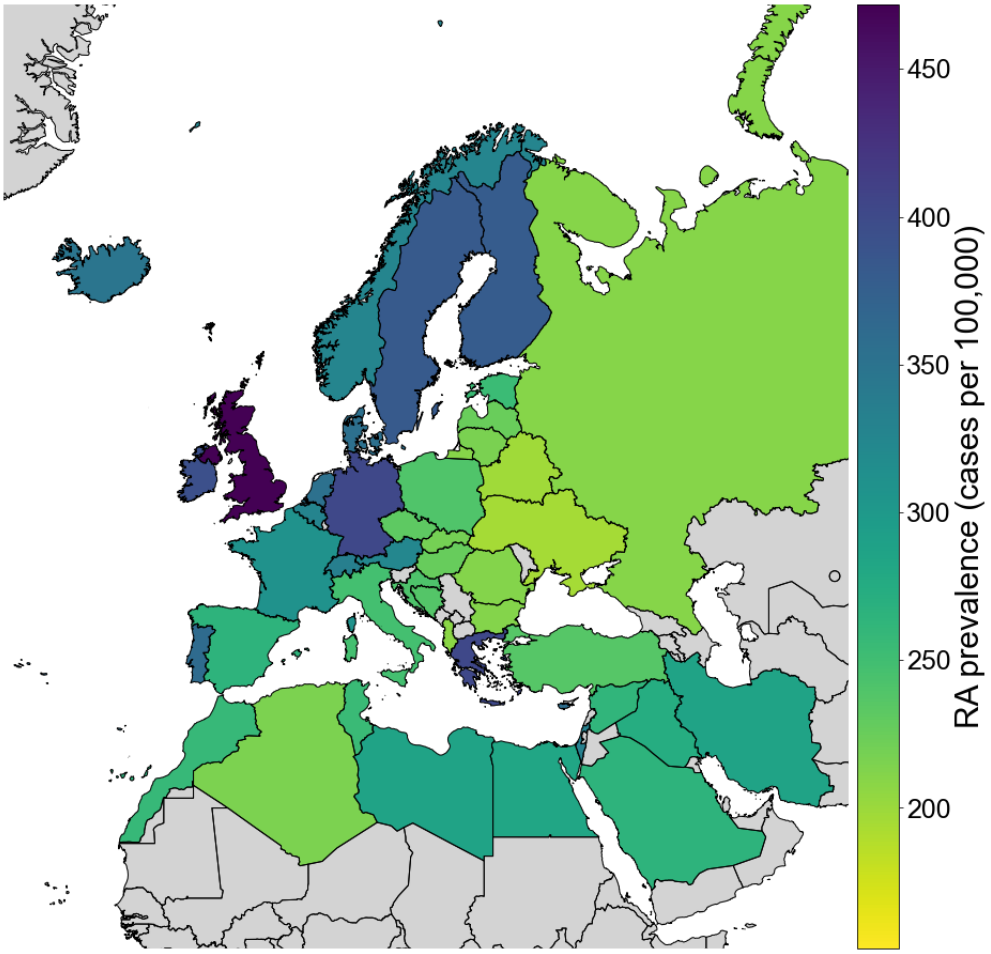
A pattern that repeatedly appears is that of lifestyle change driving changes in risk and phenotypic outcomes. We have shown that in the past environmental changes driven by lifestyle innovation inadvertently drove an increase in genetic risk for MS. Today, with increasing prevalence of MS cases observed over the last five decades (Wallin et al., 2019; GBD 2016 Neurology Collaborators, 2019), we again observe a striking correlation with changes in our environment, including lifestyle choices and improved hygiene, which no longer favours this previous genetic architecture. Instead, the fine balance of genetically-driven cells within the immune system, which are needed to combat a broad repertoire of pathogens without harming self-tissue, has been met with new challenges, including a potential absence of requirement. For example, while a population of immune cells, T helper 1 (Th1), direct strong cellular immune responses against intracellular pathogens, T helper 2 (Th2) cells mediate humoral immune responses against extracellular bacteria and parasites and further have the capacity to guide the restoring of homeostasis, thus preventing damage of the infected tissue via immune-regulatory cytokines. We have shown that the majority of selected MS-associated SNPs are associated with protection against a wide range of pathogens, consistent with strong but balanced Th1/Th2 immunity in the Bronze age, where a diversification of pathogens likely took place. In contrast, although MS pathogenesis is complex and multicellular of nature, CD4+ Th cells, in particular IFN-ɣ producing Th1 cells and IL-17-producing Th17 cells play a key role in disease development (Attfield et al., 2022). The skewed Th1/Th2 balance observed in MS may partly result from the developed world's increased sanitation, which has led to a drastically reduced burden of parasites, which the immune system had evolved to efficiently combat (Fleming and Fabry, 2007). In the case of RA, the exposure of Hunter Gatherer populations to the respiratory or gastrointestinal pathogens linked to triggering RA (Joo et al., 2019) was likely low. The new pathogenic challenges associated with agriculture, animal domestication, pastoralism, and higher population densities might have substantially increased the risk of developing RA in genetically predisposed individuals, resulting in negative selection. If true, this would present a parallel between RA in the Bronze Age and MS today, in which lifestyle changes have exposed previously favourable genetic variants as autoimmune disease risks.

More broadly, it is clear that this was a critical period in human history during which highly genetically and culturally divergent populations evolved and eventually mixed. These separate histories dictate the genetic risk and prevalence of several autoimmune diseases today. Surprisingly, the emergence of the pastoralist Steppe lifestyle may have had an impact on immune response as great as or greater than the emergence of farming during the Neolithic transition, commonly held to be the greatest lifestyle change in human history.

Supplementary Figures

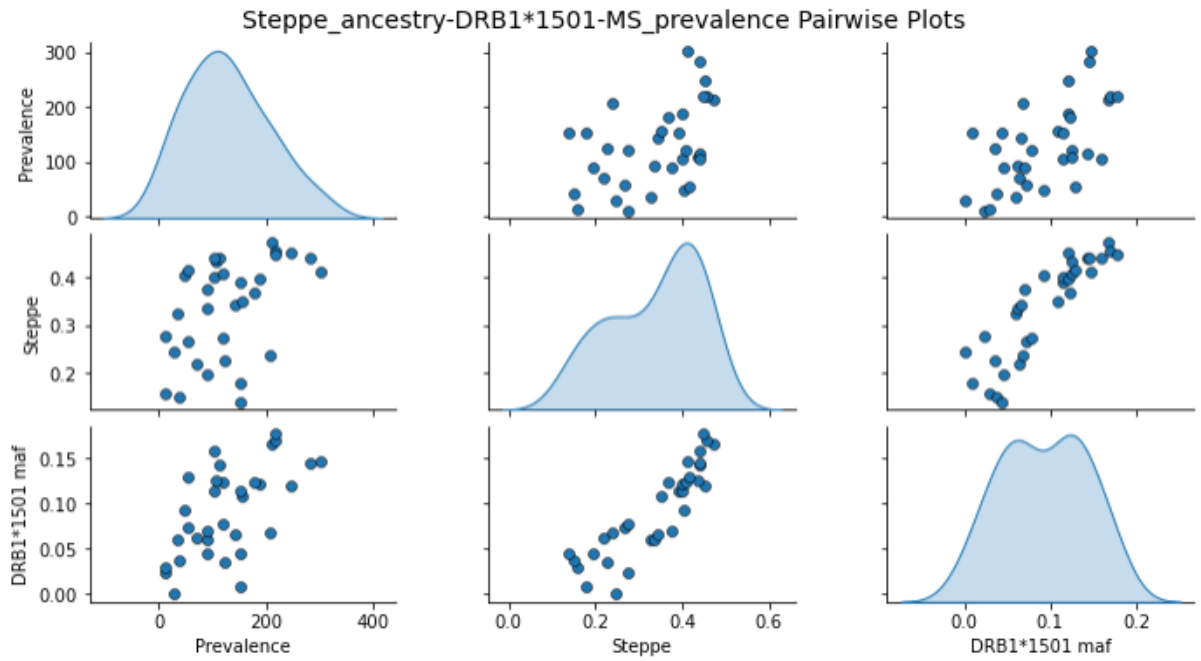


Supplementary Figure 1.1 | Ancient sample PCA, map, ancestry proportions through time for samples in Denmark.
(1) PC1 vs PC2 of the filtered Western Eurasian ancient samples included in this study. Black circled points are Danish Medieval and post-Medieval samples published here for the first time. Major component ancestry locations are labelled. (2) Map of ancient filtered Western Eurasian ancient samples included in this study (3a) Map of reference data and time transect of Denmark as in Figure 1. (3b) More recent ancient data (samples <4,200 years ago) not used as reference, showing the clines of the main ancestry components from (3a).

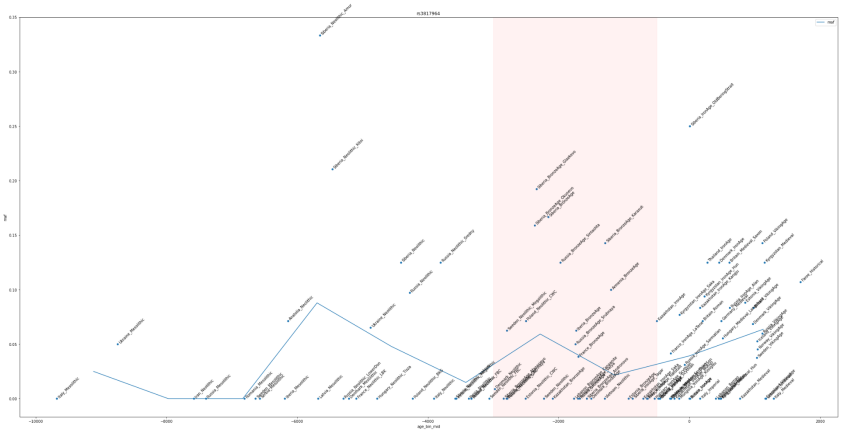


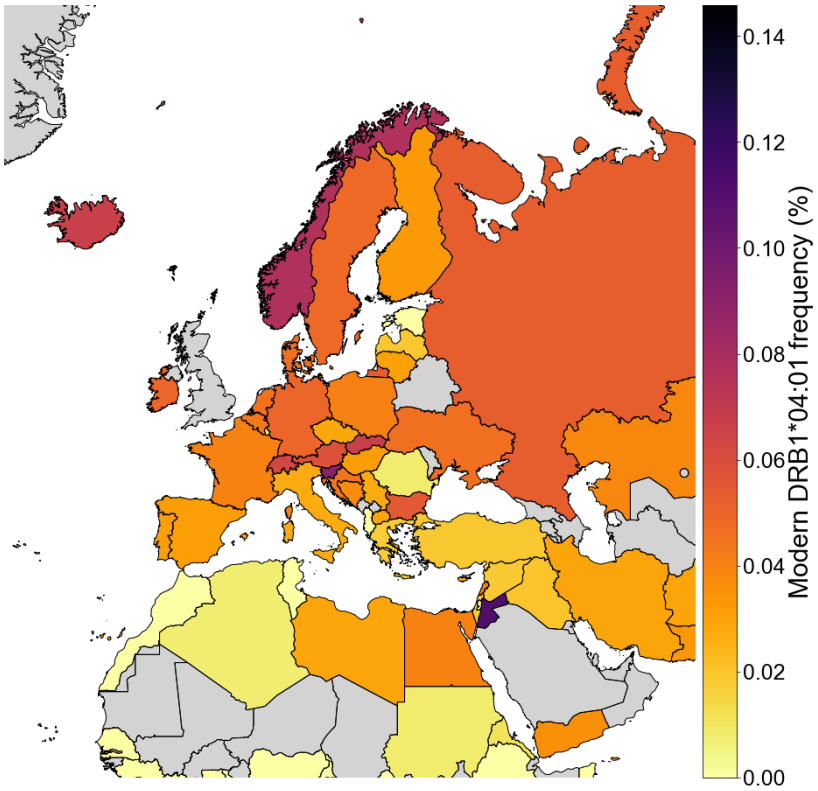
Supplementary Figure 1.2 | Modern prevalences of RA.

Modern-day geographical distribution of RA prevalence in Eurasia and North Africa. Prevalence data for RA (cases per 100,000) was obtained from Safiri et al. (2019).

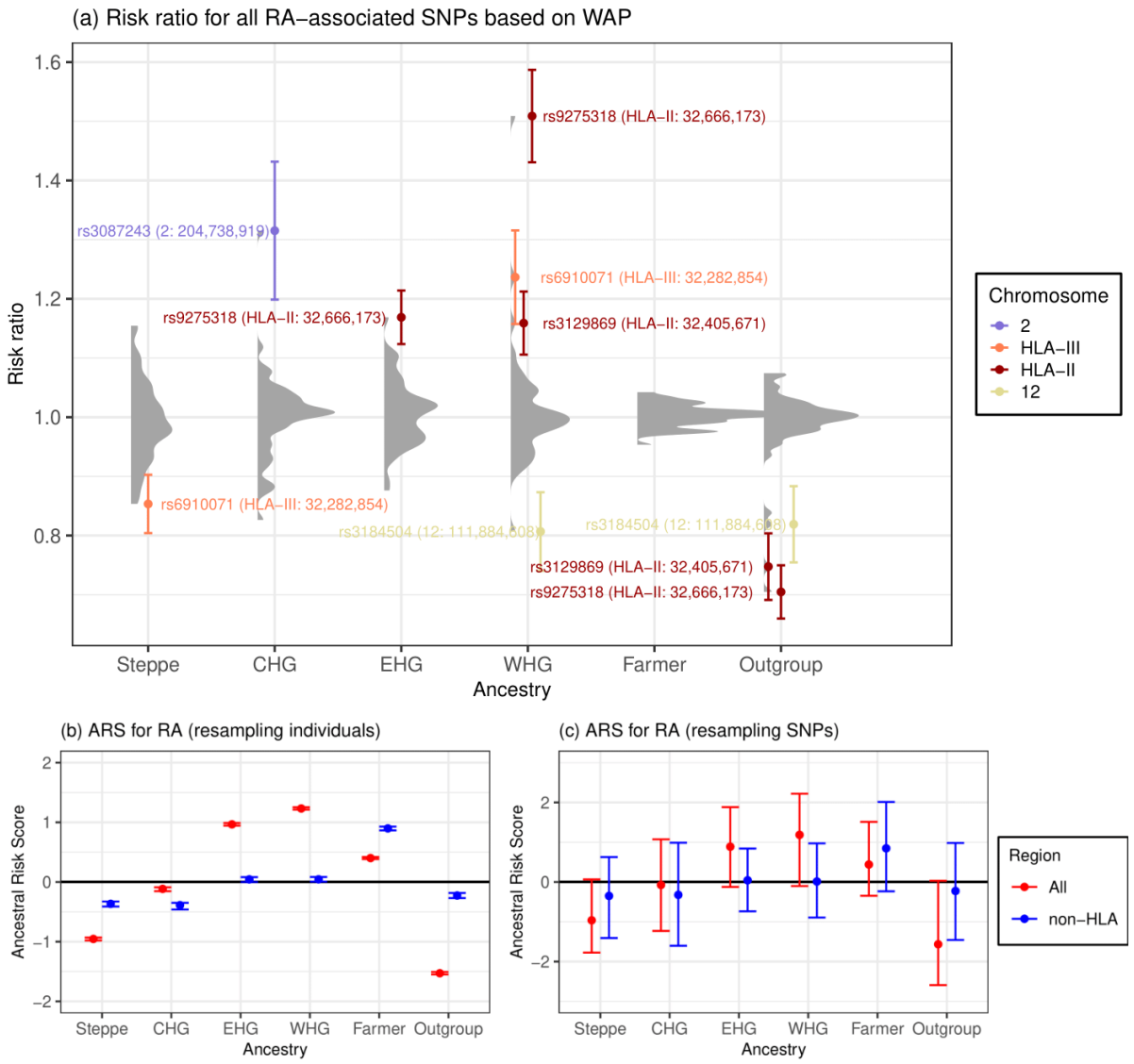


Supplementary Figure 1.3 | Association between genome-wide Steppe ancestry, MS prevalence and DRB1*15:01 frequency in modern populations in the UK Biobank.



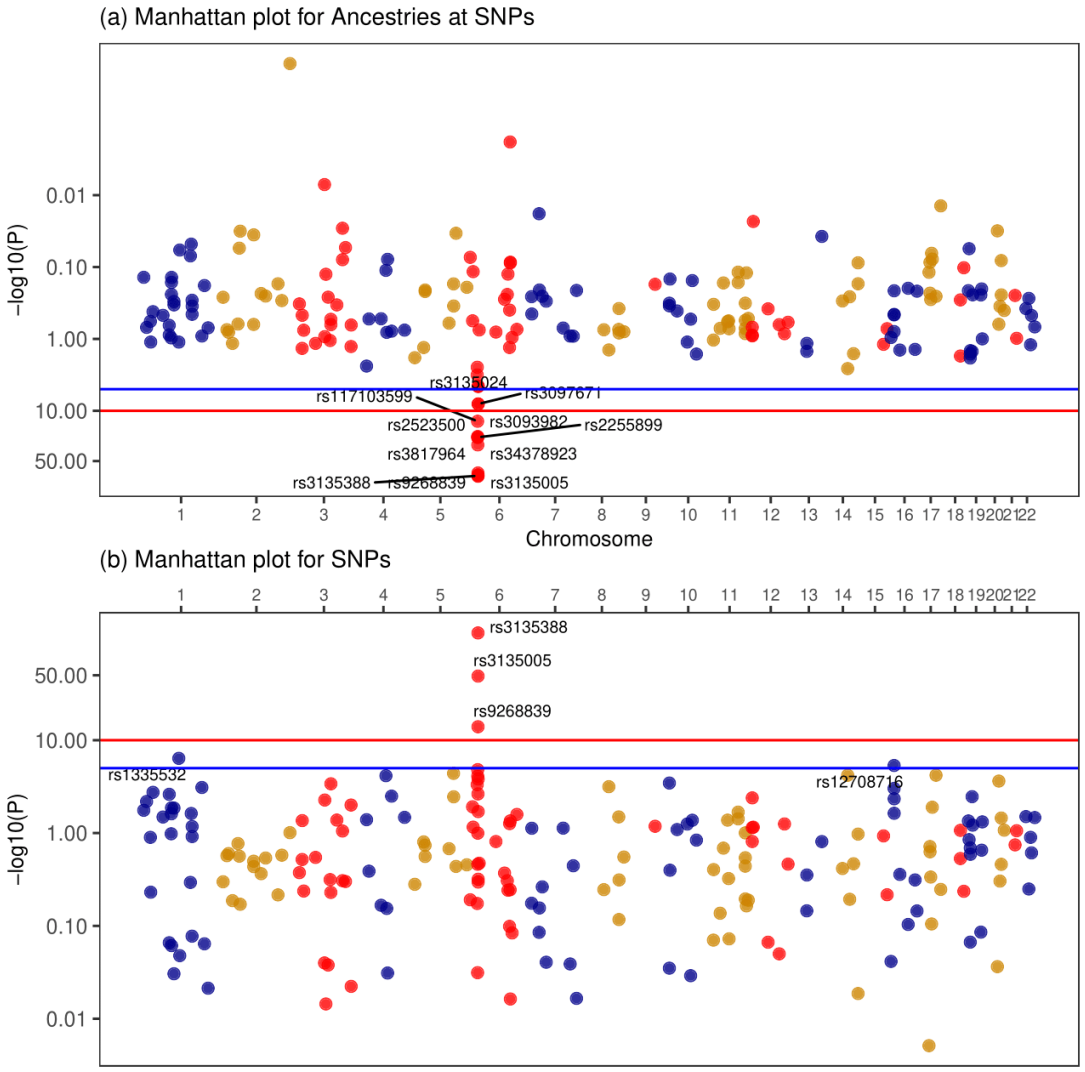


Supplementary Figure 2.1 | Ancient and modern prevalences of HLA-DRB1*04:01 (rs3817964). Top: Ancient distribution of HLA-DRB1*04:01, the largest genetic risk factor in RA. Average frequency across all populations is shown (blue line, 10 time bins) as well as the Bronze Age (red shading). Bottom: Modern distribution of HLA-DRB1*04:01 in populations in the UK Biobank. NB the tag SNPs may be less effective at tagging these types in non-European populations, so we urge caution in interpretation - especially in African populations.



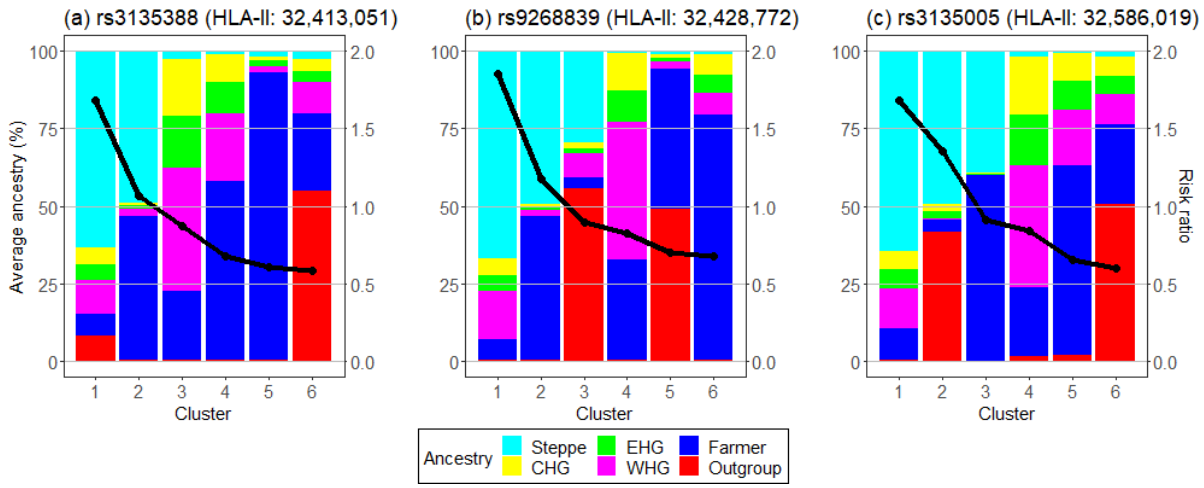
Supplementary Figure 3.1 | Associations between local ancestry and RA in a modern population.

a) Risk ratio of SNPs for RA based on weighted average prevalence (WAP; see Methods), when decomposed by inferred ancestry. A mean and standard deviation are calculated for each ancestry based on bootstrap resampling, for each chromosome. The distribution of all SNPs' risk ratios at each ancestry are shown as a raincloud plot, while only SNPs significant at the 1% level are shown individually, coloured by chromosome or HLA region, and those with risk ratio >1.1 or <0.9 are annotated with rsID, HLA region and position (build GRCh37/hg19). b-c) Genome-wide Ancestral Risk Scores (ARS, see Methods) for RA for all associated SNPs (red) or non-HLA SNPs only (blue). Confidence intervals are estimated by either bootstrapping over individuals (b, which can be interpreted as testing power to reject a null of no association between RA and ancestry) and bootstrapping over SNPs (c, which can be interpreted as testing whether ancestry is associated with RA genome-wide).



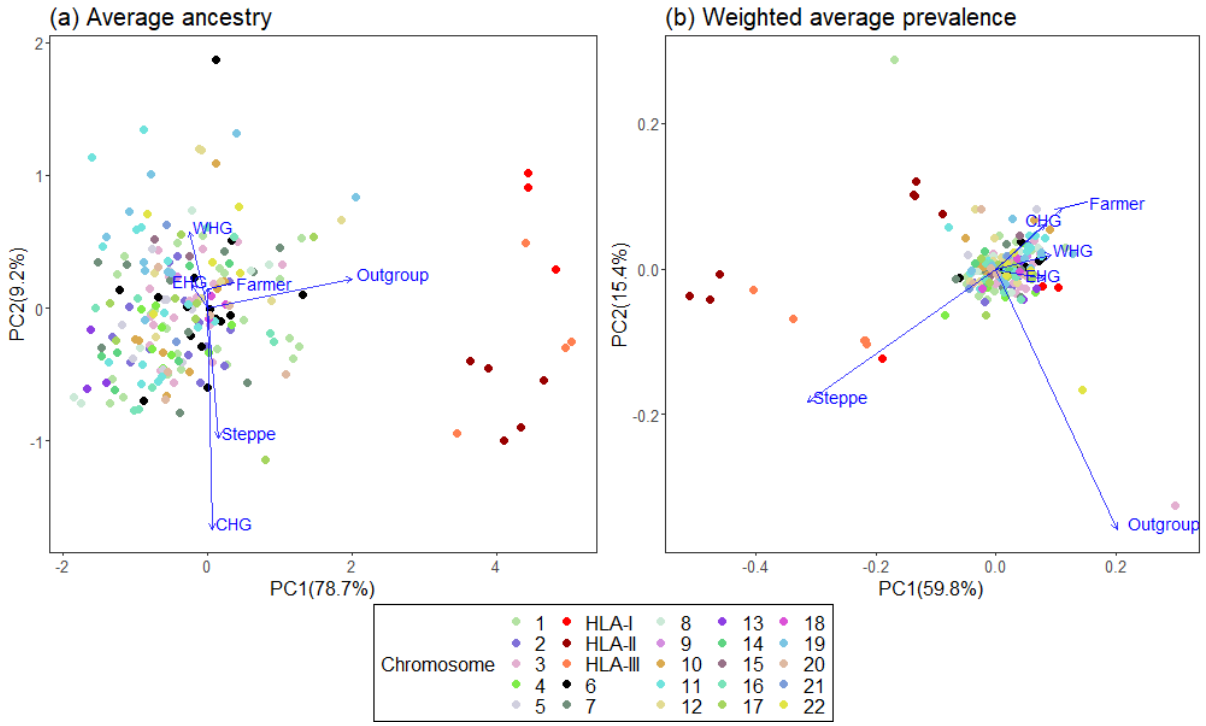
Supplementary Figure 4.1 | Association with MS risk at externally ascertained SNPs, for all 6 ancestries (top, see Methods), and SNPs (bottom).

Due to the UK Biobank being less powered (having fewer cases) than the Case-Control study from which these SNPs were ascertained, the only statistically significant associations here are in the HLA.

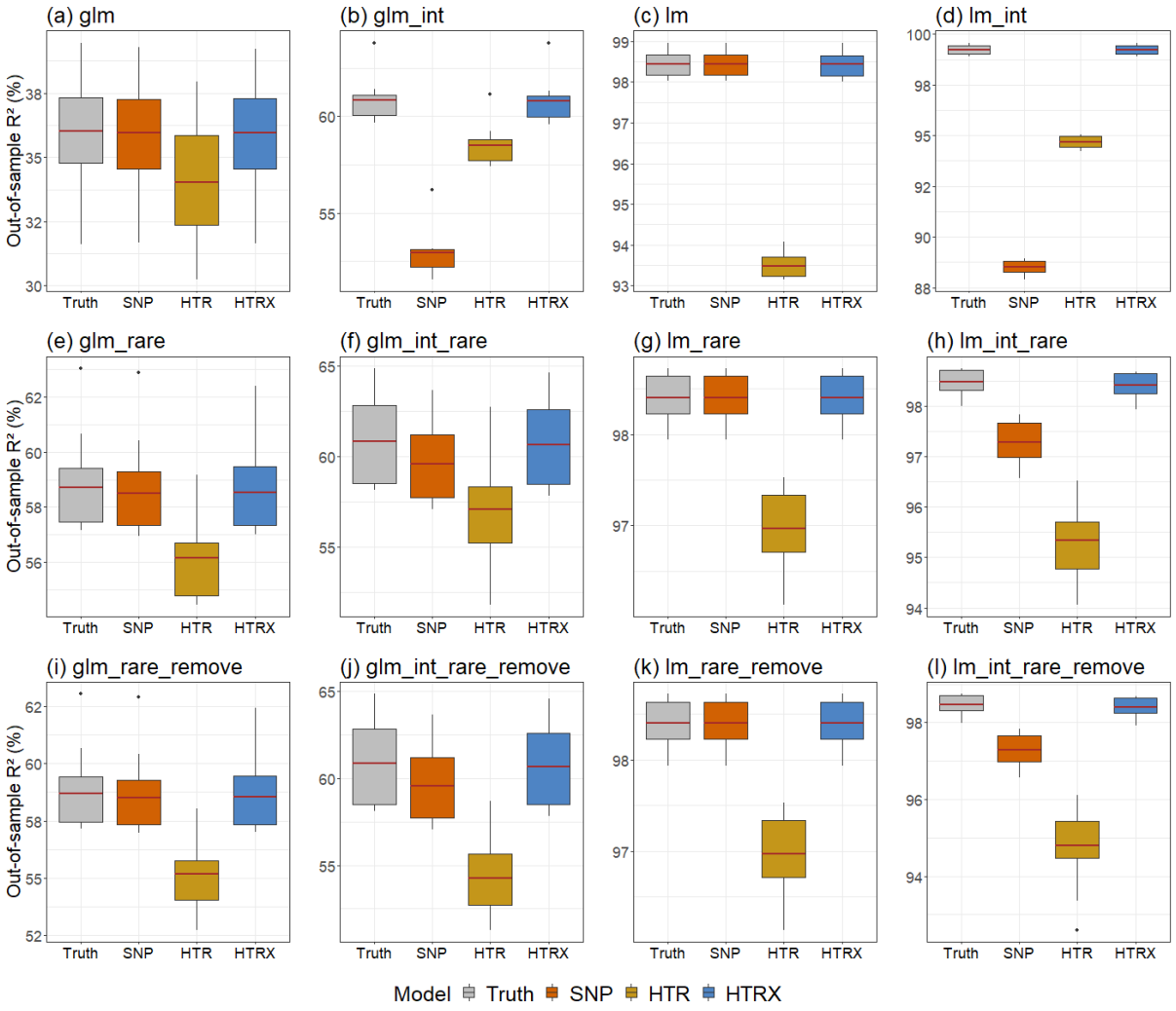


Supplementary Figure 4.2 | Comparison between MS-risk and local ancestry for 3 example SNPs.

In the HLA class II region, all SNPs share a pattern in which high Steppe ancestry is associated with high MS-risk. The risk decreases monotonically and is not present in the Steppe precursor populations (Hunter Gathers), but is with the admixed Bronze-age European populations (Steppe + Farmer).

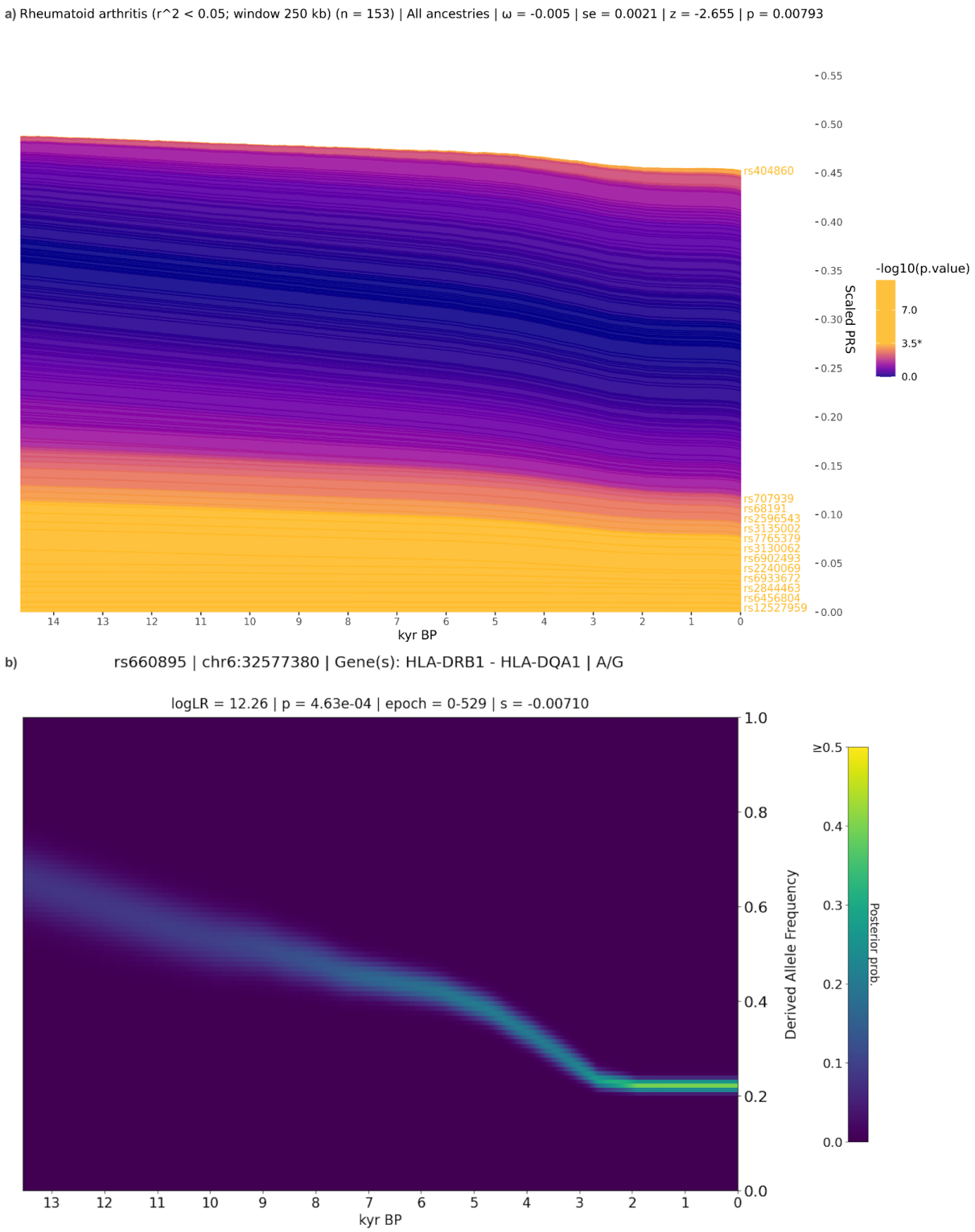


Supplementary Figure 4.3 | Decomposition of individuals ancestry at MS risk SNPs in terms of (left) the ancestry of those SNPs alone, or (right) the Weighted average prevalence of MS in each ancestry after "logit" transformation.



Supplementary Figure 4.4 | Simulation study with four SNPs showing the boxplots of out-of-sample variance (with the red line representing the average) explained by HTRX compared to GWAS, HTR and the true model.

The total variance explained by HTRX is the same as SNP and bigger than HTR when there are no interactions. When interaction (with subtitle "int") exists, HTRX significantly outperforms GWAS and HTR. In all situations, HTRX works similarly to the truth.

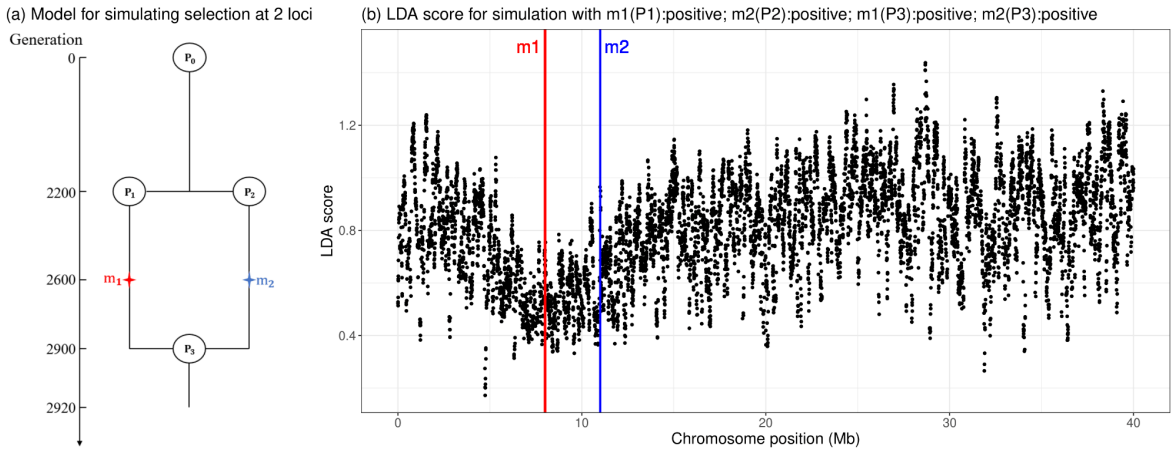


Supplementary Figure 5.1 | Evidence for selection on RA-associated SNPs.

a) Stacked line plot of the pan-ancestry PALM analysis for RA, showing the contribution of alleles to disease risk over time. Individual SNPs are stacked, with their trajectories polarised to show the frequency of the positive risk allele and weighted by their scaled effect size: when a given SNP bar becomes wider over time the risk allele has increased in frequency, and vice versa. SNPs are sorted by their marginal p-value and direction of effect, with selected SNPs that increase risk plotted on top.

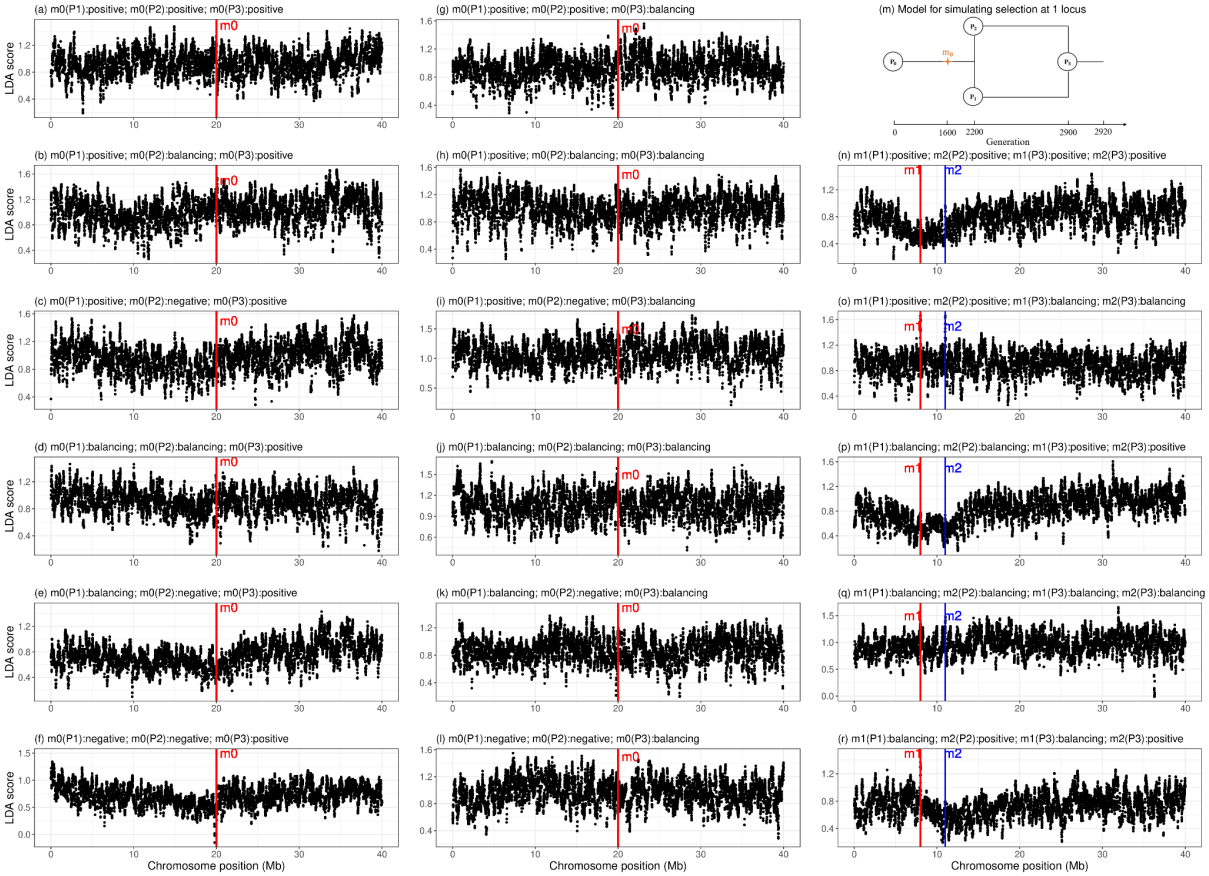
SNPs are also coloured by their marginal p-values, and significant SNPs are shown in yellow. The y-axis shows the scaled polygenic risk score (PRS), which ranges from 0 to 1, representing the maximum possible additive genetic risk in a population.

b) Posterior likelihood trajectory for rs660895, tagging HLA-DRB1*04:01, inferred by CLUES.



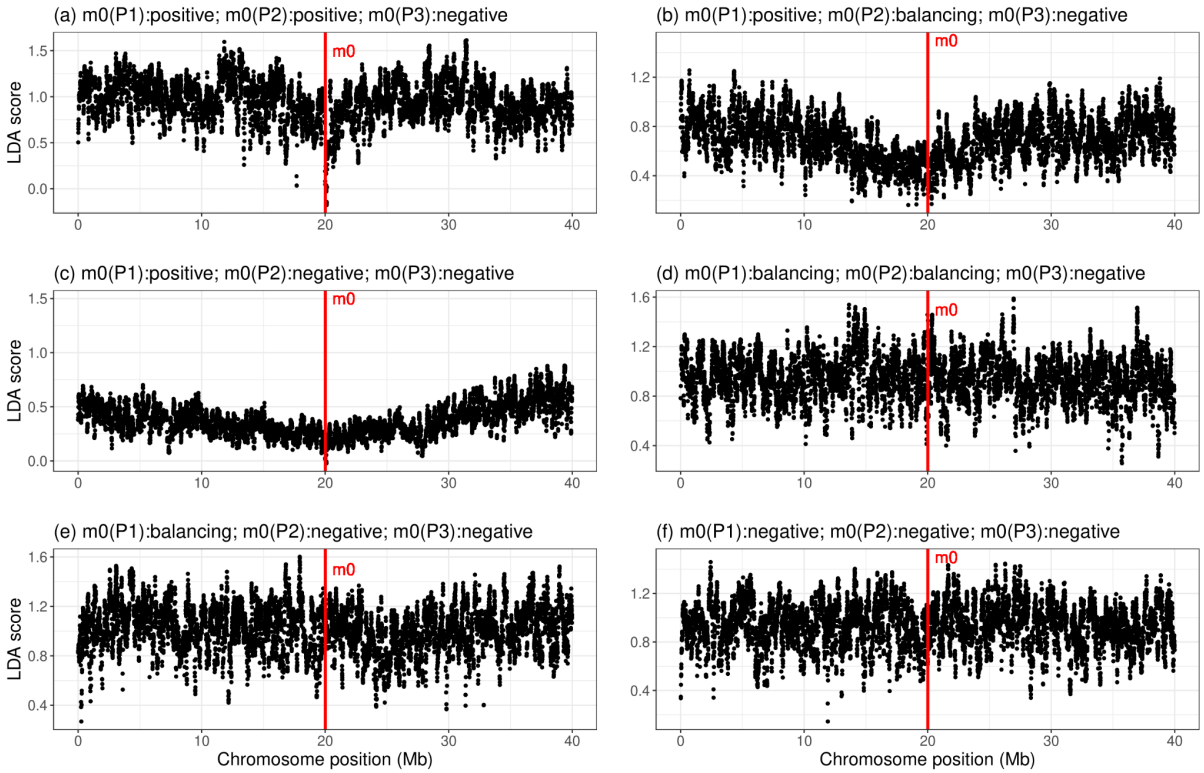
Supplementary Figure 6.1 | Simulating Low LDA score.

Left: A simulated history in which a single population splits into two (“Steppe” and “Farmer”) after 2200 generations and experiences positive selection on different loci (m_1 in P_1 and m_2 in P_2). After 2900 generations the populations merge (“Europeans”) but selection continues on *both* loci.



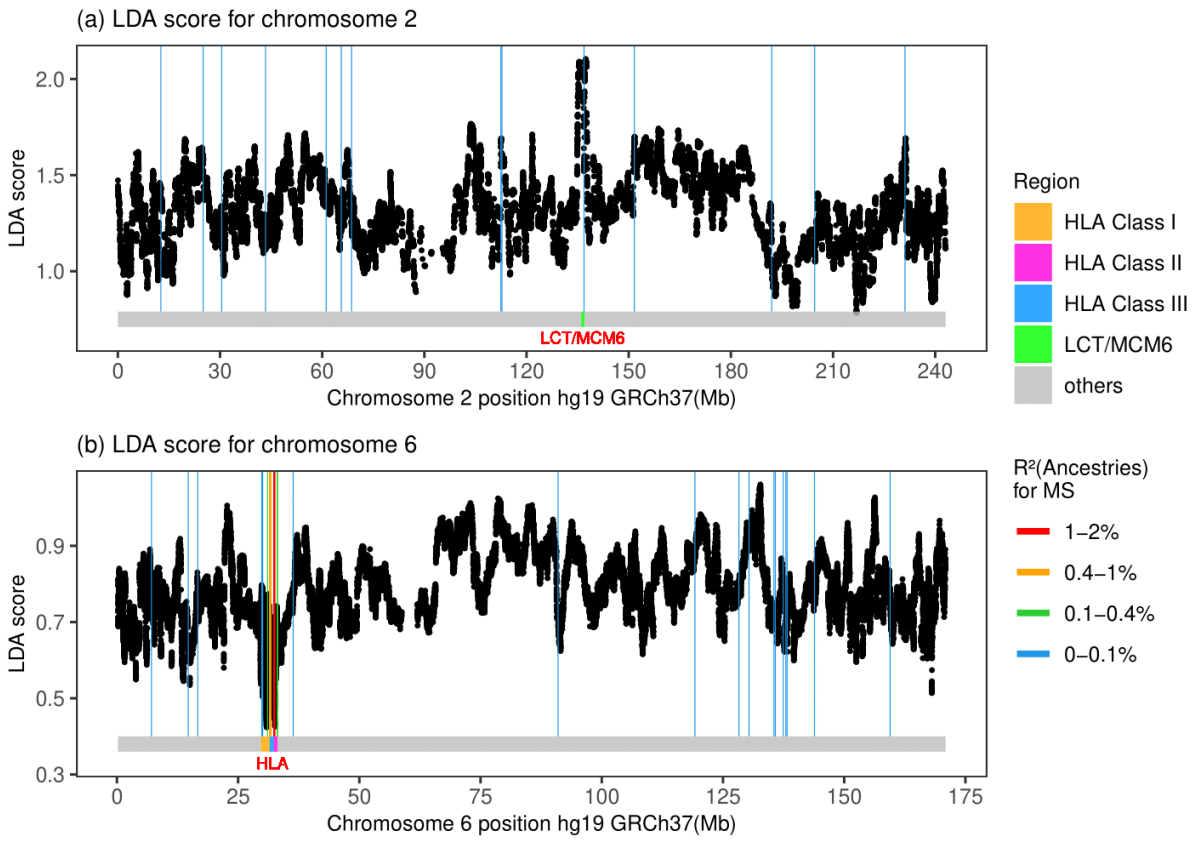
Supplementary Figure 6.2 | LDAS simulation with positive or balancing selection in the modern population.

The left two columns show simulations with a single variant satisfying the observed constraint that modern-day frequencies are not decreasing (i.e. not negative selection). The right column shows simulations with two variants, also obeying this constraint. The model for simulating 2 loci is the same as in Supplementary Figure 6.1, and that for 1 locus is in the top right of this plot (which differs only in the location of the selected variant in the separated populations).



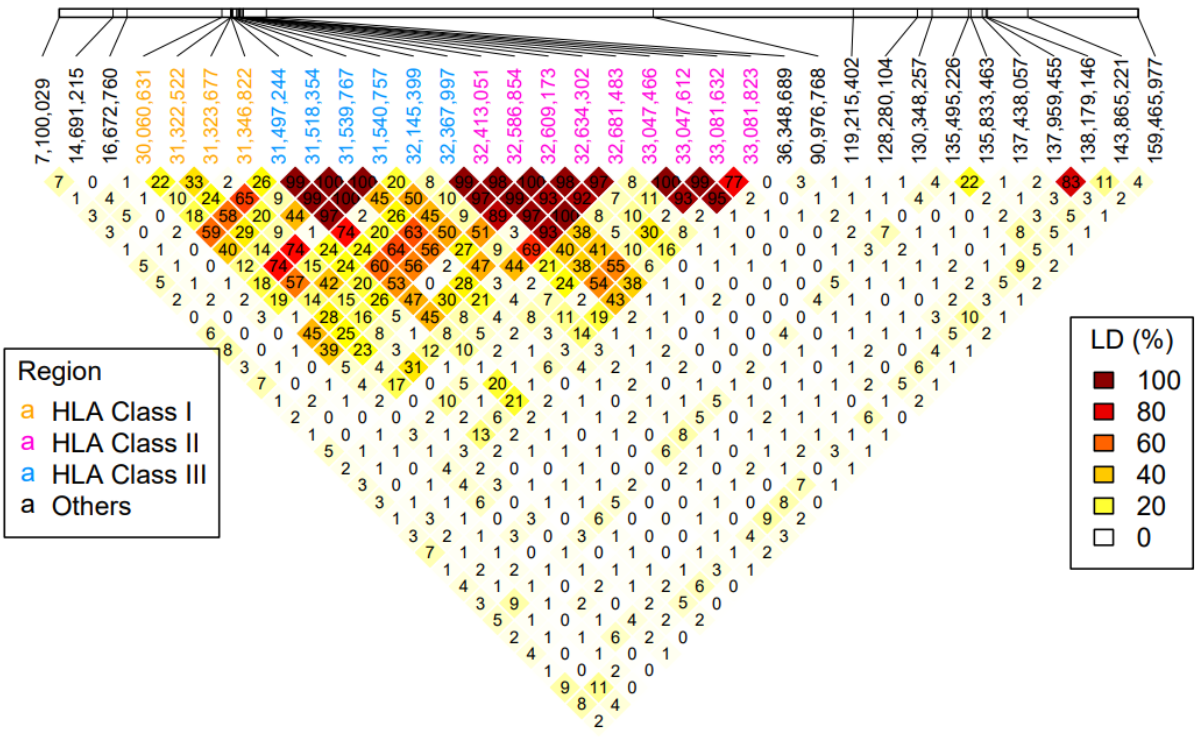
Supplementary Figure 6.3 | LDAS simulation with single locus negatively selected in the modern population.

In two cases this generates a low LDAS score, which requires recent negative selection (which is ruled out for HLA by the observed frequency trend). In this case, one ancestry dominates the region and recombination to the other conveys risk. The model used is in the top right of Supplementary Figure 6.2.

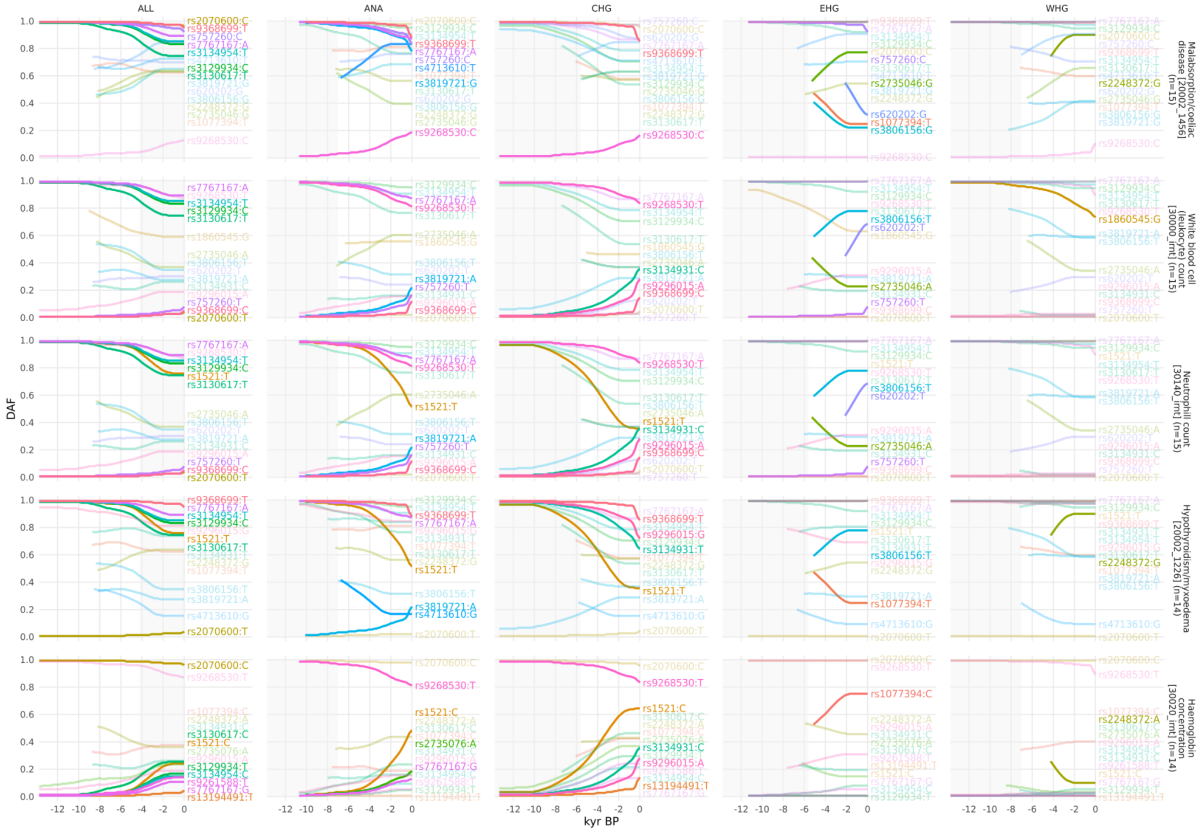


Supplementary Figure 6.4 | LDAS on chromosome 6 and 2.

LDA score is a) high in the LCT/MCM6 region while is b) low in the HLA region.

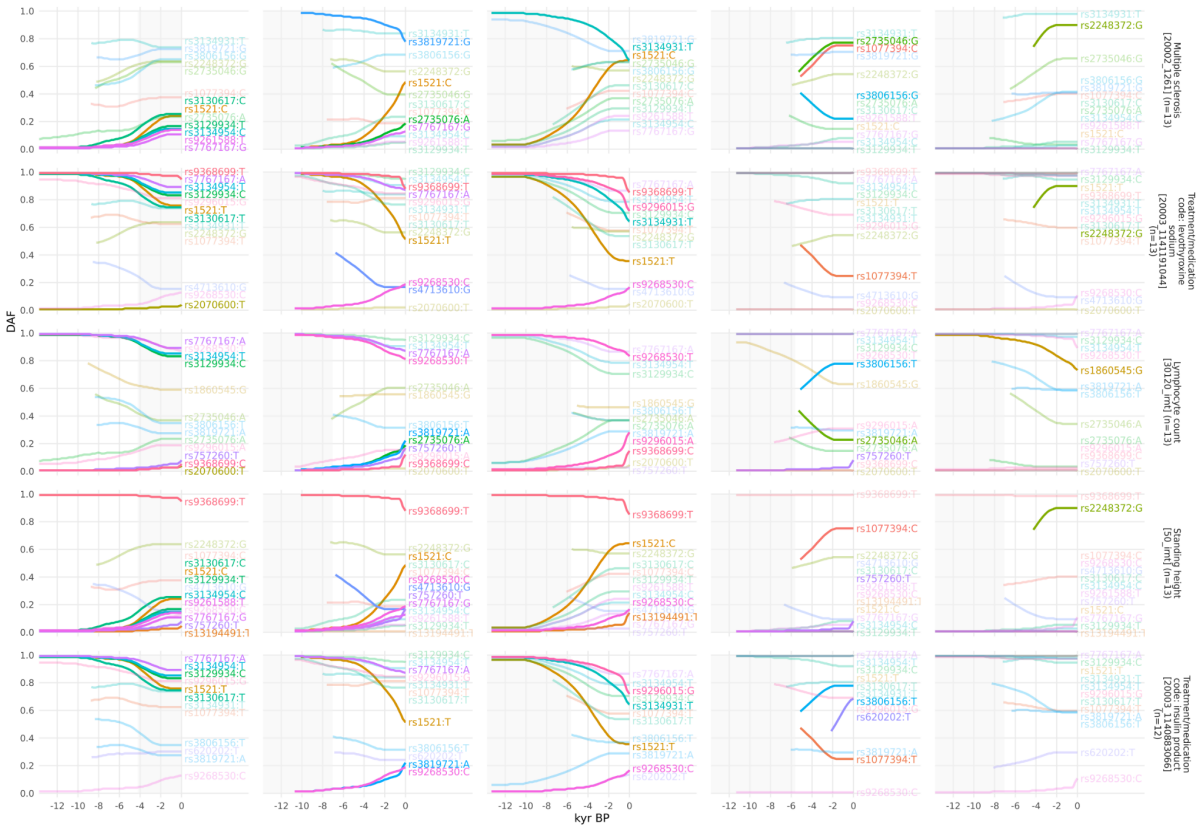


Supplementary Figure 6.5 | Pairwise Linkage Disequilibrium (LD) plot (measured by D') for all the MS-associated SNPs on chromosome 6.



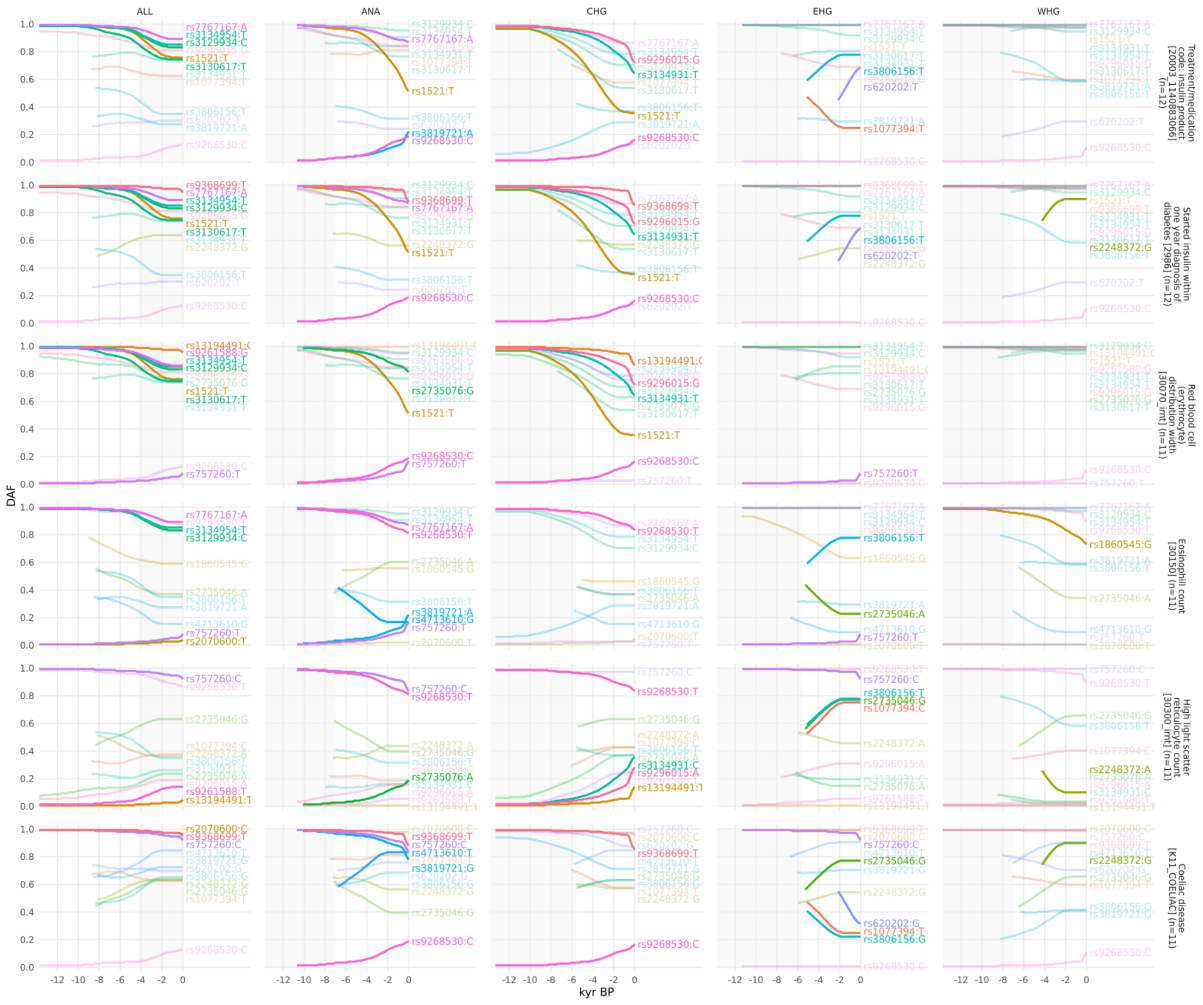
Supplementary Figure 7.1 | Allele frequency plots for positively selected MS-associated SNPs that are also associated with other phenotypes in the UK Biobank. Traits 1-5.

SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their effect on the marginal UK Biobank trait (i.e. showing the 'risk' allele). Phenotypes are ordered according to the number of common SNPs, non-significant SNPs are shown with partial transparency, portions of the trajectory with low posterior density are cropped off, and the background is shaded for the approximate time period in which the ancestry existed as an actual population.



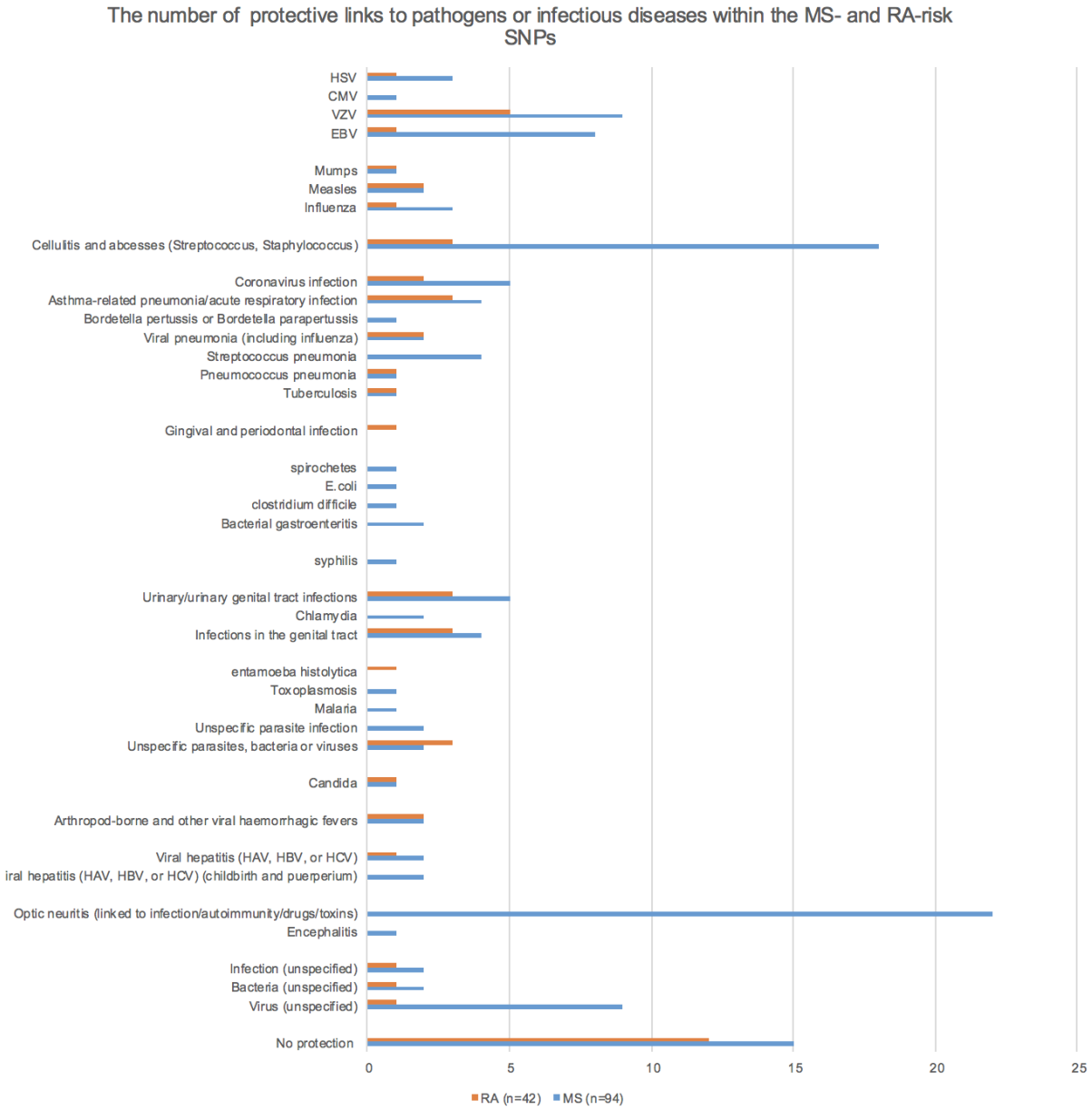
Supplementary Figure 7.2 | Allele frequency plots for positively selected MS-associated SNPs that are also associated with other phenotypes in the UK Biobank. Traits 6-10.

SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their effect on the marginal UK Biobank trait (i.e. showing the 'risk' allele). Phenotypes are ordered according to the number of common SNPs, non-significant SNPs are shown with partial transparency, portions of the trajectory with low posterior density are cropped off, and the background is shaded for the approximate time period in which the ancestry existed as an actual population. Note that many phenotypes are underpowered in the UKB GWAS, hence why MS appears as just the joint 7th in this list.



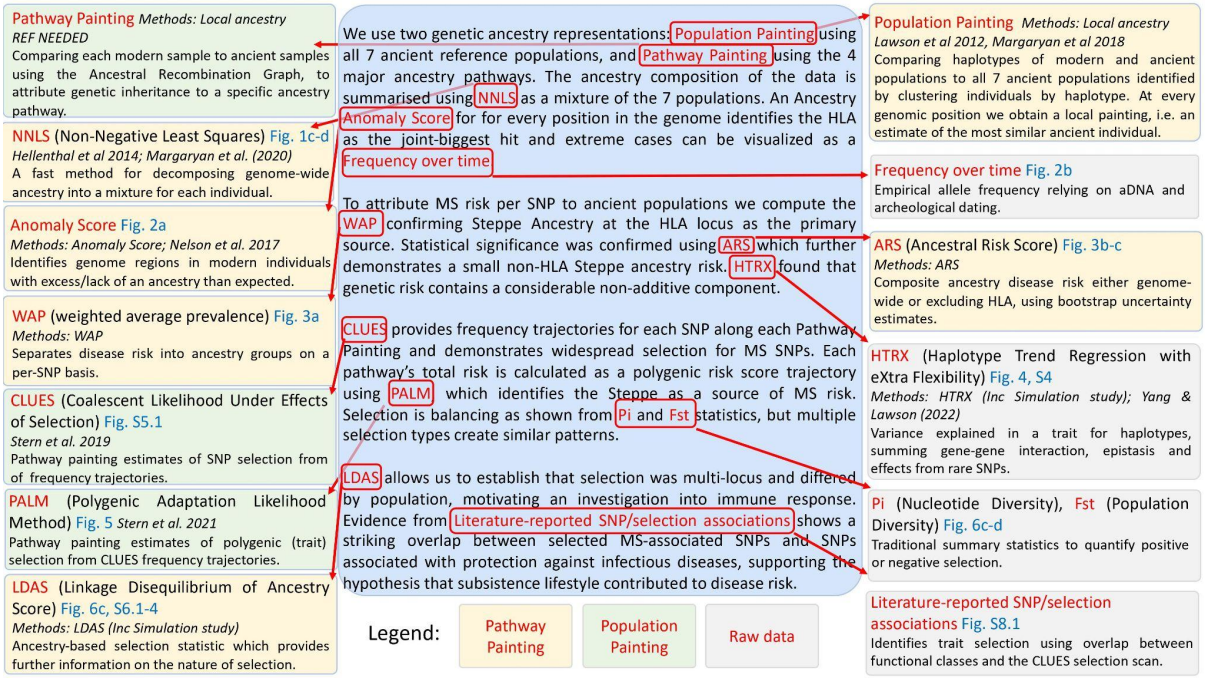
Supplementary Figure 7.3 | Allele frequency plots for positively selected MS-associated SNPs that are also associated with other phenotypes in the UK Biobank. Traits 11-15.

SNPs are shown with their maximum likelihood trajectories, and polarised by the direction of their effect on the marginal UK Biobank trait (i.e. showing the 'risk' allele). Phenotypes are ordered according to the number of common SNPs, non-significant SNPs are shown with partial transparency, portions of the trajectory with low posterior density are cropped off, and the background is shaded for the approximate time period in which the ancestry existed as an actual population.



Supplementary Figure 8.1 | The number of protective associations with pathogens or infectious diseases for the MA- and RA-associated selected SNPs.

The number of protective associations to specific pathogens and/or diseases associated with the MS- and RA-SNPs that showed statistically significant evidence for selection using CLUES. One SNP can have a link to more than one pathogen and/or disease (see ST13 and ST14 for details on each SNP). Fifteen and twelve SNPs had no detectable links to any pathogen or infectious disease in the MS and RA SNP sets, respectively.



Supplementary Figure 9.1 Methods map detailing datasets used, methods, and statistics.

A narrative of the evidence used is provided in the centre, with boxes on each side detailing the methods used. Boxes are coloured by the dataset used.

Methods

Data Generation

Overview

In order to examine variants associated with phenotypes backwards in time, we assembled a large ancient DNA dataset. Here we present new genomic data from 86 ancient individuals from Medieval and post-Medieval periods from Denmark (Supplementary Figure 1, Supplementary Note 1, ST1). The samples range in age from around the XIth to the XVIIIth century. We extracted ancient DNA from tooth cementum or petrous bone and shotgun sequenced the 86 genomes to a depth of genomic coverage ranging from 0.02 X to 1.6 X (mean = 0.39 X and median = 0.27 X). The genomes of the new 86 individuals were imputed using the 1,000 Genomes phased data as a reference panel by an imputation method designed for low coverage genomes (GLIMPSE, Rubinacci et al., 2021), and we also imputed 1,664 ancient genomes presented in the accompanying study 'Population Genomics of Stone Age Eurasia' (Allentoft et al., 2022). Depending on the specific data quality requirements for the downstream analyses, we filtered out samples with poor coverage, variant sites with low MAF and with low imputation quality (average genotype probability < 0.98). Our dataset of ancient individuals span approximately 15,000 years across Eurasia (Supplementary Figure 1).

Ancient data DNA extraction and library preparation

Laboratory work was conducted in the dedicated ancient DNA clean-room facilities at the Lundbeck Foundation GeoGenetics Centre (Globe Institute, University of Copenhagen). A total of 86 Medieval and post-Medieval human samples from Denmark (ST2) were processed using semi-automated procedures. Each sample was processed in parallel. For each extract non USER-treated and USER-treated (NEB) libraries were built (Meyer & Kircher, 2010). All libraries were sequenced on the NovaSeq6000 instrument at the GeoGenetics Sequencing Core, Copenhagen, using S4 200 cycles kits version 1.5. A more detailed description of DNA extraction and library preparation can be found in Supplementary Note 1.

Basic bioinformatics

The sequencing data was demultiplexed using the Illumina software BCL Convert (https://emea.support.illumina.com/sequencing/sequencing_software/bcl-convert.html, Illumina Inc.) . Adapter sequences were trimmed and overlapping reads were collapsed

using AdapterRemoval (2.2.4 (Schubert et al., 2016)). Single-end collapsed reads of at least 30bp and paired-end reads were mapped to the human reference genome build 37 using bwa (0.7.17 (Li & Durbin, 2009)) with seeding disabled to allow for higher sensitivity. Paired- and single-end reads for each library and lane were merged, and duplicates were marked using Picard MarkDuplicates (2.18.26, http://picard.sourceforge.net) with a pixel distance of 12000. Read depth and coverage were determined using samtools (1.10 (Li et al., 2009)) with the all sites used in the calculation (-a). Data was then merged to sample level and duplicates were marked again.

DNA authentication

In order to determine the authenticity of the ancient reads, post-mortem DNA damage patterns were quantified using mapDamage2.0 (Jónsson et al., 2013). Next, two different methods were used to estimate the levels of contamination. Firstly, we applied ContamMix in order to quantify the fraction of exogenous reads in the mitochondrial reads by comparing the mtDNA consensus genome to possible contaminant genomes (Fu et al., 2013). The consensus was constructed using an in-house perl script that used sites with at least 5x coverage, and bases were only called if observed in at least 70% of reads covering the site. Lastly, we applied ANGSD (0.931 (Korneliussen et al., 2014)) to estimate nuclear contamination by quantifying heterozygosity on the X chromosome in males. Both contamination estimates only used filtered reads with a base quality of ≥20 and mapping quality of ≥30.

Imputation

We combined the 86 newly sequenced Medieval and post-Medieval Danish individuals with 1,664 previously published ancient genomes (Allentoft et al., 2022). We then excluded individuals showing: contamination (more than 5%); low autosomal coverage (less than 0.1 X); low genome-wide average imputation genotype probability (less than 0.98), and we chose the best quality sample in a close relative pair (first or second degree relative). A total of 1,557 individuals passed all filters, and were used in downstream analyses. We restricted the analysis to SNPs with imputation INFO score ≥ 0.5 and MAF ≥ 0.05.

Kinship analysis and uniparental haplogroup inferences

READ (Monroy Kuhn et al., 2018) was used to detect the degree of relatedness between pairs of individuals.

The mtDNA haplogroups of the Medieval and post-Medieval individuals were assigned using HaploGrep2 (Weissensteiner et al., 2016). Y chromosome haplogroup assignment was inferred following the workflow already published (Scorrano et al., 2021). More details can be found in Supplementary Note 2.

Standard Population genetic analyses

The main population-genetics approach we base our inference on is Population-based painting (detailed below). However, to robustly understand population structure, we applied other standard techniques. Firstly, we used principal component analysis (PCA) (Supplementary Figure 1.1) to investigate the overall population structure of the dataset. We used plink (Purcell et al., 2007), excluding SNPs with minor allele frequency (MAF) < 0.05 in the imputed panel. Based on 1,210 ancient western Eurasia imputed genomes, the Medieval and post-Medieval samples cluster very close to each other, displaying a relatively low genetic variability and situated within the genetic variability observed in the post-Bronze Age western Eurasian populations.

We then used two additional standard methods to estimate ancestry components in our ancient samples. Firstly, we used model-based clustering (ADMIXTURE) (Shringarpure et al., 2016) (Supplementary Note 1, Figure S1.1) on a subset of 826,248 SNPs. Secondly, we used qpAdm (Patterson et al., 2012) (Supplementary Note 1 Figure S1.2 and Table S1.1) with a reference panel of three genetic ancestries (WHG, Neolithic Farmer, and Steppe) on the same 826,248 SNPs. We performed qpAdm applying the option “allsnps: YES” and a set of 7 outgroups was used as "right populations": Siberia_UpperPaleolithic_UstIshim, Siberia_UpperPaleolithic_Yana, Russia_UpperPaleolithic_Sunghir, Switzerland_Mesolithic, Iran_Neolithic, Siberia_Neolithic, USA_Beringia. We set a minimum threshold of 100,000 SNPs and only results with $p > 0.05$ only have been considered.

Population painting

Our main analysis uses chromosome painting (Lawson et al., 2012) with a panel of 6 ancient ancestries (as on the UK Biobank, see below). This allows fine-scale estimation of ancestry as a function of those populations. We ran chromosome painting on all ancient individuals not in the reference panel, using a reference panel of ancient donors grouped into populations to represent specific ancestries: Western Hunter-Gatherer (WHG), Eastern Hunter-Gatherer (EHG), Caucasus Hunter-Gatherer (CHG), Neolithic Farmer, Steppe, and African (method described in Allentoft et al., 2022 Supplementary Note 3h). Painting followed the pipeline of Margaryan et al. (2020) based on GLOBETROTTER (Hellenthal et

al. 2014), with admixture proportions estimated using Non-Negative Least squares (NNLS). NNLS explains the genome-wide haplotype matches of an individual as a mixture of the genome-wide haplotype-matches of the reference populations. This setup allows both the reference panel and any additional samples (i.e. modern) to be described using these 6 ancestries (Figure 1).

We then painted individuals born in Denmark of a typical ancestry (typical based on density-based clustering of the first 18 PCs, Allentoft et al. 2022). The reference panel used for chromosome painting was designed to capture the various components of European ancestry only, and so we urge caution in interpreting these results for non-European samples.

This dataset provides the opportunity to study the population history of Denmark from the Mesolithic to the post-Medieval period, covering around 10,000 years, which can be considered a typical Northern European population. Our results clearly demonstrate the impact of previously described demographic events, including the influx of Neolithic Farmer ancestry ~9,000 years ago and Steppe ancestry ~5,000 years ago (Allentoft et al., 2015; Haak et al., 2015). We highlight genetic continuity from the Bronze Age to the post-Medieval period (Supplementary Note 1 Figure S1.1), although qpAdm detected a small increase in the Neolithic Farmer component during the Viking Age (Supplementary Note 1 Figure S1.2 and Table S1.1), while the Medieval period marked a time of increased genetic diversity, likely reflecting increased mobility across Europe. This genetic continuity is further confirmed by the haplogroups identified in the uniparental genetic makers (Supplementary Note 2). Together, these results suggest that after the Bronze Age Steppe migration there was no other major gene flow into Denmark from populations with significantly different Neolithic and Bronze Age ancestry compositions, and therefore no changes in these ancestry components in the Danish population.

Local ancestry from Population painting

Chromosome Painting provides an estimate of the probability that an individual from each reference population is the closest match to the reference individual at every position in the genome. This provides our first estimate of local ancestry from Allentoft et al. (2022): the population of the first reference individual to coalesce with the target individual, as estimated by Chromopainter (Lawson et al., 2012). This was estimated for all “White British” individuals in the UK Biobank, using the population painting reference panel described above. We refer

to this henceforth as ‘local ancestry’, though note that the closest relative in the sample may not represent ancestry in the conventional sense.

Pathway painting

An alternative approach is to identify which of the four major ancestry pathways (Farmer, CHG, EHG, WHG) each position in the genome best matches to. This has the advantage of not forcing haplotypes to choose between “Steppe” ancestry and its ancestors, but the disadvantage of being more complex to interpret. To do this, we modelled ancestry path labels in GBR, FIN and TSI 1000G populations (The 1000 Genomes Project Consortium et al., 2015) and 1015 ancient genomes generated using a neural network to assign ancestry paths based on a sample’s nearest neighbours at the first five informative nodes of a marginal tree sequence, where an informative node is defined as one which has at least one leaf from the reference set of ancient samples described above (Allentoft et al., 2022 Supplementary Note S3i). We refer to this henceforth as ‘ancestry path labels’.

SNP associations

We aimed to generate SNP associations from previous studies for each phenotype in a consistent approach. To generate a list of SNPs associated with multiple sclerosis (MS), rheumatoid arthritis (RA) and celiac disease (CD), we used two approaches: in the first, we downloaded fine-mapped SNPs from previous association studies. For each fine-mapped SNP, if the SNP did not have an ancestry path label, we found the SNP in highest LD that did, with a minimum threshold of $r^2 \geq 0.7$ in the GBR, FIN and TSI 1000G populations using LDLinkR (Myers et al., 2020). The final SNPs used for each phenotype can be found in ST4 (MS), ST5 (RA), and ST6 (CD).

For MS, we used data from IMSGC, 2019. For non-MHC SNPs, we used the ‘discovery’ SNPs with P(joined) and OR(joined) generated in the replication phase. For MHC variants, we searched the literature for the reported HLA alleles and amino-acid polymorphisms (ST3). In total, we generated 205 SNPs which were either fine-mapped or in high LD with a fine-mapped SNP (15 MHC, 190 non-MHC).

For RA, we downloaded 57 genome-wide significant non-MHC SNPs for seropositive RA in Europeans (Ishigaki et al., 2021). We retrieved MHC associations separately (Alekseyenko et al., 2011, with associated ORs and p-values from Raychaudhuri et al., 2012). In total, we generated 51 SNPs which were either fine-mapped or in high LD with a fine-mapped SNP (3 MHC, 48 non-MHC).

For CD, we retrieved non-MHC SNPs from Trynka et al. (2011). We used MHC SNPs from Monsuur et al. (2008), with associated ORs and p-values from Gutierrez-Achury et al. (2015). In total, this generated 32 SNPs which were either fine-mapped or in high LD with a fine-mapped SNP (3 MHC, 29 non-MHC).

Secondly, because we could not always find tag SNPs for fine-mapped SNPs that were present in our ancestry path labels dataset, we found that we were losing significant signal from the HLA, therefore we generated a second set of SNP associations. We downloaded full summary statistics for each disease (MS: IMSGC, 2019; RA: Okada et al., 2014, CD: http://www.nealelab.is/uk-biobank/), restricted to sites present in the ancestry path labels dataset, and ran Plink's (PLINK v1.90b4.4, Chang et al., 2015) clump method (parameters: --clump-p1 5e-8 --clump-r2 0.05 --clump-kb 250 as in Ju and Mathieson (2021) using LD in the GBR, FIN and TSI 1000G populations (The 1000 Genomes Project Consortium et al., 2015) to extract genome-wide significant independent SNPs.

In the main text we report results for the first set of SNPs ('fine-mapped') for analyses involving local ancestry in modern data, and the second set of SNPs ('pruned') for analyses involving polygenic measures of selection (CLUES/PALM).

Anomaly Score: Regions of Unusual Ancestry

To assess which regions of ancestry were unusual, we converted the ancestry estimates to a Z-score standardized to the genome wide mean and with predictable variance. Specifically, we let $A(i,j,j)$ denote the probability of the kth ancestry ($k = 1,\ldots,K$) at the jth SNP ($j = 1,\ldots,J$) of a chromosome for the ith individual ($i = 1,\ldots,N$). We then computed the mean painting for each SNP, $A(j,k) = \frac{1}{N}\sum_{i=1}^{N} A(i,j,k)$. From this we estimated a scale parameter μ_k and deviation parameter σ_k using a block-median approach. Specifically we partitioned the genome into 0.5Mb regions, and within each, computed the mean and standard deviation of the ancestry. The parameter estimates are then the median values over the whole genome. We then computed an anomaly score for each SNP for each ancestry $Z(j,k) = (A(j,k) - \mu_k)/\sigma_k$. This is the Normal-distribution approximation to the Poisson-Binomial score for excess ancestry, for which a detailed simulation study is presented in Nelson et al. (2017).

To arrive at an anomaly score for each SNP aggregated over all ancestries, we also had to account for correlations in the ancestry paintings. Instead of scaling each ancestry deviation $A^*(j,k) = A(j,k) - \mu_k$ by its standard deviation, we instead "whitened" them, i.e. rotated the data to have an independent signal. Let $C = A^{*T} A^*$ by a $K \times K$ covariance matrix, and let $C^{-1} = UDV^T$ be the Singular Value Decomposition. Then $W = UD^{1/2}$ is the whitening matrix from which $Z = A^* W$ are normally distributed with covariance matrix diag(1) under the null that A^* is normally distributed with mean 0 and unknown covariance Σ. The "ancestry anomaly score" test statistic for each SNP is $t(j) = \sum_{k=1}^{K} Z(j,k)^2$, which is Chi-squared distributed with K degrees of freedom under the null, and we reported p-values from this.

To test for gene enrichment we formed a list of all SNPs reaching genome-wide significance ($p < 5^{-8}$) and using the R package *gprofiler2* (Kolberg et al., 2020) converted these to a unique list of genes. We then used *gost* to perform an enrichment test for each GO term, for which we used default p-value correction via the *g:Profiler SCS* method. This is an empirical correction based on performing random lookups of the same number of genes under the null, to control the error rate and ensure that 95% of reported categories (at p=0.05) are correct.

Allele Frequency Plots Over Time

To investigate how effect allele frequencies have changed over time, we extracted high effect alleles for each phenotype from the ancient data. We excluded all non-Eurasian samples, grouped them by 'groupLabel', excluded any group with fewer than 4 samples, and coloured points by ancestry proportion according to genome-wide NNLS based on chromosome painting (above).

Weighted Average Prevalence

In order to understand whether risk-conferring haplotypes evolved in the Steppe population, or in a pre- or post-dating population in which Steppe ancestry is high, we developed a statistic that could account for the origin of risk to be identified with multiple ancestry groups, which do not have to be the same set for each SNP.

We first applied k-means clustering to the dosage of each ancestry for each associated SNP an

d investigated the dosage distribution of clusters with significantly higher MS prevalence. For the target SNPs, the Elbow method (Thorndike, 1953) suggested selecting around 5-7 clusters, of which we chose 6. After performing the k-means cluster analysis, we calculated the average probability for each ancestry for case individuals. Furthermore, we calculated the prevalence of MS in each cluster, and performed a one-sample t-test to investigate whether it differs from the overall MS prevalence (0.487%). This tests whether any particular combinations of ancestry are associated with the phenotype at a SNP. Clusters with high MS risk-ratios have high Steppe components (Supplementary Figure 4.2), leading to the conclusion that Steppe ancestry alone is driving this signal.

We can then compute the Weighted Average Prevalence (WAP) which summarises these results into the ancestries. For the jth SNP, let $P_{jkm} = n_{jm}\bar{P}_{jkm}$ denote the sum of the kth ancestry probabilities of all the individuals in the mth cluster ($k, m = 1,\dots, 6$), where n_{jm} is the cluster size of the mth cluster. Let π_{jm} denote the prevalence of MS in the mth cluster, the weighted average prevalence for the kth ancestry is defined as:

$$\bar{\pi}_{jk} = \frac{P_{jkm}\pi_{jm}}{\sum_{m=1}^{6} P_{jkm}},$$

where P_{jkm} is defined as the weight for each cluster.

The standard deviation of $\bar{\pi}_{jk}$ is computed as $sd(\bar{\pi}_{jk}) = \sqrt{\sum_{m=1}^{6} w_{jkm}{}^2 \sigma_m{}^2}$, where $w_{jkm} = \frac{P_{jkm}}{\sum_{m=1}^{6} P_{jkm}}$, $\sigma_m = \frac{s(y_{jm})}{\sqrt{n_{jm}}}$ and $s(y_{jm})$ is the standard deviation of the outcome for the individuals in the mth cluster. We also test the hypothesis that H_0: $\bar{\pi}_{jk} = \bar{\pi}$ against H_1: $\bar{\pi}_{jk} \neq \bar{\pi}$, and compute the p-value as $p_{jk} = 2(1 - \phi(\frac{|\bar{\pi} - \bar{\pi}_{jk}|}{sd(\bar{\pi}_{jk})}))$.

For each ancestry, WAP measures the association of that ancestry with MS risk across all clusters. To make a clear comparison, we calculated the risk ratio (compared to the overall MS prevalence) for each ancestry at each SNP, and assigned a mean and confidence interval for the risk ratios of each ancestry at each chromosome (Figure3, Supplementary Figure 3.1 and 3.2).

PCA/UMAP Of WAP/Average Dosage

We performed principal component analysis (PCA) on the average ancestry probability and WAP at each MS-associated SNP (Supplementary Figure 4.3). The former shows that all of the HLA SNPs except three from HLA class II and III have much larger Outgroup components compared with the others. The latter analysis indicates a strong association between Steppe and MS risk. Also, Outgroup ancestry at rs10914539 from chromosome 1 exceptionally reduces the incidence of MS, while Outgroup ancestry at rs771767 (chromosome 3) and rs137956 (chromosome 22) significantly boosts MS risk.

Ancestral Risk Scores

Because panels of ancient individuals are small and geographically biased, allele frequency estimates based directly on aDNA genotype calls have low confidence (Dehasque et al., 2020). Equally, selection or drift (e.g. from population bottlenecks) mean that the allele frequency in an ancient population does not necessarily reflect the proportion of effect alleles that that ancestry eventually contributed to a modern population. Therefore, a better estimate of an ancestral contribution is to generate allele frequencies based on local ancestry: if a haplotype is under-sampled in the ancient data or undergoes subsequent positive selection, this will be reflected in an allele frequency that is higher in the estimate based on the painting than one based on the ancient data. We refer to these frequencies as “painting frequencies”.

This approach was used to estimate ancestral contributions to a range of phenotypes in Irving-Pease et al. (2022), re-capitulating already known contributions such as height, hair colour and eye colour.

All code for implementing these analyses can be found at https://github.com/will-camb/ms_paper**.**

Imputation of local ancestry

Because not all SNPs in the GWAS data were painted, for each variant in each GWAS dataset we imputed the local ancestry by taking the average of the painting values of the SNPs on either side, weighted by their physical distance (impute_ancestry.py).

Ancestral risk score

Following methods developed in Irving-Pease et al. (2022), we calculated the effect allele painting frequency for a given ancestry $f_{\{anc,i\}}$ for SNP i using the formula:

$$f_{\{anc,i\}} = \frac{\sum_{j}^{M_{effect}} Painting\ certainty_{\{j,i,anc\}}}{\sum_{j}^{M_{alt}} Painting\ certainty_{\{j,i,\ anc\}} + \sum_{j}^{M_{effect}} Painting\ certainty_{\{j,i,\ anc\}}},$$

where there are M_{effect} individuals homozygous for the effect allele, M_{alt} individuals homozygous for the other allele, and $\sum_{j}^{M_{effect}} Painting\ certainty_{\{j,i,anc\}}$ is the sum of the painting probabilities for that ancestry anc in individuals homozygous for the effect allele at SNP i. This can be interpreted as an estimate of an ancestral contribution to effect allele frequency in a modern population. Per-SNP painting frequencies can be found in ST4, ST5, and ST6.

To calculate the ancestral risk score (ARS**)** we summed over all I pruned SNPs in an additive model:

$$ARS_{anc} = \sum_{i}^{I} f_{\{anc,i\}} * \ beta_i.$$

I then ran a transformation as in Berg & Coop (2014), centering results around the ancestral mean (i.e. all ancestries) and reporting as a Z-score. To obtain 95% confidence intervals, we ran an accelerated bootstrap over loci, which accounts for the skew of data to better estimate confidence intervals (Frangos & Schucany, 1990).

GWAS of Ancestry and Genotypes

The total variance of a trait explained by genotypes (SNP values), Ancestry, and haplotypes (described below) is a measure of how well each captures the causal factors driving that trait. We therefore computed the variance explained for each data type in a “head-to-head” comparison, either at specific SNPs or SNP sets. In this section we describe the model and covariates accounted for.

We used the UK Biobank to fit GWAS models for local ancestry values and genotype values separately, using only SNPs known to be associated with the phenotype (‘fine-mapped’ SNPs). We used the following phenotype codes for each phenotype: MS: Data-Field 131043; RA: Data-Field 131849 (seropositive); CD: Data-Field 21068.

Let Y_i denote the phenotype status for the *i*th individual ($i = 1,\ldots, 399998$), which takes value 1 for a case and 0 for control, and let $\pi_i = Pr(Y_i = 1)$ denote the probability that this individual has the event. Let X_{ijk} denote the kth ancestry probability ($k = 1,\ldots, K$) for the jth SNP ($j = 1,\ldots, 205$) of the ith individual. C_{ic} is the cth predictor ($c = 1,\ldots, N_c$) for the ith

individual. We used the following logistic regression model for GWAS, which assumes the effects of alleles are additive:

$$Y_i \sim Bin(1, \pi_i);\ log(\frac{\pi_i}{1-\pi_i}) = \sum_{k=1}^{K} \beta_{jk} X_{ijk} + \sum_{c=1}^{N_c} \gamma_c C_{ic}.$$

We used N_c=20 predictors in the GWAS models, including sex, age and the first 18 PCs, which are sufficient to capture most of the population structure in the UK Biobank (Sarmanova et al., 2020).

First, we built the model with $K = 1$. By using only one ancestry probability in each model, we aimed to find the statistical significance of each SNP under each ancestry. Then, we built the model with $K = 5$, i.e. using all 6 local ancestry probabilities which sum to 1. We calculated the variance explained by each SNP by summing up the variance explained by X_{ijk}(k=1,...,5).

We considered fitting the multivariate models by using all the SNPs as covariates. However, the dataset only contains 1,982 cases. Even though only one ancestry is included, the multivariate model contains 191 predictors, which could result in overfitting problems. Therefore, the GWAS models are preferred over multivariate models.

We also fitted a logistic regression model for GWAS using the genotype data as follows:

$$Y_i \sim Bin(1, \pi_i);\ log(\frac{\pi_i}{1-\pi_i}) = \beta_j X_{ij} + \sum_{c=1}^{N_c} \gamma_c C_{ic},$$

where $X_{ij} \in \{0, 1, 2\}$ denotes the number of copies of the reference allele of the jth SNP ($j = 1,..., 205$) that the ith individual has, and C_{ic} ($c = 1,..., N_c$) denotes the covariates including age, sex and first 18 PCs for the ith individual, where N_c=20. Due to the UK Biobank being underpowered compared to the Case-Control study from which these SNPs were found, the only statistically significant (at $p < 10^{-5}$) association is in the HLA class II tagging HLA-DRB1*15:01.

GWAS comparison for trait-associated SNPs

In this section we describe how we moved from associations between observations on SNPs (either genotype values or ancestry) and a trait, to total variance explained.

We compared the variance explained by SNPs from the GWAS model using the painting data (all 6 local ancestry probabilities) with that from GWAS model using the genotype data. McFadden's pseudo R squared measure (McFadden et al., 1973) is widely used for estimating the variance explained by the logistic regression models. McFadden's pseudo R squared is defined as

$$R^2 = 1 - \frac{ln(L_M)}{lm(L_0)},$$

where L_M and 0 are the likelihoods for the fitted and the null model, respectively. Taking overfitting into account, we propose the adjusted McFadden's pseudo R squared by penalizing the number of predictors:

$$Adjusted\ R^2 = 1 - \frac{ln(L_M)/(N-k)}{ln(L_0)/(N-1)},$$

where N is the sample size and k is the number of predictors.

Specifically, $R^2(SNPs)$ is calculated as the extra variance in addition to sex, age and 18 PCs that can be explained by SNPs:

$R^2(SNPs) = R^2(sex + age + 18\,PCs + SNPs) - R^2(sex + age + 18\,PCs)$.

Notably, two SNPs stand out for explaining much larger variance than others when fitting the GWAS model using the genotype data, but overall more SNPs from GWAS painting explain more than 0.1% variance, which indicates the painting data are probably more efficient for estimating the effect sizes of SNPs and detecting significant SNPs. Also, some SNPs from GWAS models using painting data explain almost the same amount of variance, suggesting that these SNPs consist of very similar ancestries.

Quantifying selection via historical allele frequencies from Pathway Painting

The historical trajectory of SNP frequencies is a strong signal of selection when ancient DNA data are available. This is the main purpose of our Pathway Painting method, and can be used to infer selection at individual loci, and combined into a polygenic score by analysing sets of SNPs associated with a trait.

Firstly, we inferred allele frequency trajectories and selection coefficients for a set of LD-puned genome-wide significant trait associated variants using a modified version of the software CLUES (Coalescent Likelihood Under Effects of Selection) (Stern et al., 2019) To account for population structure in our samples, we applied a novel chromosome painting

technique based on inference of a sample's nearest neighbours in the marginal trees of an ARG that contains labelled individuals (Irving-Pease et al., 2022). We ran CLUES using a time-series of imputed aDNA genotype probabilities obtained from 1,015 ancient West Eurasian samples that passed all quality control filters. We produced four additional models for each trait associated variant, by conditioning the analysis on one of the four ancestral path labels from our chromosome painting model: either Western Hunter-Gatherers (WHG), Eastern Hunter-Gatherers (EHG), Caucasus Hunter-Gatherers (CHG), or Anatolian farmers (ANA).

Secondly, were able to infer polygenic selection gradients (ω) and p-values for each trait, i.e. of MS, CD and RA, in all ancestral paths, using the software PALM (Polygenic Adaptation Likelihood Method) (Stern et al., 2021). Full methods and results can be found in Supplementary Note 6.

Linkage Disequilibrium of Ancestry (LDA) and LDA Score (LDAS)

In population genetics, linkage disequilibrium (LD) is defined as the non-random association of alleles at different loci in a given population (Slatkin, 2008). Just like the values of the genotype, ancestries can be correlated along the genome, and further, deviations from the expected length distribution for a particular ancestry is a signal of selection, dated by the affected ancestry. We propose an ancestry linkage disequilibrium (LDA) approach to measure the association of ancestries between SNPs, and an LDA Score (LDAS) to quantify deviations from the null hypothesis that ancestry is inherited at random across loci.

LDA is defined in terms of local ancestry. Let $A(i, j, k)$ denote the probability of the kth ancestry ($k = 1,\ldots, K$) at the jth SNP ($j = 1,\ldots, J$) of a chromosome for the ith individual ($i = 1,\ldots, N$).

We define the distance between SNP l and m as the average L_2 norm between ancestries at those SNPs. Specifically we compute the L_2 norm for the ith genome as

$$D_i(l, m) = ||A(i, l, \cdot) - A(i, m, \cdot)||_2 = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(A(i, l, k) - A(i, m, k))^2}.$$

Then we compute the distance between SNP l and m by averaging $D_i(l, m)$:

$$D(l, m) = \frac{1}{N}\sum_{i=1}^{N} D_i(l, m).$$

We define $D^*(l,m)$ as the theoretical distance between SNP l and m if there were no linkage disequilibrium of ancestry (LDA) between them. $D^*(l,m)$ is estimated by

$$D^*(l,m) \approx \frac{1}{N}\sum_{i=1}^{N} ||A(i^*,l,\cdot) - A(i,m,\cdot)||_2,$$

where $i^* \in \{1,\ldots,N\}$ are re-sampled without replacement at SNP l. Using the empirical distribution of ancestry probabilities accounts for variability in both the average ancestry and its distribution across SNPs. Ancestry assignment can be very precise in regions of the genome where our reference panel matches our data, and uncertain in others where we only have distant relatives of the underlying populations.

The LDA between SNP l and m is a similarity, defined in terms of the negative distance $-D(l,m)$ normalized by the expected value $D^*(l,m)$ under no LD, as:

$$LDA(l,m) = \frac{D^*(l,m) - D(l,m)}{D^*(l,m)}.$$

LDA therefore takes an expected value 0 when haplotypes are randomly assigned at different SNPs, and positive values when the ancestries of haplotypes are correlated.

LDA is a pairwise quantity. To arrive at a per-SNP property, we define the LDA score (LDAS) of SNP j as the total LDA of this SNP with the rest of the genome, i.e. the integral of the LDA for that SNP. Because this quantity decreases to zero as we move away from the target SNP, this is in practice computed within an XcM-window (we use $X = 5$ as LDA is approximately zero outside this region in our data) on both sides of the SNP. Note that we measure this quantity in terms of the genetic distance, and therefore LDAS is measuring the length of ancestry-specific haplotypes compared to individual-level recombination rates.

As a technical note, when the SNPs approach either end of the chromosome, they no longer have a complete window, which results in a smaller LDAS. This would be appropriate for measuring total ancestry correlations, but to make LDAS useful for detecting anomalous SNPs, we use the LDAS of the symmetric side of the SNP to estimate the LDAS within the non-existent window.

$$LDAS(j;X) = \begin{cases} \int_{gd(j)-X}^{gd(j)+X} LDA(j,l)\,dgd & \text{if } X \le gd(j) \le tg - X, \\ \int_{0}^{gd(j)+X} LDA(j,l)\,dgd + \int_{2gd(j)}^{gd(j)+X} LDA(j,l)\,dgd & \text{if } gd(j) < X, \\ \int_{gd(j)-X}^{tg} LDA(j,l)\,dgd + \int_{gd(j)-X}^{2gd(j)-tg} LDA(j,l)\,dgd & \text{if } gd(j) > tg - X. \end{cases}$$

where $gd(l)$ is the genetic distance (i.e. position in cM) of SNP l, and tg is the total genetic distance of a chromosome. We also assume the LDA on either end of the chromosome equals the LDA of the SNP closest to the end: $LDA(j, gd = 0) = LDA(j, l_{min(gd)})$ and $LDA(j, gd = td) = LDA(j, l_{max(gd)})$, where gd is the genetic distance, $l_{min(gd)}$ and $l_{max(gd)}$ are the indexes of the SNP with the smallest and largest genetic distance, respectively.

The integral $\int_{gd(j)-X}^{gd(j)+X} LDA(j,l)dgd$ is computed assuming linear interpolation of the LDA score between adjacent SNPs.

LDA thus quantifies the correlations between the ancestry of two SNPs, measuring the proportion of individuals who have experienced a recombination leading to a change in ancestry, relative to the genome-wide baseline. The LDA score is the total amount of genome in LDA with each SNP (measured in recombination map distance).

Simulation study for LDA and LDAS

An ancient population P_0 evolved for 2200 generations before splitting into two sub-populations P_1 (Steppe) and P_2 (Farmer). After evolving 400 generations, we added mutation "m_1" and "m_2" at the different locus in P_1 and P_2. Both added mutations were then positively selected in the following 300 generations, after which they merged to P_3, where both added mutations experienced strong positive selection for 20 generations. Finally, we sampled 1000 individuals from P_3 to compute their ancestry proportions of P_1 and P_2 using the "chromosome painting" technique, and calculated the LDA score of the simulated chromosome positions.

The above describes the simulation in Supplementary Figure 6.1.

We investigated balancing selection at 2 loci as well. The balancing selection in P_1 and P_2 ensured the mutated allele reaches around 50% frequency, while positive selection made

the mutated allele become almost the only allele. In P_3, if m_1 or m_2 was positively selected, its frequency reached more than 80% regardless of whether the allele experienced balancing or positive selection in P_1 or P_2, because we set a strong positive selection. If m_1 or m_2 was balancing selected in P_3, its frequency slightly increased, e.g. if m_1 underwent balancing selection in P_1, it had 25% frequency when P_3 was created, and the frequency reached around 37.5% after 20 generations of balancing selection in P_3.

The results (Supplementary Figure 6.2) show that positive selection in P_3 resulted in low LDA scores around the selected locus, if this allele was not uncommon (i.e. had 50% (balancing selection) or 100% frequency (positive selection) in subpopulation P_1 or P_2). Note that the balancing selection in P_1 or P_2 worked the same as “weak positive selection”, because m_1 and m_2 were rare when they first occurred, which were positively selected until 50% frequency.

We also performed simulations for selection at a single locus (Supplementary Figure 6.2&6.3).

Stage 1: We added a mutation m_1 in the 1600 generation in P_0, which then underwent balancing selection until generation 2200, when P_0 split into P_1 and P_2, where the frequency of m_1 was around 50%.

Stage 2: Then we explored different combinations of positive, balancing and negative selection of m_1 in P_1 and P_2. the frequency of m_1 reached 80%, 50% and 20% when it was positively, balancing or negatively selected, respectively, until generation 2899. Here we sampled 20 individuals each in P_1 and P_2 as the ancient samples.

Stage 3: Then P_1 and P_2 merged into P_3 in generation 2900. In P_3, for each combination of selection in Stage 2, we simulated positive, balancing and negative selection for m_1. The selection lasted for 20 generations, and then we sampled 4000 individuals from P_3 as the modern population.

Results: when m_1 was positively selected in only one of P_1 and P_2, and it experienced negative selection in P_3, the LDA scores around the locus of m_1 were low. Otherwise, no abnormal LDA scores were found at m_1.

Discussion

Introduction

In this thesis, I aimed to assess ancestral genetic contributions to modern phenotypes, in order to explain modern day disease load and distribution and shed light on the origins and evolution, including selection, of polygenic phenotypes. I aimed to do this using local ancestry assignments in a very large modern panel rather than directly from ancient DNA, in order to mitigate sampling bias and intervening selection or drift and to increase power, thereby giving a more direct measure of this than has been performed previously. I took advantage of recent advances in the sequencing of large numbers of ancient samples in Europe, as well as sophisticated understanding of the ancestry contributions to modern European populations from the Mesolithic, Neolithic and Bronze Age combined with ongoing research on the cultural practices and lifestyles of these populations. I focussed on Multiple Sclerosis (MS) in more depth, as an autoimmune disease showing a heterogeneous geographic distribution.

In this thesis, I have reported results based on painting the UK Biobank using a panel of ancient genomes to investigate ancestry component distributions and differential ancestral contributions to modern genetic risk. In Chapter One, I described modifications to Chromopainter (Lawson et al., 2012) to enable painting on a Biobank scale, including the efficient storage of local painting probabilities; I reported results for genome-wide ancestry proportions in Great Britain, derived using Non-Negative Least Squares. In Chapter Two, I outlined how it is possible to utilise non-British individuals in the UK Biobank to investigate ancestry in other countries in Eurasia and North Africa; I reported results of geographic variations in ancestry components for these countries.

In Chapter Three, I introduced a new statistic, analogous to a polygenic risk score, derived from the local painting probabilities for each ancestry. I reported results for traits already known to be over-dispersed in the ancient populations using PRS calculated directly using ancient genomes, including physical traits such as hair and skin colour, height, and BMI, psychological traits, and diseases like diabetes.

In Chapter Four, I applied these methods in a more in-depth way to MS, in collaboration with a team of others, to investigate the origins of genetic risk for MS. We found that risk can be largely attributed to genetic variants deriving from populations from the Pontic-Caspian Steppe, which came to Europe in the Bronze Age ~5,000 years ago and is associated with the Yamnaya, Corded Ware and Bell Beaker cultures, as well as the Afanasievo culture eastwards, and others. We showed that over time the cumulative risk of MS has increased,

mainly in the period 5,000-2,000 years ago, driven by increased risk in CHG ancestry present in the Steppe populations. We attempted to link this positive selection to the lifestyle of these first pastoralists, including exposure to novel pathogens. The distribution of Steppe ancestry in modern populations may explain at least part of the north-south gradient of this disease observed today.

This Chapter will first outline the theoretical implications of the research presented here, including how the research fits into existing theoretical frameworks and contributes to existing knowledge in the field. It will then discuss methodological implications, including limitations and potential improvements in future studies. Next, I discuss the practical implications, including recommendations for future studies using data generated here. Finally, there will be a conclusion reflecting on the main points discussed in this Chapter, as well as the role similar research has to play in the future.

Theoretical Implications

The results reported here add substantially to existing knowledge in several areas.

Firstly, the inference of local ancestry on a Biobank scale (i.e. hundreds of thousands of genomes) has not, to my knowledge, been attempted before. At a theoretical level, the ability to generate local ancestry labels on this scale enables the incorporation of LAI into statistical tests that require large numbers for power, such as GWAS or admixture mapping. This is discussed further below.

Although ChromoPainter is now somewhat dated, having recently passed its ten year anniversary and with newer LAI methods that work on a Biobank scale having been published during the course of this research (e.g. Hilmarsson et al., 2021), the power and novelty of the approach used here is in the incorporation of an external reference panel consisting of ancient individuals. This has only become possible recently, as the number of ancient samples has grown large enough to capture a significant degree of the genetic variation inherent in an ancient population, and our understanding of the demographic contributions to European prehistory have made confidence in such a panel possible. Using aDNA in this way ensures that local ancestry assignments are (1) easier to interpret than inferred ancestries, as the groupings are historically and genetically meaningful; and (2) enables the use of ancestries that are less diverged than those traditionally used in LAI, i.e. continent-scale (e.g. European vs African ancestry).

Because ChromoPainter has not been used with an ancient reference dataset such as this before, modelling was performed to test its accuracy. Sequence evolution was simulated under a model of European demography, with individuals sampled approximately corresponding to the ancient samples that were used in the actual reference panel. Although the accuracy of LAI was not hugely high, as expected using relatively small numbers of ancient individuals from ancestries which are not entirely diverged, the accuracy recorded was in line with other LAI methods (Schubert et al., 2020). This provides a degree of confidence that the local ancestry probabilities inferred in this work are robust and reliable. Furthermore, ancestry-specific differences in accuracy are not a cause for concern, as the downstream statistics account for these differences both genome-wide (NNLS) and locally (ARS).

Having applied ChromoPainter to the UK Biobank, it became possible to map the genome-wide (“global”) ancestry components present in the reference panel in modern

individuals; through reverse geocoding, the subtle signatures of these populations in modern people was observed for the first time at high resolution across Britain, clearly distinguishing Scotland and Wales from England, and showing gradients even within these countries. The fact that there are still detectable geographic differences in ancestry proportions today is clearly surprising given the age of these populations, and likely reflects demographic events involving later populations which themselves had differing proportions of the older ancestries. Performing the same analysis at the country level across Eurasia again revealed geographic differences which have hitherto been shown for some ancestries (e.g. Sikora et al., 2019) but never in such detail. The practical implications of these findings are discussed below.

The assumption that the 'white British' cohort in the UK Biobank as defined by Bycroft et al., (2018) (self-reported white British participants with outliers in PC space excluded) is of a 'single ancestry' is now firmly rejected, as has been shown in recent studies which have questioned whether population stratification has driven empirical results (Berg et al., 2019a; Sohail et al., 2019). Here, we have demonstrated a potential driver of this stratification, with diverse ancestries within the 'white British' cohort which vary geographically; these will cause geographic differences in allele frequencies which correlate with any variable that is also geographically structured. This questions the entire premise of a 'British' ethnicity, as used in e.g. medical and demographic studies.

Perhaps the main contribution of this thesis is developing methods to use this large dataset of local ancestry labels in modern individuals to estimate ancestral contributions to both single variants and polygenic traits. As outlined in the Introduction, most estimates of the latter are premised on polygenic risk scores calculated on grouped ancient samples or individuals; this faces myriad problems: exporting of effect sizes across time and space; differing LD patterns in ancient populations, meaning tag SNPs are less likely to tag the true effect; sampling bias of ancient samples (e.g. more likely to be elite individuals); intervening selection or drift skewing the estimation of the extent of the contribution. Here, I present a framework for a better estimate of this contribution, the Ancestral Risk Score. This statistic avoids or minimises each of the problems outlined above: it avoids exporting effect sizes over space and time, as it is merely concerned with with risk in a modern population; it does not rely on LD patterns in ancient individuals as it is calculated in the modern sample; it reduces sampling bias as one reference individual can be used as the donor a disproportionately high number of times (so, if a haplotype is under-sampled but present in the reference panel, it will be painted as the nearest neighbour the correct number of times in the modern panel); and finally intervening selection and drift are accounted for by

calculating the statistic in the modern population.

There is one study which has attempted a similar approach to that used here. Marnetto et al. (2022) use an approach utilising allele frequencies rather than haplotypes in order to assess the 'similarity' between a modern individual from the Estonian Biobank (Leitsalu et al., 2015) and an ancestral population at a specific genomic location; this statistic, *covA*, is averaged over a region around a putative GWAS hit, and is then used as a predictor for traits of interest. They are able to recover several known results, in line with the analyses presented here: height and hip/waist circumferences (higher in Yamnaya/Steppe ancestry, lower in WHG), and hair colour with some differences (e.g. whereas they found Yamnaya/Steppe ancestry had a high score for black hair, I found it had a high score for blonde). Some of these differences are due to the different ancestries included: both methods compare to an 'ancient sample mean' and use different ancient samples to represent each ancestry. However, there are also some more significant differences: we find opposite effects for cholesterol - Marnetto et al. (2022) find Yamnaya scores highly whereas WHG scores the lowest, whereas I found that WHG scores the highest by a considerable margin, with Yamnaya scoring the lowest (Chapter Three, Figure 1).

It is unclear how to consolidate these differences, given the very small number of studies performed in this field to date. However, there are several factors which I would argue advantages this study over Marnetto et al. (2022). The panel of ancient genomes used here is the largest of its kind ever used in such a study, and as a result is able to capture both more ancestries (ten versus four) and to represent each ancestry better. Furthermore, the size of the UK Biobank and subsequently ~8-fold higher number of modern samples gives much more power to our tests. A haplotype-based approach is also favourable over one which uses allele frequencies averaged over arbitrary lengths of the genome, as haplotypes capture ancestry information much better than allele frequencies, and the lengths that Marnetto et al. (2022) average over may include differing ancestries (especially over the larger intervals used). Finally, their use of variants directly from the GWAS catalogue is problematic as entries are not pruned or finemapped, and so may not be independent signals. Given all of these advantages, I suggest that a haplotype-based approach in a larger modern cohort with more ancient samples, using GWAS hits from single well-powered studies, is a more promising approach.

When this approach has been applied to MS, it has illuminated a so-far unknown association of MS with Steppe ancestry and somewhat inversely, an association of rheumatoid arthritis (RA) with hunter-gatherer ancestry. Having shown that Steppe ancestry is higher in northern

Europe than southern Europe (Chapter Two), this may explain (or partly explain) the modern-day geographic distribution of MS prevalence across Europe. Although infectious disease GWAS are severely lacking or under-powered in the literature, we were able to link many of the variants under selection with protection to a range of pathogens, and speculate that the signal of selection was in response to novel pathogens exposure due to lifestyle and technological innovations, such as massively increased exposure to livestock and consumption of unpasteurized milk and under-cooked meat. This hypothesis, that autoimmune diseases are the product of past selection against specific pathogens, has been stated many times (Liston et al., 2021). However, this usually takes the form of observing an overlap between genes influencing autoimmune disease and susceptibility to pathogenic infection (summarised in Matzaraki et al., 2017); with few studies showing statistical evidence for selection of these variants (with some exceptions, e.g. Zhernakova et al., 2010). Here, we have provided robust statistical evidence for polygenic selection, and identified both the timing and location of this signal. This is the first time that the genetic risk for an autoimmune disease has been localised to a specific ancestry in time and space, and raises the intriguing possibility of being able to identify cultural, technological or pathogenic changes that drove the change in this risk.

This study has practical implications, which are discussed below but, more broadly, this changes our understanding of autoimmune disease: it points towards ancestry-mediated geographic differences in prevalence due to differing selection pressures in the past combined with the incidental details of demographic change. This is arguably a new way of thinking about disease evolution; where previous studies have mainly looked at differences in susceptibility between modern populations, by decomposing ancestry in a modern population we are able to interrogate changes in past populations that no longer exist. It remains to be seen how many other diseases underwent population-specific evolution. Moving forward, geographically structured phenotypes would be an obvious place to continue.

The fact that MS risk increased significantly in the Bronze Age may not be a coincidence. Given that this period was a time of increased infectious diseases (Chapter Four, Discussion), it may be the case that other autoimmune diseases can trace their origins to this period. The Bronze Age appears to have been critical in ‘fine-tuning’ the modern immune system, to the extent that, although the “Neolithic Revolution” is usually taken to have had the greatest lasting impact on human health, perhaps the Bronze Age deserves further examination.

I decided to focus in more detail on a specific disease rather than casting the net more widely, as is more common in these types of study, in order to try to better understand the mechanism behind the selection and attempt to link it to a specific pressure. Although we were able to tie selection to a specific time and population, we were unable to find a single (pathogenic) pressure driving the MS-associated variants higher in frequency. In the absence of better association studies for infectious diseases or knowledge of the past distributions of pathogenic variants, it is difficult to know whether this is a result - i.e. a range of pathogens drove this evolution - or merely the result of being underpowered. For now, we can merely speculate that this ancestry-specific selection was likely driven by pathogenic pressure. We were also unable to find a link between the variants that were under selection versus those that were not. It remains to be seen whether the selected variants are interacting or not, although the results based on HTRX suggest that, within the HLA region, where nearly all the selected SNPs are located, interactions are important.

On a theoretical level, these results prompt the question of why or whether they are an improvement on studies which test for selection alone, and how these fields relate to each other. Both the major advantage and disadvantage of the ARS over tests for selection is that it explicitly tests for different contributions from given ancestries to genetic risk for a trait, and not for selection directly. While this is often due to selection (as in the case of MS), this is clearly not always the case.

This ambiguity with regards to selection can be useful: tests for selection have high thresholds for significance, and often signals of selection cannot be detected due to admixture or the decay of selective sweeps. If the question being asked is about the differential contribution of risk by ancestries, the ARS is a better approach. This is often the case, for example in explaining geographic differences in risk, which will not always (or even often) be due to past selection events; bottlenecks and drift may explain these differences. The ability to pick up these more subtle signatures, by calculating the statistic in a large modern cohort, gives the test greater sensitivity and applicability. This can be seen as a tradeoff between sensitivity and specificity - we increase the former at the expense of the latter. Finally, the ARS is extremely quick to calculate, providing convenient targets for testing for selection.

Even when there is a signal for selection, it is not always clear how to interpret this. The common central claim of these tests - that natural selection has acted on the focal trait - is often questionable. For example, it is extremely unlikely that the MS phenotype was under positive selection; rather, it is much more likely that a trait with related underlying genetic

architecture (i.e. shared variants or variants in high LD with MS risk-conferring variants) was being selected for instead, in this case protection against an infection or infections. This problem is perhaps more immediately acute for autoimmune disorders, where positive selection has been hypothesised (Liston et al., 2021), but the point holds that linking selection to a trait, even a polygenic trait, is extremely difficult. The weaker claim made by the ARS, that an ancestry has contributed differently to the genetic risk for a phenotype, is easier to justify.

Having said that there is an advantage in these methods over tests for selection, it is also reassuring to note that they corroborate the findings here. As expected, if there is evidence of positive selection in a specific ancestry, we would expect a higher contribution from that ancestry to modern risk. In Chapter Four, we tested for polygenic selection using inferred allele frequencies based on a novel chromosome painting technique which uses a neural network to assign haplotypes an ancestry path by looking at informative nodes of an ARG containing labelled individuals. This is an entirely different approach for LAI, and yet when we ran CLUES on the MS-associated variants, there was strong statistical evidence for positive selection, which could be accounted for by selection in the Steppe ancestry. Other approaches using the same ancestry labels (LDA and HTRX) also concurred with ARS results. Finally, these results were neutral with regard to which GWAS data were used.

Methodological Implications

There are several limitations to this work and its conclusions.

The first set of limitations is common to all studies which use aDNA. Although these data can give us a good idea of the genetics of people living in the past, there are several sources of potential bias: the sampling of ancient individuals is not random, for example elite individuals are more likely to be sampled, and cultures with burial practices are favoured over those which used cremation; the number of samples for each population differs, causing statistical power differences between ancestries (e.g. the ability to detect genetic risk contributions or selection); post-mortem DNA damage introduces alleles not reflective of the sample's actual genetic diversity (Allentoft et al., 2012); there is imputation bias towards populations used in the imputation reference panel (Günther & Nettelblad, 2019); splitting of people into 'populations' or 'ancestries', although data-driven, can be viewed as somewhat arbitrary based on the level at which splitting is imposed (Lawson, 2015); and the assumption of a correlation between genetics and culture, although often caveated, persists (Eisenmann et al., 2018). Although we have attempted to limit each of these, they are all present in this study.

ChromoPainter is designed to find coalescent tracts, or genealogically nearest neighbours, in a reference panel; it is not explicitly designed for LAI. If populations were well separated in the past (i.e. have a very old split time and no intervening admixture), existed at the same time, and samples were of homogeneous ancestry, this would not be a problem - the nearest neighbour in the reference panel would correspond to the local ancestry of that locus. However, in reality these conditions do not hold. The populations under consideration here are often admixed, and may be more recent mixtures of other populations in the reference panel (e.g. Steppe is a mixture of EHG and CHG). ChromoPainter attempts to control for the admixture of samples by learning copying priors, but we are still imposing categories that may not make total sense - for example, WHG and EHG existed in a continuum rather than as strictly separate groups. The temporal nature of the sampling should also theoretically mean that we observe 'masking' whereby samples coalesce with the more recent population and never with the older population that mixed to form it. In reality, if a haplotype is present in two populations we observe that ChromoPainter copies from both, but with a higher probability of copying from the more recent population. Downstream analyses such as ARS attempt to control for this by computing effect allele frequencies based on the relative rather than absolute painting of the effect versus the alternate allele in a given ancestry. While this should avoid bias (i.e. there is no reason to think an effect allele will be painted more than an

alternate allele due to masking), it may result in power differences between ancestries. This is exacerbated by smaller sample sizes in some ancestries, particularly the three hunter-gatherer groups.

There are however reasons to believe that these power differences between ancestries are not significant. If they were, we would expect the ancestries most susceptible to this to be closer to the ancient mean across all ancestries for each phenotype; this is not observed. WHG (unmasked) and EHG (masked by Yamnaya) show similar patterns of risk contribution, as we would expect for these similar ancestries. And finally many of the results here are recapitulated in other studies, giving us good reason to believe that we have power to detect signals in these ancestries as well as the larger, more recent ancestries.

Perhaps a better way to implement LAI using populations that are spread over time and space is to think about ancestry paths rather than assigning each ancestry separately, as was done in the alternative LAI in Chapter Four. Intuitively, this is a better way to think about ancestry, as routes backwards in time which may encompass several distinct populations. However, a potential drawback of this is that a model is needed with which to relate populations and decide on potential paths for inference. Here, we used Jones et al. (2015), but such a model is not available for ancestry in many parts of the world, and even for Europe this model is almost certainly over-simplified and could be wrong. Furthermore, methods based on reconstructing ARGs are currently computationally taxing and often unable to incorporate aDNA; even when they are, the path assignment is complex and with its own biases (Allentoft et al., 2022). In the future, for studies considering complex ancestry rather than simple cases of recent admixture between continent-scale ancestry, this is the likely direction that will be taken.

Even once local ancestry has been inferred and ancestral contributions to a complex phenotype estimated, misinterpretation of the results is very feasible or even likely, in a similar way to studies which reconstruct PRS for ancient samples. While the ARS is calculated on haplotypes which passed through a given ancestral population, this does not mean *per se* that it is a reflection of the phenotype or even genotype of that population, both because intervening drift, selection and sampling bias can skew the relative proportions of painted haplotypes at each locus, and because SNP effects are highly context dependent (both environmental and genomic). For example, high ARS for Farmer ancestry for traits related to mental health disorders, such as anxious feelings or irritability (Chapter Three), should not be interpreted as implying that Neolithic Farmers were more irritable or anxious. Likewise, we are not claiming that MS was selected for directly or that Steppe populations

had higher incidences of MS (Chapter Four). This is an easy and simplistic interpretation that we must be vigilant in emphasising is unjustified by these analyses.

Another methodological takeaway from this study is that there is great value in using the best powered GWAS studies for the association data. This sounds like an obvious recommendation but often, when the goal is to scan for selection signals in many phenotypes, a short-cut is taken and a single source is used for GWAS data, e.g. the UK Biobank. While this is understandable as an attempt to cut down on the hours required for researching GWAS studies, and downloading and formatting the required data, there are considerable costs: these biobanks are often very under-powered with respect to rarer diseases, and so many interesting signals will be overlooked even though theoretically they have been investigated. For example, MS was not found to be under statistically significant selection using data from the UK Biobank GWAS (http://www.nealelab.is/uk-biobank/), but was using a better powered study (IMSG, 2019).

Finally, geneticists often analyse and interpret genetic data and publish their findings without any reference to the intermediate, functional processes which link the genetic variants to the phenotypic outcome. There is an implicit assumption that these findings will eventually filter down to the functional and clinical research settings, with an abstract ‘benefit’ to health research. However, in reality this rarely happens: evolutionary genomics papers are inaccessible to many wet lab and clinical researchers, either due to technical misunderstandings or because they are unaware of them. Even if a researcher is aware of and understands one of these papers, it is often unclear how to interpret or use the results in a useful manner: a huge amount of expertise is required to know how to utilise the results effectively, which unsurprisingly is often lacking. My final methodological recommendation is therefore that these studies and geneticists working in phenotypic analysis of aDNA must do more to cross interdisciplinary boundaries, especially if the stated aim of the research is to impact human health. We must collaborate with clinicians and functional geneticists, and work harder to promote our work more widely through talks, interviews and articles.

Practical Implications

There are practical implications to the results presented in this thesis. These revolve around recommendations for future research and the practical applications of the data generated and findings reported here.

The first is that the dataset generated (painting probabilities for 410,000 individuals in the UK Biobank at ~550,000 sites) can be used for a wide variety of applications which go far beyond the analyses performed here. The painting probabilities can be imputed for any site on the genome, meaning the dataset can be expanded to any variant of interest, while the large number of samples painted provides enough statistical power for most tests, providing a significant resource for future studies. For example, the dataset could be used as a covariate in GWAS to control for ancestry at a local level; it could be used in admixture mapping studies to locate causal variants or assess ancestry-specific genetic risk; or it could be used to investigate the ancestry and origins of single alleles (including indels and inversions). In addition, each of the analyses performed here (e.g. ARS calculation, ancestry-GWAS etc) could be applied to any additional phenotype for which association data is available.

In Chapter Two, I presented methods to select individuals from the UK Biobank born in non-british countries and of a ‘typical ancestral background’ for that country, based on density-based scans of the first 18 PCs. Biobanks are extremely powerful and large sources of data, and yet due to requirements to control for ancestry as a confounder in statistical tests, often admixed or foreign individuals are discarded as a first step. As well as exacerbating existing inequalities in the amount of research resources directed towards different populations, with associated shortfalls in the quality of research conducted on non-European populations, this process results in a loss of statistical power. The ability to select individuals in this way therefore enables their inclusion in future studies, and can be applied to any large dataset of mixed ancestry where place of birth information is available.

While the methods and datasets presented here have practical implications for future research, so do the specific results. This falls into two categories: clinical, and social. The observation that MS risk has changed significantly at least twice in human history, both coincident with changes in lifestyle and environment, is intriguing: firstly in the Bronze Age, when we observe increasing genetic risk over time (Chapter Four), and secondly in the last five decades (Wallin et al., 2019; GBD 2016 Neurology Collaborators, 2019). It appears that a recently changing lifestyle has resulted in the realisation of genetic risk which appeared

thousands of years ago; I speculate that this may be what caused the selection against RA in the Bronze Age, as standing genetic variation became maladaptive due to changing environmental and lifestyle conditions. The link between changing risk and lifestyle, either through changes in risk allele frequencies or the conversion of this genetic variation to disease risk, fits into the increasingly prevalent view of MS as a 'lifestyle disease' (Jakimovski et al., 2019). Many of the environmental and lifestyle risk factors for MS interact with the HLA (Alfredsson and Olsson, 2019), suggesting that mechanistically they confer risk by interacting with the immune system and influence adaptive/innate immunity. This is increasingly an area of research for understanding MS aetiology, risk prediction, and disease management (Alfredsson and Olsson, 2019).

What, though, are the practical implications of understanding the specific history of MS risk presented here? Most of the results in this thesis can be viewed as basic science, furthering our understanding of the origins of an autoimmune disease. The practical implications are admittedly speculative. It is my hope that in the coming years the specific drivers of the positive selection for MS-associated variants will be identified, as more data on the distribution of pathogens in the past and GWAS results for pathogenic protection become available. For now, experiments which aim to directly dampen immune response in MS patients, such as studies using parasitic worms (helminth) (Dixit et al., 2017), may do well to target therapy at possible infections in the Steppe population: for example, the worms and parasites present in this population would have differed substantially from those infecting early Hunter-gatherers or farmers, due to the very specific diets and high meat consumption; some studies have tried infecting patients with swine worms with modest success (Jouvin & Kinet, 2012, Hansen et al., 2017), but as the Steppe population did not maintain or consume pigs, an obvious recommendation would be to change to worms which result from eating undercooked beef and dairy products instead.

Perhaps the most useful practical implication however is in risk prediction. The fact that there are ancestry-specific effects (Chapter Four, ancestry-GWAS) provides the opportunity to exploit these data for individual risk prediction. We have shown that ancestry-specific haplotypes improve out-of-sample prediction over tagging SNPs, which are used for PRS construction in clinical settings; these non-contiguous haplotypes (Chapter Four, HTRX) may offer a pathway to a better understanding of MS risk.

The second category of practical implications of these results is the social consequences of understanding better where a disease originated and therefore why people today are suffering from it. I hope that this can go some way to reconciling the 'randomness' many

people feel after diagnosis: while it may still feel unfair and unjust, perhaps the knowledge that these genetic variants once protected your ancestors against a range of pathogens might help, in a small way, to come to terms with the diagnosis. I also hope that this knowledge can go some way to mediating wider stigma around genetically-influenced diseases, in which genetic variants which increase risk are no longer seen as historical aberrations but as the result of complex and fascinating histories of selection.

Finally, this study has linked aDNA to medical applications more directly than nearly all previous studies incorporating aDNA. Although there is a very long way to go in translating these methods and results into clinical settings, another important step has been taken along this pathway. This both justifies funding for these types of basic research studies, many of which claim abstract health benefits in their grant applications, and provides hope for the future that our field will one day be able to deliver on this promise and improve people's lives.

Conclusion

In this chapter I have discussed how the work presented in this thesis has added to existing knowledge. The development and use of LAI on a Biobank scale, using an external reference panel consisting of ancient individuals, is novel. This dataset of local ancestry probabilities in the UK Biobank allowed the description of ancestry gradients across Great Britain, and across Eurasia using a new technique to select individuals of a 'typical ancestral background' for each country represented in the UK Biobank. This firmly rejected the view of a homogenous 'white British' population, confirming results from previous studies. I used this dataset to develop a new way of estimating the contributions of ancient genetic ancestries to polygenic disease risk today, and applied this to a variety of phenotypes; many of the drawbacks of previous methods were avoided. This is the first time that the genetic risk for an autoimmune disease has been localised to a specific ancestry in time and space, which raises the possibility of learning more about the cultural, environmental or pathogenic drivers of this signal in the past.

Methodologically, some limitations of this method became clear. aDNA presents a variety of challenges which are already well described. Furthermore ChromoPainter, although state-of-the-art previously, is no longer the best method for LAI; new methods, which take advantage of full ARG inference and can incorporate aDNA samples, will be used in the future. When applying tests for polygenic phenotypes, the choice of GWAS summary stats (and pruning methods) are important and often overlooked. Finally, studies which are able to incorporate direct links to the clinical setting, usually via collaboration with clinicians from the outset of the project, will have a greater impact on improving human health.

Practically, the dataset here can be used for a variety of future research questions, including GWAS and admixture mapping, while the methods can be applied to a wider selection of phenotypes, including in other Biobanks. The ability to include admixed individuals in studies using Biobanks is also an important step, offering another route to reducing the neglect of under-studied groups (i.e. non white and European in origin) in scientific research. The MS-specific results have clinical applications in terms of risk prediction and treatment, and social implications regarding how we view autoimmune diseases.

It is my belief that ancestry considerations, including local ancestry on a Biobank scale powered by increasingly large aDNA reference panels, provides the opportunity to significantly advance our understanding of the genetic basis for complex human phenotypes, including their origins. This will ultimately help us to understand why humans harbour genetic

mutations which are harmful, and why populations exhibit differences in genetic risk. These questions are of both fundamental evolutionary and practical, clinical interest. I hope that the work presented in this thesis goes a small way to realising that ambition.

Bibliography

Abra Brisbin *et al.* (2012) 'PCAdmix: Principal Components-Based Assignment of Ancestry Along Each Chromosome in Individuals with Admixed Ancestry from Two or More Populations', *Human Biology*, 84(4), pp. 343–364. Available at: https://doi.org/10.3378/027.084.0401.

Adams, C.I.M. *et al.* (2019) 'Beyond Biodiversity: Can Environmental DNA (eDNA) Cut It as a Population Genetics Tool?', *Genes*, 10(3), p. 192. Available at: https://doi.org/10.3390/genes10030192.

Adams, M.B. ed., 1990. 'The wellborn science: eugenics in Germany, France, Brazil, and Russia', Oxford University Press on Demand

Adeyemo, A. *et al.* (2021) 'Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps', *Nature Medicine*, 27(11), pp. 1876–1884. Available at: https://doi.org/10.1038/s41591-021-01549-6.

Adler, G. *et al.* (2017) 'Bosnian study of *APOE* distribution (BOSAD): a comparison with other European populations', *Annals of Human Biology*, 44(6), pp. 568–573. Available at: https://doi.org/10.1080/03014460.2017.1346708.

Alcina, A. *et al.* (2012) 'Multiple Sclerosis Risk Variant HLA-DRB1*1501 Associates with High Expression of DRB1 Gene in Different Human Populations', *PLoS ONE*. Edited by F. Palau, 7(1), p. e29819. Available at: https://doi.org/10.1371/journal.pone.0029819.

Alekseyenko, A.V. *et al.* (2011) 'Causal graph-based analysis of genome-wide association data in rheumatoid arthritis', *Biology Direct*, 6(1), p. 25. Available at: https://doi.org/10.1186/1745-6150-6-25.

Alexander, D.H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Research*, 19(9), pp. 1655–1664. Available at: https://doi.org/10.1101/gr.094052.109.

Alfredsson, L. and Olsson, T. (2019) 'Lifestyle and Environmental Factors in Multiple Sclerosis', *Cold Spring Harbor Perspectives in Medicine*, 9(4). Available at: https://doi.org/10.1101/cshperspect.a028944.

Allen, J.S., Bruss, J. and Damasio, H. (2005) 'The aging brain: The cognitive reserve hypothesis and hominid evolution', *American Journal of Human Biology*, 17(6), pp. 673–689. Available at: https://doi.org/10.1002/ajhb.20439.

Allentoft, M.E. *et al.* (2015) 'Population genomics of Bronze Age Eurasia', *Nature*, 522(7555), pp. 167–172. Available at: https://doi.org/10.1038/nature14507.

Allentoft, M.E. *et al.* (2022) *Population Genomics of Stone Age Eurasia*. preprint. Evolutionary Biology. Available at: https://doi.org/10.1101/2022.05.04.490594.

Anisimova, M. (ed.) (2019) *Evolutionary Genomics: Statistical and Computational Methods*. New York, NY: Springer New York (Methods in Molecular Biology). Available at: https://doi.org/10.1007/978-1-4939-9074-0.

Araújo, A., Reinhard, K. and Ferreira, L.F. (2015) 'Palaeoparasitology — Human Parasites in Ancient Material', in *Advances in Parasitology*. Elsevier, pp. 349–387. Available at: https://doi.org/10.1016/bs.apar.2015.03.003.

Atkinson, E.G. *et al.* (2021) 'Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power', *Nature Genetics*, 53(2), pp. 195–204. Available at: https://doi.org/10.1038/s41588-020-00766-y.

Attfield, K.E. *et al.* (2022) 'The immunology of multiple sclerosis', *Nature Reviews Immunology* [Preprint]. Available at: https://doi.org/10.1038/s41577-022-00718-z.

Benton, M.L. *et al.* (2021) 'The influence of evolutionary history on human health and disease', *Nature Reviews Genetics*, 22(5), pp. 269–283. Available at: https://doi.org/10.1038/s41576-020-00305-9.

Berg, J.J. and Coop, G. (2014) 'A Population Genetic Signal of Polygenic Adaptation', *PLoS Genetics*. Edited by M. W. Feldman, 10(8), p. e1004412. Available at: https://doi.org/10.1371/journal.pgen.1004412.

Berg, J.J. *et al.* (2019) 'Reduced signal for polygenic adaptation of height in UK Biobank', *eLife*. Edited by M. Nordborg et al., 8, p. e39725. Available at: https://doi.org/10.7554/eLife.39725.

Bergström, A. *et al.* (2020) 'Insights into human genetic variation and population history from 929 diverse genomes', p. 10.

Berisa, T. and Pickrell, J.K. (2015) 'Approximately independent linkage disequilibrium blocks in human populations', *Bioinformatics*, p. btv546. Available at: https://doi.org/10.1093/bioinformatics/btv546.

Bersaglieri, T. *et al.* (2004) 'Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene', *The American Journal of Human Genetics*, 74(6), pp. 1111–1120. Available at: https://doi.org/10.1086/421051.

Betti, L. *et al.* (2020) 'Climate shaped how Neolithic farmers and European hunter-gatherers interacted after a major slowdown from 6,100 bce to 4,500 bce', *Nature Human Behaviour*, 4(10), pp. 1004–1010. Available at: https://doi.org/10.1038/s41562-020-0897-7.

Bihrmann, K. *et al.* (2018) ‘Small-scale geographical variation in multiple sclerosis: A case-control study using Danish register data 1971–2013’, *Multiple Sclerosis and Related Disorders*, 23, pp. 40–45. Available at: https://doi.org/10.1016/j.msard.2018.04.021.

Bitarello, B.D. and Mathieson, I. (2020) ‘Supplemental Material for Bitarello and Mathieson, 2020’. GSA Journals. Available at: https://doi.org/10.25387/G3.12795887.

Bjornevik, K. *et al.* (2022) ‘Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis’, *Science*, 375(6578), pp. 296–301. Available at: https://doi.org/10.1126/science.abj8222.

Booker, T.R., Yeaman, S. and Whitlock, M.C. (2019) *Global adaptation confounds the search for local adaptation*. preprint. Evolutionary Biology. Available at: https://doi.org/10.1101/742247.

Bos, K.I. *et al.* (2014) ‘Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis’, *Nature*, 514(7523), pp. 494–497. Available at: https://doi.org/10.1038/nature13591.

Boyle, E.A., Li, Y.I. and Pritchard, J.K. (2017) ‘An Expanded View of Complex Traits: From Polygenic to Omnigenic’, *Cell*, 169(7), pp. 1177–1186. Available at: https://doi.org/10.1016/j.cell.2017.05.038.

Brace, S. *et al.* (2019) ‘Ancient genomes indicate population replacement in Early Neolithic Britain’, *Nature Ecology & Evolution*, 3(5), pp. 765–771. Available at: https://doi.org/10.1038/s41559-019-0871-9.

Brinkworth, J. F. (2017) ‘Infectious Disease and the Diversification of the Human Genome’, *Human Biology*, 89(1), pp. 47–65. Available at: https://doi.org/10.13110/humanbiology.89.1.03.

Broushaki, F. *et al.* (2016) ‘Early Neolithic genomes from the eastern Fertile Crescent’, *Science*, 353(6298), pp. 499–503. Available at: https://doi.org/10.1126/science.aaf7943.

Browning, B.L. and Browning, S.R. (2013) ‘Detecting Identity by Descent and Estimating Genotype Error Rates in Sequence Data’, *The American Journal of Human Genetics*, 93(5), pp. 840–851. Available at: https://doi.org/10.1016/j.ajhg.2013.09.014.

Burchard, E.G. *et al.* (2003) ‘The Importance of Race and Ethnic Background in Biomedical Research and Clinical Practice’, *New England Journal of Medicine*, 348(12), pp. 1170–1175. Available at: https://doi.org/10.1056/NEJMsb025007.

Bycroft, C. *et al.* (2018) ‘The UK Biobank resource with deep phenotyping and genomic data’, *Nature*, 562(7726), pp. 203–209. Available at: https://doi.org/10.1038/s41586-018-0579-z.

Cappellini, E. *et al.* (2018) 'Ancient Biomolecules and Evolutionary Inference', *Annual Review of Biochemistry*, 87(1), pp. 1029–1060. Available at: https://doi.org/10.1146/annurev-biochem-062917-012002.

Carmi, S. *et al.* (2014) 'Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins', *Nature Communications*, 5(1), p. 4835. Available at: https://doi.org/10.1038/ncomms5835.

Carvalho-Wells, A.L. *et al.* (2010) 'Interactions between age and apoE genotype on fasting and postprandial triglycerides levels', *Atherosclerosis*, 212(2), pp. 481–487. Available at: https://doi.org/10.1016/j.atherosclerosis.2010.06.036.

Cassidy, L.M. *et al.* (2016) 'Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome', *Proceedings of the National Academy of Sciences*, 113(2), pp. 368–373. Available at: https://doi.org/10.1073/pnas.1518445113.

Cassidy, L.M. *et al.* (2020) 'A dynastic elite in monumental Neolithic society', *Nature*, 582(7812), pp. 384–388. Available at: https://doi.org/10.1038/s41586-020-2378-6.

Chang, C.C. *et al.* (2015) 'Second-generation PLINK: rising to the challenge of larger and richer datasets', *GigaScience*, 4(1), p. 7. Available at: https://doi.org/10.1186/s13742-015-0047-8.

Chi, C. *et al.* (2019) 'Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry', *PLOS Genetics*. Edited by W. Bush, 15(1), p. e1007808. Available at: https://doi.org/10.1371/journal.pgen.1007808.

Chiang, C.W.K. *et al.* (2018) 'Genomic history of the Sardinian population', *Nature Genetics*, 50(10), pp. 1426–1434. Available at: https://doi.org/10.1038/s41588-018-0215-8.

Choi, S.W. and O'Reilly, P.F. (2019) 'PRSice-2: Polygenic Risk Score software for biobank-scale data', *GigaScience*, 8(7), p. giz082. Available at: https://doi.org/10.1093/gigascience/giz082.

Comabella, M. *et al.* (2008) 'Identification of a Novel Risk Locus for Multiple Sclerosis at 13q31.3 by a Pooled Genome-Wide Scan of 500,000 Single Nucleotide Polymorphisms', *PLoS ONE*. Edited by K. Gwinn, 3(10), p. e3490. Available at: https://doi.org/10.1371/journal.pone.0003490.

CORBO, R.M. and SCACCHI, R. (1999) 'Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a "thrifty" allele?', *Annals of Human Genetics*, 63(4), pp. 301–310. Available at: https://doi.org/10.1046/j.1469-1809.1999.6340301.x.

CORBO, R.M. *et al.* (1995) 'Apolipoprotein E polymorphism in Italy investigated in native plasma by a simple polyacrylamide gel isoelectric focusing technique. Comparison with

frequency data of other European populations', *Annals of Human Genetics*, 59(2), pp. 197–209. Available at: https://doi.org/10.1111/j.1469-1809.1995.tb00741.x.

Corder, E.H. *et al.* (1993) 'Gene Dose of Apolipoprotein E Type 4 Allele and the Risk of Alzheimer's Disease in Late Onset Families', *Science*, 261(5123), pp. 921–923. Available at: https://doi.org/10.1126/science.8346443.

Cotsapas, C. and Mitrovic, M. (2018) 'Genome-wide association studies of multiple sclerosis', *Clinical & Translational Immunology*, 7(6), p. e1018. Available at: https://doi.org/10.1002/cti2.1018.

Cox, S.L. *et al.* (2019) 'Genetic contributions to variation in human stature in prehistoric Europe', *Proceedings of the National Academy of Sciences*, 116(43), pp. 21484–21492. Available at: https://doi.org/10.1073/pnas.1910606116.

Cox, S.L. *et al.* (no date) 'Predicting skeletal stature using ancient DNA', p. 20.

Cree, B.A.C., Goodin, D.S. and Hauser, S.L. (2003) 'Neuromyelitis Optica', *Semin Neurol*, 22(02), pp. 105–122.

Cristescu, B. and Boyce, M.S. (2013) 'Focusing Ecological Research for Conservation', *AMBIO*, 42(7), pp. 805–815. Available at: https://doi.org/10.1007/s13280-013-0410-x.

Cubas, M. *et al.* (2020) 'Latitudinal gradient in dairy production with the introduction of farming in Atlantic Europe', *Nature Communications*, 11(1), p. 2036. Available at: https://doi.org/10.1038/s41467-020-15907-4.

Damgaard, P. de B. *et al.* (2018) '137 ancient human genomes from across the Eurasian steppes', *Nature*, 557(7705), pp. 369–374. Available at: https://doi.org/10.1038/s41586-018-0094-2.

Danke, N.A. and Kwok, W.W. (2003) 'HLA Class II-Restricted $CD4^+$ T Cell Responses Directed Against Influenza Viral Antigens Postinfluenza Vaccination', *The Journal of Immunology*, 171(6), p. 3163. Available at: https://doi.org/10.4049/jimmunol.171.6.3163.

Das, H. *et al.* (no date) 'MICA Engagement by Human V𝛾2V𝛿2 T Cells Enhances Their Antigen-Dependent Effector Function', p. 11.

de-Almada, B.V.P. *et al.* (2012) 'Protective effect of the APOE-e3 allele in Alzheimer's disease', *Brazilian Journal of Medical and Biological Research*, 45(1), pp. 8–12. Available at: https://doi.org/10.1590/S0100-879X2011007500151.

Deane, K.D. *et al.* (2017) 'Genetic and environmental risk factors for rheumatoid arthritis', *Individuals at risk of rheumatoid arthritis—the evolving story*, 31(1), pp. 3–18. Available at: https://doi.org/10.1016/j.berh.2017.08.003.

Dehasque, M. *et al.* (2020) 'Inference of natural selection from ancient DNA', *Evolution Letters*, 4(2), pp. 94–108. Available at: https://doi.org/10.1002/evl3.165.

Diao, L. and Chen, K.C. (2012) 'Local Ancestry Corrects for Population Structure in *Saccharomyces cerevisiae* Genome-Wide Association Studies', *Genetics*, 192(4), pp. 1503–1511. Available at: https://doi.org/10.1534/genetics.112.144790.

Dimopoulos, E.A. *et al.* (2020) *HAYSTAC: A Bayesian framework for robust and rapid species identification in high-throughput sequencing data*. preprint. Bioinformatics. Available at: https://doi.org/10.1101/2020.12.16.419085.

Ding, Y. *et al.* (2022) 'Polygenic scoring accuracy varies across the genetic ancestry continuum in all human populations', *bioRxiv*, p. 2022.09.28.509988. Available at: https://doi.org/10.1101/2022.09.28.509988.

Dixit, A. *et al.* (2017) 'Novel Therapeutics for Multiple Sclerosis Designed by Parasitic Worms', *International Journal of Molecular Sciences*, 18(10). Available at: https://doi.org/10.3390/ijms18102141.

Donnelly, P. and Tavaré, S. (1987) 'The population genealogy of the infinitely-many neutral alleles model', *Journal of Mathematical Biology*, 25(4), pp. 381–391. Available at: https://doi.org/10.1007/BF00277163.

Dou, Y. *et al.* (2020) 'Expanding Diversity of Susceptible Hosts in Peste Des Petits Ruminants Virus Infection and Its Potential Mechanism Beyond', *Frontiers in Veterinary Science*, 7. Available at: https://www.frontiersin.org/articles/10.3389/fvets.2020.00066.

Dulberger, C.L. *et al.* (2017) 'Human Leukocyte Antigen F Presents Peptides and Regulates Immunity through Interactions with NK Cell Receptors', *Immunity*, 46(6), pp. 1018-1029.e7. Available at: https://doi.org/10.1016/j.immuni.2017.06.002.

Duncan, L.E., Ostacher, M. and Ballon, J. (2019) 'How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete', *Neuropsychopharmacology*, 44(9), pp. 1518–1523. Available at: https://doi.org/10.1038/s41386-019-0389-5.

Duncan, L. *et al.* (2019) 'Analysis of polygenic risk score usage and performance in diverse human populations', *Nature Communications*, 10(1), p. 3328. Available at: https://doi.org/10.1038/s41467-019-11112-0.

Efron, B. (1979) 'Bootstrap Methods: Another Look at the Jackknife', *The Annals of Statistics*, 7(1,), pp. 1–26.

Efron, B. (1987) 'Better Bootstrap Confidence Intervals', *Journal of the American Statistical Association*, 82(397), pp. 171–185. Available at: https://doi.org/10.2307/2289144.

Egan, J.J. *et al.* (1995) 'Epstein-Barr virus replication within pulmonary epithelial cells in cryptogenic fibrosing alveolitis.', *Thorax*, 50(12), pp. 1234–1239. Available at: https://doi.org/10.1136/thx.50.12.1234.

Eisenmann, S. *et al.* (2018) 'Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis', *Scientific Reports*, 8(1), p. 13003. Available at: https://doi.org/10.1038/s41598-018-31123-z.

Entezami, P. *et al.* (2011) 'Historical Perspective on the Etiology of Rheumatoid Arthritis', *Current Concepts in the Treatment of the Rheumatoid Hand, Wrist and Elbow*, 27(1), pp. 1–10. Available at: https://doi.org/10.1016/j.hcl.2010.09.006.

Esteller-Cucala, P. *et al.* (2020) 'Genomic analysis of the natural history of attention-deficit/hyperactivity disorder using Neanderthal and ancient Homo sapiens samples', *Scientific Reports*, 10(1), p. 8622. Available at: https://doi.org/10.1038/s41598-020-65322-4.

Evershed, R.P. *et al.* (2022) 'Dairying, diseases and the evolution of lactase persistence in Europe', *Nature*, 608(7922), pp. 336–345. Available at: https://doi.org/10.1038/s41586-022-05010-7.

Excoffier, L., Hofer, T. and Foll, M. (2009) 'Detecting loci under selection in a hierarchically structured population', *Heredity*, 103(4), pp. 285–298. Available at: https://doi.org/10.1038/hdy.2009.74.

Fares, A. (2011) 'Seasonality of tuberculosis', *Journal of Global Infectious Diseases*. Wolters Kluwer Medknow Publications.

Feigin, V.L. *et al.* (2019) 'Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016', *The Lancet Neurology*, 18(5), pp. 459–480. Available at: https://doi.org/10.1016/S1474-4422(18)30499-X.

Fejerman, L. *et al.* (2012) 'Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas', *Human Molecular Genetics*, 21(8), pp. 1907–1917. Available at: https://doi.org/10.1093/hmg/ddr617.

Feldman, M. *et al.* (2019) 'Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia', *Nature Communications*, 10(1), p. 1218. Available at: https://doi.org/10.1038/s41467-019-09209-7.

Finch, C.E. and Stanford, C.B. (2004) 'Meat-Adaptive Genes and the Evolution of Slower Aging in Humans', *The Quarterly Review of Biology*, 79(1), pp. 3–50. Available at: https://doi.org/10.1086/381662.

Fischer, A. & Kristiansen, K. (2002) ‘The Neolithisation of Denmark. 150 years of debate’, J.R. Collis Publications

Fleming, J. and Fabry, Z. (2007) ‘The hygiene hypothesis and multiple sclerosis’, *Annals of Neurology*, 61(2), pp. 85–89. Available at: https://doi.org/10.1002/ana.21092.

Fort, J. (2015) ‘Demic and cultural diffusion propagated the Neolithic transition across different regions of Europe’, J. R. Soc. Interface 12

Fowler, C., Harding, J. & Hofmann, D. (2015) ‘Introduction’ in The Oxford Handbook of Neolithic Europe (eds. Fowler, C., Harding, J. & Hofmann, D.), Oxford University Press

Frangos, C.C. and Schucany, W.R. (1990) ‘Jackknife estimation of the bootstrap acceleration constant’, *Computational Statistics & Data Analysis*, 9(3), pp. 271–281. Available at: https://doi.org/10.1016/0167-9473(90)90109-U.

Fu, Q. *et al.* (2013) ‘A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes’, *Current Biology*, 23(7), pp. 553–559. Available at: https://doi.org/10.1016/j.cub.2013.02.044.

Fu, Q. *et al.* (2015) ‘An early modern human from Romania with a recent Neanderthal ancestor’, *Nature*, 524(7564), pp. 216–219. Available at: https://doi.org/10.1038/nature14558.

Fu, Q. *et al.* (2016) ‘The genetic history of Ice Age Europe’, *Nature*, 534(7606), pp. 200–205. Available at: https://doi.org/10.1038/nature17993.

Fugger, L., Jensen, L.T. and Rossjohn, J. (2020) ‘Challenges, Progress, and Prospects of Developing Therapies to Treat Autoimmune Diseases’, *Cell*, 181(1), pp. 63–80. Available at: https://doi.org/10.1016/j.cell.2020.03.007.

Fullerton, S.M. *et al.* (2000) ‘Apolipoprotein E Variation at the Sequence Haplotype Level: Implications for the Origin and Maintenance of a Major Human Polymorphism’, *The American Journal of Human Genetics*, 67(4), pp. 881–900. Available at: https://doi.org/10.1086/303070.

Fumagalli, M. *et al.* (2011) ‘Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution’, *PLOS Genetics*, 7(11), p. e1002355. Available at: https://doi.org/10.1371/journal.pgen.1002355.

Gale, R. and Martyn, N. (no date) ‘MIGRANT STUDIES IN MULTIPLE SCLEROSIS’, p. 24.

Galinsky, K.J. *et al.* (2016) ‘Population Structure of UK Biobank and Ancient Eurasians Reveals Adaptation at Genes Influencing Blood Pressure’, *The American Journal of Human Genetics*, 99(5), pp. 1130–1139. Available at: https://doi.org/10.1016/j.ajhg.2016.09.014.

Garrison, E. *et al.* (2018) ‘Variation graph toolkit improves read mapping by representing genetic variation in the reference’, *Nature Biotechnology*, 36(9), pp. 875–879. Available at: https://doi.org/10.1038/nbt.4227.

Gaulton, K.J. *et al.* (2015) ‘Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci’, *Nature Genetics*, 47(12), pp. 1415–1425. Available at: https://doi.org/10.1038/ng.3437.

Gokhman, D. *et al.* (2019) ‘Reconstructing Denisovan Anatomy Using DNA Methylation Maps’, *Cell*, 179(1), pp. 180-192.e10. Available at: https://doi.org/10.1016/j.cell.2019.08.035.

Goldberg, A. *et al.* (2017) ‘Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations’, *Proceedings of the National Academy of Sciences*, 114(10), pp. 2657–2662. Available at: https://doi.org/10.1073/pnas.1616392114.

Gompo, T.R. *et al.* (2020) ‘Risk factors of tuberculosis in human and its association with cattle TB in Nepal: A one health approach’, *One Health*, 10, p. 100156. Available at: https://doi.org/10.1016/j.onehlt.2020.100156.

Graves, C.J. and Weinreich, D.M. (2017) ‘Variability in Fitness Effects Can Preclude Selection of the Fittest’, *Annual Review of Ecology, Evolution, and Systematics*, 48(1), pp. 399–417. Available at: https://doi.org/10.1146/annurev-ecolsys-110316-022722.

Gregersen, J.W. *et al.* (2006) ‘Functional epistasis on a common MHC haplotype associated with multiple sclerosis’, *Nature*, 443(7111), pp. 574–577. Available at: https://doi.org/10.1038/nature05133.

Gretzinger, J. *et al.* (2022) ‘The Anglo-Saxon migration and the formation of the early English gene pool’, *Nature*, 610(7930), pp. 112–119. Available at: https://doi.org/10.1038/s41586-022-05247-2.

Griffiths, R.C. (1991) ‘The Two-Locus Ancestral Graph’, *Lecture Notes-Monograph Series*, 18, pp. 100–117.

Grytten, N. *et al.* (2006) ‘A 50-year follow-up of the incidence of multiple sclerosis in Hordaland County, Norway’, *Neurology*, 66(2), pp. 182–186. Available at: https://doi.org/10.1212/01.wnl.0000195549.95448.b9.

Grønlie, S.A. *et al.* (2000) ‘Multiple sclerosis in North Norway, and first appearance in an indigenous population’, *Journal of Neurology*, 247(2), pp. 129–133. Available at: https://doi.org/10.1007/PL00007793.

Guellil, M. *et al.* (no date) ‘Ancient herpes simplex 1 genomes reveal recent viral structure in Eurasia’, *Science Advances*, 8(30), p. eabo4435. Available at: https://doi.org/10.1126/sciadv.abo4435.

Guggisberg, M. (2018) in Oxford Handbook of the European Iron Age (eds Rebay-Sailsbury, K., Haselgrove, C. & Wells, P.), Oxford Univ. Press

Günther, T. and Nettelblad, C. (2019) 'The presence and impact of reference bias on population genomic studies of prehistoric human populations', *PLOS Genetics*. Edited by A. Di Rienzo, 15(7), p. e1008302. Available at: https://doi.org/10.1371/journal.pgen.1008302.

Gutierrez-Achury, J. *et al.* (2015) 'Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease', *Nature Genetics*, 47(6), pp. 577–578. Available at: https://doi.org/10.1038/ng.3268.

Haak, W. *et al.* (2015) 'Massive migration from the steppe was a source for Indo-European languages in Europe', *Nature*, 522(7555), pp. 207–211. Available at: https://doi.org/10.1038/nature14317.

Hamid, I. *et al.* (2021) 'Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde', *eLife*. Edited by M. Przeworski et al., 10, p. e63177. Available at: https://doi.org/10.7554/eLife.63177.

Hancock, A.M. *et al.* (2008) 'Adaptations to Climate in Candidate Genes for Common Metabolic Disorders', *PLoS Genetics*. Edited by D.A. Petrov, 4(2), p. e32. Available at: https://doi.org/10.1371/journal.pgen.0040032.

Hansen, C.S. *et al.* (2017) 'Trichuris suis secrete products that reduce disease severity in a multiple sclerosis model', 62(1), pp. 22–28. Available at: https://doi.org/10.1515/ap-2017-0002.

Haworth, S. *et al.* (2019) 'Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis', *Nature Communications*, 10(1), p. 333. Available at: https://doi.org/10.1038/s41467-018-08219-1.

He, Z. *et al.* (2020) 'Detecting and Quantifying Natural Selection at Two Linked Loci from Time Series Data of Allele Frequencies with Forward-in-Time Simulations', *Genetics*, 216(2), pp. 521–541. Available at: https://doi.org/10.1534/genetics.120.303463.

Hellenthal, G. *et al.* (2014) 'A Genetic Atlas of Human Admixture History', *Science*, 343(6172), pp. 747–751. Available at: https://doi.org/10.1126/science.1243518.

Henn, B.M. *et al.* (2015) 'Estimating the mutation load in human genomes', *Nature Reviews Genetics*, 16(6), pp. 333–343. Available at: https://doi.org/10.1038/nrg3931.

Heyd, V. (2017) 'Kossinna's smile', *Antiquity*. 2017/04/04 edn, 91(356), pp. 348–359. Available at: https://doi.org/10.15184/aqy.2017.21.

Higham, T. *et al.* (2011) 'The earliest evidence for anatomically modern humans in northwestern Europe', *Nature*, 479(7374), pp. 521–524. Available at: https://doi.org/10.1038/nature10484.

Hill, A.V.S. (2001) 'The Genomics and Genetics of Human Infectious Disease Susceptibility', *Annual Review of Genomics and Human Genetics*, 2(1), pp. 373–400. Available at: https://doi.org/10.1146/annurev.genom.2.1.373.

Hilmarsson, H. *et al.* (2021) 'High Resolution Ancestry Deconvolution for Next Generation Genomic Data', *bioRxiv*, p. 2021.09.19.460980. Available at: https://doi.org/10.1101/2021.09.19.460980.

Hofmanová, Z. *et al.* (2016) 'Early farmers from across Europe directly descended from Neolithic Aegeans', *Proceedings of the National Academy of Sciences*, 113(25), pp. 6886–6891. Available at: https://doi.org/10.1073/pnas.1523951113.

Hubisz, M.J., Williams, A.L. and Siepel, A. (2020) 'Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph', *PLOS Genetics*, 16(8), p. e1008895. Available at: https://doi.org/10.1371/journal.pgen.1008895.

Hudson, R.R. (1983) 'Properties of a neutral allele model with intragenic recombination', *Theoretical Population Biology*, 23(2), pp. 183–201. Available at: https://doi.org/10.1016/0040-5809(83)90013-8.

Huebbe, P. *et al.* (2011) '*APOE* ε *4* is associated with higher vitamin D levels in targeted replacement mice and humans', *The FASEB Journal*, 25(9), pp. 3262–3270. Available at: https://doi.org/10.1096/fj.11-180935.

Hydes, T.J. *et al.* (2015) 'The interaction of genetic determinants in the outcome of HCV infection: evidence for discrete immunological pathways: Immunological pathways act discretely to bring about HCV clearance', *Tissue Antigens*, 86(4), pp. 267–275. Available at: https://doi.org/10.1111/tan.12650.

Immel, A. *et al.* (2021) 'Genome-wide study of a Neolithic Wartberg grave community reveals distinct HLA variation and hunter-gatherer ancestry', *Communications Biology*, 4(1), p. 113. Available at: https://doi.org/10.1038/s42003-020-01627-4.

International Multiple Sclerosis Genetics Consortium *et al.* (2019) 'Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility', *Science*, 365(6460), p. eaav7188. Available at: https://doi.org/10.1126/science.aav7188.

Ioannidis, A.G. *et al.* (2020) 'Native American gene flow into Polynesia predating Easter Island settlement', *Nature*, 583(7817), pp. 572–577. Available at: https://doi.org/10.1038/s41586-020-2487-2.

Irving-Pease, E.K. *et al.* (2021) ‘Quantitative Human Paleogenetics: What can Ancient DNA Tell us About Complex Trait Evolution?’, *Frontiers in Genetics*, 12, p. 703541. Available at: https://doi.org/10.3389/fgene.2021.703541.

Irving-Pease, E.K. *et al.* (2022) ‘The Selection Landscape and Genetic Legacy of Ancient Eurasians’, *bioRxiv*, p. 2022.09.22.509027. Available at: https://doi.org/10.1101/2022.09.22.509027.

Ishigaki, K. *et al.* (2021) *Trans-ancestry genome-wide association study identifies novel genetic mechanisms in rheumatoid arthritis*. preprint. Genetic and Genomic Medicine. Available at: https://doi.org/10.1101/2021.12.01.21267132.

Itan, Y. *et al.* (2009) ‘The Origins of Lactase Persistence in Europe’, *PLoS Computational Biology*. Edited by M.M. Tanaka, 5(8), p. e1000491. Available at: https://doi.org/10.1371/journal.pcbi.1000491.

Jablonski, N.G. and Chaplin, G. (2017) ‘The colours of humanity: the evolution of pigmentation in the human lineage’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1724), p. 20160349. Available at: https://doi.org/10.1098/rstb.2016.0349.

Jacobs, B.M. *et al.* (2021) ‘Gene-Environment Interactions in Multiple Sclerosis: A UK Biobank Study’, *Neurology - Neuroimmunology Neuroinflammation*, 8(4), p. e1007. Available at: https://doi.org/10.1212/NXI.0000000000001007.

Jakimovski, D. *et al.* (2019) ‘Lifestyle-based modifiable risk factors in multiple sclerosis: review of experimental and clinical findings’, *Neurodegenerative Disease Management*, 9(3), pp. 149–172. Available at: https://doi.org/10.2217/nmt-2018-0046.

Jensen, T.Z. *et al.* (2018) *Stone Age ‘chewing gum’ yields 5,700 year-old human genome and oral microbiome*. preprint. Genomics. Available at: https://doi.org/10.1101/493882.

Johnsen, P.V. *et al.* (2021) ‘A new method for exploring gene–gene and gene–environment interactions in GWAS with tree ensemble methods and SHAP values’, *BMC Bioinformatics*, 22(1), p. 230. Available at: https://doi.org/10.1186/s12859-021-04041-7.

Jones, E.R. *et al.* (2015) ‘Upper Palaeolithic genomes reveal deep roots of modern Eurasians’, *Nature Communications*, 6(1), p. 8912. Available at: https://doi.org/10.1038/ncomms9912.

Jones, E.R. *et al.* (2017) ‘The Neolithic Transition in the Baltic Was Not Driven by Admixture with Early European Farmers’, *Current Biology*, 27(4), pp. 576–582. Available at: https://doi.org/10.1016/j.cub.2016.12.060.

Jónsson, H. *et al.* (2013) ‘mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters’, *Bioinformatics*, 29(13), pp. 1682–1684. Available at: https://doi.org/10.1093/bioinformatics/btt193.

Joo, Y.B. *et al.* (2019) ‘Respiratory viral infections and the risk of rheumatoid arthritis’, *Arthritis Research & Therapy*, 21(1), p. 199. Available at: https://doi.org/10.1186/s13075-019-1977-9.

Jouvin, M.-H. and Kinet, J.-P. (2012) ‘Trichuris suis ova: Testing a helminth-based therapy as an extension of the hygiene hypothesis’, *Journal of Allergy and Clinical Immunology*, 130(1), pp. 3–10. Available at: https://doi.org/10.1016/j.jaci.2012.05.028.

Ju, D. and Mathieson, I. (2021) ‘The evolution of skin pigmentation-associated variation in West Eurasia’, *Proceedings of the National Academy of Sciences*, 118(1), p. e2009227118. Available at: https://doi.org/10.1073/pnas.2009227118.

Katoh, H. *et al.* (2015) ‘Mumps Virus Is Released from the Apical Surface of Polarized Epithelial Cells, and the Release Is Facilitated by a Rab11-Mediated Transport System’, *Journal of Virology*. Edited by D.S. Lyles, 89(23), pp. 12026–12034. Available at: https://doi.org/10.1128/JVI.02048-15.

Kelleher, J., Etheridge, A.M. and McVean, G. (2016) ‘Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes’, *PLOS Computational Biology*. Edited by Y.S. Song, 12(5), p. e1004842. Available at: https://doi.org/10.1371/journal.pcbi.1004842.

Kelleher, J. *et al.* (2019) ‘Inferring whole-genome histories in large population datasets’, *Nature Genetics*, 51(9), pp. 1330–1338. Available at: https://doi.org/10.1038/s41588-019-0483-y.

Kenney, A.D. *et al.* (2017) ‘Human Genetic Determinants of Viral Diseases’, *Annual Review of Genetics*, 51(1), pp. 241–263. Available at: https://doi.org/10.1146/annurev-genet-120116-023425.

Kerner, G. *et al.* (2021) ‘Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years’, *The American Journal of Human Genetics*, 108(3), pp. 517–524. Available at: https://doi.org/10.1016/j.ajhg.2021.02.009.

Kingman, J.F., Koch, G. and Spizzichino, F. (1982) ‘Exchangeability and the evolution of large populations’, *Exchangeability in probability and statistics*, 91, p. 112.

Kingman, J. F. C. (1982) ‘On the genealogy of large populations’, *Journal of Applied Probability*. 2016/07/14 edn, 19(A), pp. 27–43. Available at: https://doi.org/10.2307/3213548.

Kingman, J.F.C. (1982) ‘The coalescent’, *Stochastic Processes and their Applications*, 13(3), pp. 235–248. Available at: https://doi.org/10.1016/0304-4149(82)90011-4.

Kislenko, A. & Tatarintseva, N. (1999) ‘The eastern Ural steppe at the end of the Stone Age. Late prehistoric exploitation of the Eurasian steppe’ 183–216

Kitajima, H., Sonoda, M. and Yamamoto, K. (2012) ‘HLA and SNP haplotype mapping in the Japanese population’, *Genes & Immunity*, 13(7), pp. 543–548. Available at: https://doi.org/10.1038/gene.2012.35.

Klejn, L.S. *et al.* (2018) ‘Discussion: Are the Origins of Indo-European Languages Explained by the Migration of the Yamnaya Culture to the West?’, *European Journal of Archaeology*, 21(1), pp. 3–17. Available at: https://doi.org/10.1017/eaa.2017.35.

Koch-Henriksen, N. *et al.* (2018) ‘Incidence of MS has increased markedly over six decades in Denmark particularly with late onset and in women’, *Neurology*, 90(22), pp. e1954–e1963. Available at: https://doi.org/10.1212/WNL.0000000000005612.

Korneliussen, T.S., Albrechtsen, A. and Nielsen, R. (2014) ‘ANGSD: Analysis of Next Generation Sequencing Data’, *BMC Bioinformatics*, 15(1), p. 356. Available at: https://doi.org/10.1186/s12859-014-0356-4.

Krause-Kyora, B. *et al.* (2018) ‘Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans’, *Nature Communications*, 9(1), p. 1569. Available at: https://doi.org/10.1038/s41467-018-03857-x.

Laland, K.N., Odling-Smee, J. and Feldman, M.W. (2001) ‘Cultural niche construction and human evolution’, *Journal of Evolutionary Biology*, 14(1), pp. 22–33. Available at: https://doi.org/10.1046/j.1420-9101.2001.00262.x.

Landis, B.C. *et al.* (2022) ‘Coronavirus Disease 2019, Eye Pain, Headache, and Beyond’, *Journal of Neuro-Ophthalmology*, 42(1). Available at: https://journals.lww.com/jneuro-ophthalmology/Fulltext/2022/03000/Coronavirus_Disease_2019,_Eye_Pain,_Headache,_and.4.aspx.

Langton, D.J. *et al.* (2021) ‘The influence of HLA genotype on the severity of COVID‐19 infection’, *HLA*, 98(1), pp. 14–22. Available at: https://doi.org/10.1111/tan.14284.

Lanz, T.V. *et al.* (2022) ‘Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM’, *Nature*, 603(7900), pp. 321–327. Available at: https://doi.org/10.1038/s41586-022-04432-7.

Lawson, D.J., van Dorp, L. and Falush, D. (2018) ‘A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots’, *Nature Communications*, 9(1), p. 3258. Available at: https://doi.org/10.1038/s41467-018-05257-7.

Lawson, D.J. (2015) ‘108Populations in Statistical Genetic Modelling and Inference’, in P. Kreager et al. (eds) *Population in the Human Sciences: Concepts, Models, Evidence*. Oxford

University Press, p. 0. Available at: https://doi.org/10.1093/acprof:oso/9780199688203.003.0004.

Lawson, D.J. *et al.* (2012) 'Inference of Population Structure using Dense Haplotype Data', *PLoS Genetics*. Edited by G.P. Copenhaver, 8(1), p. e1002453. Available at: https://doi.org/10.1371/journal.pgen.1002453.

Lazaridis, I. *et al.* (2016) 'Genomic insights into the origin of farming in the ancient Near East', *Nature*, 536(7617), pp. 419–424. Available at: https://doi.org/10.1038/nature19310.

Leitsalu, L. *et al.* (2015) 'Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu', *International Journal of Epidemiology*, 44(4), pp. 1137–1147. Available at: https://doi.org/10.1093/ije/dyt268.

Lenz, T.L. *et al.* (2016) 'Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection', *Molecular Biology and Evolution*, 33(10), pp. 2555–2564. Available at: https://doi.org/10.1093/molbev/msw127.

Leonardi, M. *et al.* (2017) 'Evolutionary Patterns and Processes: Lessons from Ancient DNA', *Systematic Biology*, 66(1), pp. e1–e29. Available at: https://doi.org/10.1093/sysbio/syw059.

Leslie, S. *et al.* (2015) 'The fine-scale genetic structure of the British population', *Nature*, 519(7543), pp. 309–314. Available at: https://doi.org/10.1038/nature14230.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/bioinformatics/btp324.

Li, H. and Durbin, R. (2011) 'Inference of human population history from individual whole-genome sequences', *Nature*, 475(7357), pp. 493–496. Available at: https://doi.org/10.1038/nature10231.

Li, H. and Stephan, W. (2006) 'Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila', *PLOS Genetics*, 2(10), p. e166. Available at: https://doi.org/10.1371/journal.pgen.0020166.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25(16), pp. 2078–2079. Available at: https://doi.org/10.1093/bioinformatics/btp352.

Li, N. and Stephens, M. (2003) 'Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data', *Genetics*, 165(4), pp. 2213–2233. Available at: https://doi.org/10.1093/genetics/165.4.2213.

Lin, W.-H.W. *et al.* (2021) 'Primary differentiated respiratory epithelial cells respond to apical measles virus infection by shedding multinucleated giant cells', *Proceedings of the National Academy of Sciences*, 118(11), p. e2013264118. Available at: https://doi.org/10.1073/pnas.2013264118.

Lindfors, K. *et al.* (2019) 'Coeliac disease', *Nature Reviews Disease Primers*, 5(1), p. 3. Available at: https://doi.org/10.1038/s41572-018-0054-z.

Lipson, M. *et al.* (2017) 'Parallel palaeogenomic transects reveal complex genetic history of early European farmers', *Nature*, 551(7680), pp. 368–372. Available at: https://doi.org/10.1038/nature24476.

Liston, A. *et al.* (2021) 'Human immune diversity: from evolution to modernity', *Nature Immunology*, 22(12), pp. 1479–1489. Available at: https://doi.org/10.1038/s41590-021-01058-1.

Liu, L. *et al.* (2013) 'Robust methods for population stratification in genome wide association studies', *BMC Bioinformatics*, 14(1), p. 132. Available at: https://doi.org/10.1186/1471-2105-14-132.

Liu, Y. *et al.* (2021) 'Insights into human history from the first decade of ancient human genomics', *Science*, 373, pp. 1479–1484. Available at: https://doi.org/10.1126/science.abi8202.

Lombard, J.E. *et al.* (2021) 'Human-to-Cattle Mycobacterium tuberculosis Complex Transmission in the United States', *Frontiers in Veterinary Science*, 8, p. 691192. Available at: https://doi.org/10.3389/fvets.2021.691192.

Loos, R.J.F. (2020) '15 years of genome-wide association studies and no signs of slowing down', *Nature Communications*, 11(1), p. 5900. Available at: https://doi.org/10.1038/s41467-020-19653-5.

Lucotte, G *et al.* (1997) 'Pattern of gradient of apolipoprotein E allele *4 frequencies in western Europe', *Human biology, 69(2), 253–262*

MacArthur, J. *et al.* (2017) 'The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)', *Nucleic Acids Research*, 45(D1), pp. D896–D901. Available at: https://doi.org/10.1093/nar/gkw1133.

Mahajan, A. *et al.* (2014) 'Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility', *Nature Genetics*, 46(3), pp. 234–244. Available at: https://doi.org/10.1038/ng.2897.

Majander, K. *et al.* (2020) ‘Ancient Bacterial Genomes Reveal a High Diversity of Treponema pallidum Strains in Early Modern Europe’, *Current Biology*, 30(19), pp. 3788-3803.e10. Available at: https://doi.org/10.1016/j.cub.2020.07.058.

Malkova, A. *et al.* (2020) ‘The opposite effect of human leukocyte antigen genotypes in sarcoidosis and tuberculosis: a narrative review of the literature’, *ERJ Open Research*, 6(3), pp. 00155–02020. Available at: https://doi.org/10.1183/23120541.00155-2020.

Manolio, T.A. *et al.* (2009) ‘Finding the missing heritability of complex diseases’, *Nature*, 461(7265), pp. 747–753. Available at: https://doi.org/10.1038/nature08494.

Maples, B.K. *et al.* (2013) ‘RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference’, *The American Journal of Human Genetics*, 93(2), pp. 278–288. Available at: https://doi.org/10.1016/j.ajhg.2013.06.020.

Marciniak, S. *et al.* (2022) ‘An integrative skeletal and paleogenomic analysis of stature variation suggests relatively reduced health for early European farmers’, *Proceedings of the National Academy of Sciences*, 119(15), p. e2106743119. Available at: https://doi.org/10.1073/pnas.2106743119.

Margaryan, A. *et al.* (2020) ‘Population genomics of the Viking world’, *Nature*, 585(7825), pp. 390–396. Available at: https://doi.org/10.1038/s41586-020-2688-8.

Marnetto, D. *et al.* (2020) ‘Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals’, *Nature Communications*, 11(1), p. 1628. Available at: https://doi.org/10.1038/s41467-020-15464-w.

Marnetto, D. *et al.* (2021) *Ancestral contributions to contemporary European complex traits*. preprint. Genomics. Available at: https://doi.org/10.1101/2021.08.03.454888.

Marnetto, D. *et al.* (2022) ‘Ancestral genomic contributions to complex traits in contemporary Europeans’, *Current Biology*, 32(6), pp. 1412-1419.e3. Available at: https://doi.org/10.1016/j.cub.2022.01.046.

Marrie, R.A. *et al.* (2018) ‘Lower prevalence of multiple sclerosis in First Nations Canadians’, *Neurology: Clinical Practice*, 8(1), pp. 33–39. Available at: https://doi.org/10.1212/CPJ.0000000000000418.

Martin, A.R. *et al.* (2019) ‘Clinical use of current polygenic risk scores may exacerbate health disparities’, *Nature Genetics*, 51(4), pp. 584–591. Available at: https://doi.org/10.1038/s41588-019-0379-x.

Martiniano, R. *et al.* (2016) ‘Genomic signals of migration and continuity in Britain before the Anglo-Saxons’, *Nature Communications*, 7(1), p. 10326. Available at: https://doi.org/10.1038/ncomms10326.

Martiniano, R. *et al.* (2017) 'The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods', *PLOS Genetics*, 13(7), p. e1006852. Available at: https://doi.org/10.1371/journal.pgen.1006852.

Martiniano, R. *et al.* (2019) *Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph*. preprint. Genomics. Available at: https://doi.org/10.1101/782755.

Mathieson, I. and Scally, A. (2020) 'What is ancestry?', *PLOS Genetics*. Edited by J. Flint, 16(3), p. e1008624. Available at: https://doi.org/10.1371/journal.pgen.1008624.

Mathieson, I. and Terhorst, J. (2022) *Direct detection of natural selection in Bronze Age Britain*. preprint. Evolutionary Biology. Available at: https://doi.org/10.1101/2022.03.14.484330.

Mathieson, I. *et al.* (2015) 'Genome-wide patterns of selection in 230 ancient Eurasians', *Nature*, 528(7583), pp. 499–503. Available at: https://doi.org/10.1038/nature16152.

Mathieson, I. *et al.* (2018) 'The genomic history of southeastern Europe', *Nature*, 555(7695), pp. 197–203. Available at: https://doi.org/10.1038/nature25778.

Mathieson, S. and Mathieson, I. (2018) '*FADS1* and the Timing of Human Adaptation to Agriculture', *Molecular Biology and Evolution*. Edited by E. Heyer, 35(12), pp. 2957–2970. Available at: https://doi.org/10.1093/molbev/msy180.

Matzaraki, V. *et al.* (2017) 'The MHC locus and genetic susceptibility to autoimmune and infectious diseases', *Genome Biology*, 18(1), p. 76. Available at: https://doi.org/10.1186/s13059-017-1207-1.

McClellan, J. and King, M.-C. (2010) 'Genetic Heterogeneity in Human Disease', *Cell*, 141(2), pp. 210–217. Available at: https://doi.org/10.1016/j.cell.2010.03.032.

McColl, H. *et al.* (2018) 'The prehistoric peopling of Southeast Asia', *Science*, 361(6397), pp. 88–92. Available at: https://doi.org/10.1126/science.aat3628.

McFadden, D. (1973) 'Conditional logit analysis of qualitative choice behavior'.

McVean, G. (2009) 'A Genealogical Interpretation of Principal Components Analysis', *PLOS Genetics*, 5(10), p. e1000686. Available at: https://doi.org/10.1371/journal.pgen.1000686.

McVean, G.A.T. and Cardin, N.J. (2005) 'Approximating the coalescent with recombination', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459), pp. 1387–1393. Available at: https://doi.org/10.1098/rstb.2005.1673.

Melchior, L. *et al.* (2010) 'Genetic Diversity among Ancient Nordic Populations', *PLOS ONE*, 5(7), p. e11898. Available at: https://doi.org/10.1371/journal.pone.0011898.

Meyer, M. and Kircher, M. (2010) 'Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing', *Cold Spring Harbor Protocols*, 2010(6), p. pdb.prot5448. Available at: https://doi.org/10.1101/pdb.prot5448.

Mittnik, A. *et al.* (2018) 'The genetic prehistory of the Baltic Sea region', *Nature Communications*, 9(1), p. 442. Available at: https://doi.org/10.1038/s41467-018-02825-9.

Monroy Kuhn, J.M., Jakobsson, M. and Günther, T. (2018) 'Estimating genetic kin relationships in prehistoric populations', *PLOS ONE*. Edited by F. Calafell, 13(4), p. e0195491. Available at: https://doi.org/10.1371/journal.pone.0195491.

Monsuur, A.J. *et al.* (2008) 'Effective Detection of Human Leukocyte Antigen Risk Alleles in Celiac Disease Using Tag Single Nucleotide Polymorphisms', *PLoS ONE*. Edited by P. Heutink, 3(5), p. e2270. Available at: https://doi.org/10.1371/journal.pone.0002270.

Mostafavi, H. *et al.* (2020) 'Variable prediction accuracy of polygenic scores within an ancestry group', *eLife*. Edited by R. Loos, M.B. Eisen, and P. O'Reilly, 9, p. e48376. Available at: https://doi.org/10.7554/eLife.48376.

Myers, T.A., Chanock, S.J. and Machiela, M.J. (2020) 'LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations', *Frontiers in Genetics*, 11, p. 157. Available at: https://doi.org/10.3389/fgene.2020.00157.

Nakatsuka, N. *et al.* (2020) 'Two genetic variants explain the association of European ancestry with multiple sclerosis risk in African-Americans', *Scientific Reports*, 10(1), p. 16902. Available at: https://doi.org/10.1038/s41598-020-74035-7.

Narasimhan, V.M. *et al.* (2017) 'Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes', *Nature Communications*, 8(1), p. 303. Available at: https://doi.org/10.1038/s41467-017-00323-y.

Nelson, R.M. *et al.* (2017) 'Genomewide analysis of admixture and adaptation in the Africanized honeybee', *Molecular Ecology*, 26(14), pp. 3603–3617. Available at: https://doi.org/10.1111/mec.14122.

Nielsen, R. (2005) 'Molecular Signatures of Natural Selection', *Annual Review of Genetics*, 39(1), pp. 197–218. Available at: https://doi.org/10.1146/annurev.genet.39.073003.112420.

Nikitin, A.G. *et al.* (2019) 'Interactions between earliest Linearbandkeramik farmers and central European hunter gatherers at the dawn of European Neolithization', *Scientific Reports*, 9(1), p. 19544. Available at: https://doi.org/10.1038/s41598-019-56029-2.

Ning, C. *et al.* (2019) ‘Ancient Genomes Reveal Yamnaya-Related Ancestry and a Potential Source of Indo-European Speakers in Iron Age Tianshan’, *Current Biology*, 29(15), pp. 2526-2532.e4. Available at: https://doi.org/10.1016/j.cub.2019.06.044.

Nordborg, M. (2001) ‘Coalescent theory. Handbook of statistical genetics’, *Balding D, Bishop M& Cannings C*, pp. 179–212.

Olalde, I. and Posth, C. (2020) ‘Latest trends in archaeogenetic research of west Eurasians’, *Genetics of Human Origin*, 62, pp. 36–43. Available at: https://doi.org/10.1016/j.gde.2020.05.021.

Olalde, I. *et al.* (2014) ‘Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European’, *Nature*, 507(7491), pp. 225–228. Available at: https://doi.org/10.1038/nature12960.

Olalde, I. *et al.* (2018) ‘The Beaker phenomenon and the genomic transformation of northwest Europe’, *Nature*, 555(7695), pp. 190–196. Available at: https://doi.org/10.1038/nature25738.

Olsson, T., Barcellos, L.F. and Alfredsson, L. (2017) ‘Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis’, *Nature Reviews Neurology*, 13(1), pp. 25–36. Available at: https://doi.org/10.1038/nrneurol.2016.187.

Oriá, R.B. *et al.* (2007) ‘Role of apolipoprotein E4 in protecting children against early childhood diarrhea outcomes and implications for later development’, *Medical Hypotheses*, 68(5), pp. 1099–1107. Available at: https://doi.org/10.1016/j.mehy.2006.09.036.

Orlando, L. *et al.* (2021) ‘Ancient DNA analysis’, *Nature Reviews Methods Primers*, 1(1), p. 14. Available at: https://doi.org/10.1038/s43586-020-00011-0.

O’Connor, L.J. *et al.* (2019) ‘Extreme Polygenicity of Complex Traits Is Explained by Negative Selection’, *The American Journal of Human Genetics*, 105(3), pp. 456–476. Available at: https://doi.org/10.1016/j.ajhg.2019.07.003.

O’Dushlaine, C.T. *et al.* (2010) ‘Population structure and genome-wide patterns of variation in Ireland and Britain’, *European Journal of Human Genetics*, 18(11), pp. 1248–1254. Available at: https://doi.org/10.1038/ejhg.2010.87.

Page, A.E. *et al.* (2016) ‘Reproductive trade-offs in extant hunter-gatherers suggest adaptive mechanism for the Neolithic expansion’, *Proceedings of the National Academy of Sciences*, 113(17), pp. 4694–4699. Available at: https://doi.org/10.1073/pnas.1524031113.

Patterson, N. *et al.* (2012) ‘Ancient Admixture in Human History’, *Genetics*, 192(3), pp. 1065–1093. Available at: https://doi.org/10.1534/genetics.112.145037.

Patterson, N. *et al.* (2022) ‘Large-scale migration into Britain during the Middle to Late Bronze Age’, *Nature*, 601(7894), pp. 588–594. Available at: https://doi.org/10.1038/s41586-021-04287-4.

Pearson, J.F. *et al.* (2014) ‘Multiple sclerosis in New Zealand Māori’, *Multiple Sclerosis Journal*, 20(14), pp. 1892–1895. Available at: https://doi.org/10.1177/1352458514535130.

Pedersen, C.-E.T. *et al.* (2017) ‘The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit’, *Genetics*, 205(2), pp. 787–801. Available at: https://doi.org/10.1534/genetics.116.193821.

Pedregosa, F. *et al.* (no date) ‘Scikit-learn: Machine Learning in Python’, *MACHINE LEARNING IN PYTHON*, p. 6.

Peschl, P. *et al.* (2017) ‘Myelin Oligodendrocyte Glycoprotein: Deciphering a Target in Inflammatory Demyelinating Diseases’, *Frontiers in Immunology*, 8, p. 529. Available at: https://doi.org/10.3389/fimmu.2017.00529.

Petkeviciene, J. *et al.* (2012) ‘Associations between Apolipoprotein E Genotype, Diet, Body Mass Index, and Serum Lipids in Lithuanian Adult Population’, *PLoS ONE*. Edited by O.Y. Gorlova, 7(7), p. e41525. Available at: https://doi.org/10.1371/journal.pone.0041525.

Pickrell, J.K. and Reich, D. (2014) ‘Toward a new history and geography of human genes informed by ancient DNA’, *Trends in Genetics*, 30(9), pp. 377–389. Available at: https://doi.org/10.1016/j.tig.2014.07.007.

Pierron, D. *et al.* (2018) ‘Strong selection during the last millennium for African ancestry in the admixed population of Madagascar’, *Nature Communications*, 9(1), p. 932. Available at: https://doi.org/10.1038/s41467-018-03342-5.

Pillai, N.E. *et al.* (2014) ‘Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations’, *Human Molecular Genetics*, 23(16), pp. 4443–4451. Available at: https://doi.org/10.1093/hmg/ddu149.

Plomin, R., DeFries, J.C. and Loehlin, J.C. (1977) ‘Genotype-environment interaction and correlation in the analysis of human behavior.’, *Psychological Bulletin*, 84, pp. 309–322. Available at: https://doi.org/10.1037/0033-2909.84.2.309.

Pontremoli, C. *et al.* (2019) ‘Possible European Origin of Circulating Varicella Zoster Virus Strains’, *The Journal of Infectious Diseases*, p. jiz227. Available at: https://doi.org/10.1093/infdis/jiz227.

Price, A.L. *et al.* (2006) ‘Principal components analysis corrects for stratification in genome-wide association studies’, *Nature Genetics*, 38(8), pp. 904–909. Available at: https://doi.org/10.1038/ng1847.

Price, A.L. *et al.* (2009) ‘Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations’, *PLoS Genetics*. Edited by J.K. Pritchard, 5(6), p. e1000519. Available at: https://doi.org/10.1371/journal.pgen.1000519.

Price, T. D. (2000) ‘Europe’s First Farmers’, Cambridge University Press

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) ‘Inference of Population Structure Using Multilocus Genotype Data’, *Genetics*, 155(2), pp. 945–959. Available at: https://doi.org/10.1093/genetics/155.2.945.

Privé, F., Arbel, J. and Vilhjálmsson, B.J. (2020) ‘LDpred2: better, faster, stronger’, *Bioinformatics*, 36(22–23), pp. 5424–5431. Available at: https://doi.org/10.1093/bioinformatics/btaa1029.

Prohaska, A. *et al.* (2019) ‘Human Disease Variation in the Light of Population Genomics’, *Cell*, 177(1), pp. 115–131. Available at: https://doi.org/10.1016/j.cell.2019.01.052.

Purcell, S. *et al.* (2007) ‘PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses’, *The American Journal of Human Genetics*, 81(3), pp. 559–575. Available at: https://doi.org/10.1086/519795.

Racimo, F., Berg, J.J. and Pickrell, J.K. (2018) ‘Detecting Polygenic Adaptation in Admixture Graphs’, *Genetics*, 208(4), pp. 1565–1584. Available at: https://doi.org/10.1534/genetics.117.300489.

Racimo, F. *et al.* (2019) *A geostatistical approach to modelling human Holocene migrations in Europe using ancient DNA*. preprint. Evolutionary Biology. Available at: https://doi.org/10.1101/826149.

Rall, S.C., Weisgraber, K.H. and Mahley, R.W. (1982) ‘Human apolipoprotein E. The complete amino acid sequence.’, *Journal of Biological Chemistry*, 257(8), pp. 4171–4178. Available at: https://doi.org/10.1016/S0021-9258(18)34702-1.

Ramagopalan, S.V. and Ebers, G.C. (2009) ‘Multiple sclerosis: major histocompatibility complexity and antigen presentation’, *Genome Medicine*, 1(11), p. 105. Available at: https://doi.org/10.1186/gm105.

Rascovan, N. *et al.* (2019) ‘Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline’, *Cell*, 176(1–2), pp. 295-305.e10. Available at: https://doi.org/10.1016/j.cell.2018.11.005.

Rasmussen, M.D. *et al.* (2014) ‘Genome-Wide Inference of Ancestral Recombination Graphs’, *PLOS Genetics*, 10(5), p. e1004342. Available at: https://doi.org/10.1371/journal.pgen.1004342.

Rasmussen, S. *et al.* (2015) 'Early Divergent Strains of Yersinia pestis in Eurasia 5,000 Years Ago', *Cell*, 163(3), pp. 571–582. Available at: https://doi.org/10.1016/j.cell.2015.10.009.

Raudvere, U. *et al.* (2019) 'g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)', *Nucleic Acids Research*, 47(W1), pp. W191–W198. Available at: https://doi.org/10.1093/nar/gkz369.

Raychaudhuri, S. *et al.* (2012) 'Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis', *Nature Genetics*, 44(3), pp. 291–296. Available at: https://doi.org/10.1038/ng.1076.

Reales, G. *et al.* (2017) 'A tale of agriculturalists and hunter-gatherers: Exploring the thrifty genotype hypothesis in native South Americans', *American Journal of Physical Anthropology*, 163(3), pp. 591–601. Available at: https://doi.org/10.1002/ajpa.23233.

Rivas, M.A. *et al.* (2018) 'Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population', *PLOS Genetics*. Edited by S.M. Williams, 14(5), p. e1007329. Available at: https://doi.org/10.1371/journal.pgen.1007329.

Roach, J.C. *et al.* (2010) 'Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing', *Science*, 328(5978), pp. 636–639. Available at: https://doi.org/10.1126/science.1186802.

Robson, M.I. and Mundlos, S. (2019) 'Jumping retroviruses nudge TADs apart', *Nature Genetics*, 51(9), pp. 1304–1305. Available at: https://doi.org/10.1038/s41588-019-0491-y.

Rosenberg, N.A. *et al.* (2002) 'Genetic Structure of Human Populations', *Science*, 298(5602), pp. 2381–2385. Available at: https://doi.org/10.1126/science.1078311.

Rosenstock, E. *et al.* (2019) 'Human stature in the Near East and Europe ca. 10,000–1000 BC: its spatiotemporal development in a Bayesian errors-in-variables model', *Archaeological and Anthropological Sciences*, 11(10), pp. 5657–5690. Available at: https://doi.org/10.1007/s12520-019-00850-3.

Rubinacci, S. *et al.* (2021) 'Efficient phasing and imputation of low-coverage sequencing data using large reference panels', *Nature Genetics*, 53(1), pp. 120–126. Available at: https://doi.org/10.1038/s41588-020-00756-0.

Sabeti, P.C. *et al.* (2002) 'Detecting recent positive selection in the human genome from haplotype structure', *Nature*, 419(6909), pp. 832–837. Available at: https://doi.org/10.1038/nature01140.

Sabin, S. *et al.* (2020) ‘A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex’, *Genome Biology*, 21(1), p. 201. Available at: https://doi.org/10.1186/s13059-020-02112-1.

Safiri, S. *et al.* (2019) ‘Global, regional and national burden of rheumatoid arthritis 1990–2017: a systematic analysis of the Global Burden of Disease study 2017’, *Annals of the Rheumatic Diseases*, 78(11), pp. 1463–1471. Available at: https://doi.org/10.1136/annrheumdis-2019-215920.

Salter-Townshend, M. and Myers, S. (no date) ‘Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups’, p. 21.

Sankararaman, S. *et al.* (2008) ‘Estimating Local Ancestry in Admixed Populations’, *The American Journal of Human Genetics*, 82(2), pp. 290–303. Available at: https://doi.org/10.1016/j.ajhg.2007.09.022.

Sarmanova, A., Morris, T. and Lawson, D.J. (2020) *Population stratification in GWAS meta-analysis should be standardized to the best available reference datasets*. preprint. Genetics. Available at: https://doi.org/10.1101/2020.09.03.281568.

Scally, A. and Durbin, R. (2012) ‘Revising the human mutation rate: implications for understanding human evolution’, *Nature Reviews Genetics*, 13(10), pp. 745–753. Available at: https://doi.org/10.1038/nrg3295.

Schaffner, S.F. *et al.* (2005) ‘Calibrating a coalescent simulation of human genome sequence variation’, *Genome Research*, 15(11), pp. 1576–1583. Available at: https://doi.org/10.1101/gr.3709305.

Schiffels, S. *et al.* (2016) ‘Iron Age and Anglo-Saxon genomes from East England reveal British migration history’, *Nature Communications*, 7(1), p. 10408. Available at: https://doi.org/10.1038/ncomms10408.

Schubert, M., Lindgreen, S. and Orlando, L. (2016) ‘AdapterRemoval v2: rapid adapter trimming, identification, and read merging’, *BMC Research Notes*, 9(1), p. 88. Available at: https://doi.org/10.1186/s13104-016-1900-2.

Schubert, R., Andaleon, A. and Wheeler, H.E. (2020) ‘Comparing local ancestry inference models in populations of two- and three-way admixture’, *PeerJ*, 8, pp. e10090–e10090. Available at: https://doi.org/10.7717/peerj.10090.

Schurz, H. *et al.* (2022) ‘Multi-ancestry meta-analysis of host genetic susceptibility to tuberculosis identifies shared genetic architecture’, *medRxiv*, p. 2022.08.26.22279009. Available at: https://doi.org/10.1101/2022.08.26.22279009.

Scorrano, G. *et al.* (2021) 'The genetic and cultural impact of the Steppe migration into Europe', *Annals of Human Biology*, 48(3), pp. 223–233. Available at: https://doi.org/10.1080/03014460.2021.1942984.

Scott, A. *et al.* (2022) 'Emergence and intensification of dairying in the Caucasus and Eurasian steppes', *Nature Ecology & Evolution*, 6(6), pp. 813–822. Available at: https://doi.org/10.1038/s41559-022-01701-6.

Sheridan, A. and Schier, W. (2015) 'The Oxford Handbook of Neolithic Europe, edited by Chris Fowler, Jan Harding, and Daniela Hofmann', *Archaeological Journal*, 173, pp. 1–2. Available at: https://doi.org/10.1080/00665983.2015.1112681.

Sheridan, J. A. (2010) 'Landscapes in Transition', Oxbow (eds Finlayson, B. & Warren, G.) 89–105

Shinde, V. *et al.* (2019) 'An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers', *Cell*, 179(3), pp. 729-735.e10. Available at: https://doi.org/10.1016/j.cell.2019.08.048.

Shriner, D. (2013) 'Overview of Admixture Mapping', *Current Protocols in Human Genetics*, 76(1). Available at: https://doi.org/10.1002/0471142905.hg0123s76.

Shringarpure, S.S. *et al.* (2016) 'Efficient analysis of large datasets and sex bias with ADMIXTURE', *BMC Bioinformatics*, 17(1), p. 218. Available at: https://doi.org/10.1186/s12859-016-1082-x.

Sikora, M. *et al.* (2017) 'Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers', p. 5.

Sikora, M. *et al.* (2019) 'The population history of northeastern Siberia since the Pleistocene', *Nature*, 570(7760), pp. 182–188. Available at: https://doi.org/10.1038/s41586-019-1279-z.

Singh, P. *et al.* (2018) 'Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis', *Clinical Gastroenterology and Hepatology*, 16(6), pp. 823-836.e2. Available at: https://doi.org/10.1016/j.cgh.2017.06.037.

Skoglund, P. and Mathieson, I. (2018) 'Ancient Genomics of Modern Humans: The First Decade', *Annual Review of Genomics and Human Genetics*, 19(1), pp. 381–404. Available at: https://doi.org/10.1146/annurev-genom-083117-021749.

Skoglund, P. *et al.* (2012) 'Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe', *Science*, 336(6080), pp. 466–469. Available at: https://doi.org/10.1126/science.1216304.

Skoglund, P. *et al.* (2014) ‘Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers’, *Science*, 344(6185), pp. 747–750. Available at: https://doi.org/10.1126/science.1253448.

Slatkin, M. (2008) ‘Linkage disequilibrium — understanding the evolutionary past and mapping the medical future’, *Nature Reviews Genetics*, 9(6), pp. 477–485. Available at: https://doi.org/10.1038/nrg2361.

Slim, L. *et al.* (2022) ‘A systematic analysis of gene–gene interaction in multiple sclerosis’, *BMC Medical Genomics*, 15(1), p. 100. Available at: https://doi.org/10.1186/s12920-022-01247-3.

Smith, G.D. *et al.* (2009) ‘Lactase persistence-related genetic variant: population substructure and health outcomes’, *European Journal of Human Genetics*, 17(3), pp. 357–367. Available at: https://doi.org/10.1038/ejhg.2008.156.

Smolen, J.S. *et al.* (2018) ‘Rheumatoid arthritis’, *Nature Reviews Disease Primers*, 4(1), p. 18001. Available at: https://doi.org/10.1038/nrdp.2018.1.

Sohail, M. *et al.* (2019) ‘Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies’, *eLife*. Edited by M. Nordborg et al., 8, p. e39702. Available at: https://doi.org/10.7554/eLife.39702.

Spanish Consortium on the Genetics of Coeliac Disease (CEGEC) *et al.* (2011) ‘Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease’, *Nature Genetics*, 43(12), pp. 1193–1201. Available at: https://doi.org/10.1038/ng.998.

Speidel, L. *et al.* (2019) ‘A method for genome-wide genealogy estimation for thousands of samples’, *Nature Genetics*, 51(9), pp. 1321–1329. Available at: https://doi.org/10.1038/s41588-019-0484-x.

Spyrou, M.A. *et al.* (2018) ‘Analysis of 3800-year-old Yersinia pestis genomes suggests Bronze Age origin for bubonic plague’, *Nature Communications*, 9(1), p. 2234. Available at: https://doi.org/10.1038/s41467-018-04550-9.

Stern, A.J., Wilton, P.R. and Nielsen, R. (2019) ‘An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data’, *PLOS Genetics*. Edited by R.D. Hernandez, 15(9), p. e1008384. Available at: https://doi.org/10.1371/journal.pgen.1008384.

Stern, A.J. *et al.* (2021) ‘Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies’, *The American Journal of Human Genetics*, 108(2), pp. 219–239. Available at: https://doi.org/10.1016/j.ajhg.2020.12.005.

Strittmatter, W.J. *et al.* (1993) ‘Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease.’, *Proceedings of the National Academy of Sciences*, 90(5), pp. 1977–1981. Available at: https://doi.org/10.1073/pnas.90.5.1977.

Sudlow, C. *et al.* (2015) ‘UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age’, *PLOS Medicine*, 12(3), p. e1001779. Available at: https://doi.org/10.1371/journal.pmed.1001779.

Svensson, E. *et al.* (2021) ‘Genome of Peştera Muierii skull shows high diversity and low mutational load in pre-glacial Europe’, *Current Biology*, 31(14), pp. 2973-2983.e9. Available at: https://doi.org/10.1016/j.cub.2021.04.045.

Tajima, F. (1983) ‘EVOLUTIONARY RELATIONSHIP OF DNA SEQUENCES IN FINITE POPULATIONS’, *Genetics*, 105(2), pp. 437–460. Available at: https://doi.org/10.1093/genetics/105.2.437.

Tam, V. *et al.* (2019) ‘Benefits and limitations of genome-wide association studies’, *Nature Reviews Genetics*, 20(8), pp. 467–484. Available at: https://doi.org/10.1038/s41576-019-0127-1.

Tervi, A. *et al.* (2022) ‘Large registry based analysis of genetic predisposition to tuberculosis identifies genetic risk factors at HLA’, *Human Molecular Genetics*, p. ddac212. Available at: https://doi.org/10.1093/hmg/ddac212.

The 1000 Genomes Project Consortium *et al.* (2015) ‘A global reference for human genetic variation’, *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

the International Multiple Sclerosis Genetics Consortium (2015) ‘Class II HLA interactions modulate genetic risk for multiple sclerosis’, *Nature Genetics*, 47(10), pp. 1107–1113. Available at: https://doi.org/10.1038/ng.3395.

the RACI consortium *et al.* (2014) ‘Genetics of rheumatoid arthritis contributes to biology and drug discovery’, *Nature*, 506(7488), pp. 376–381. Available at: https://doi.org/10.1038/nature12873.

Thompson, A.J. *et al.* (2018) ‘Multiple sclerosis’, *The Lancet*, 391(10130), pp. 1622–1636. Available at: https://doi.org/10.1016/S0140-6736(18)30481-1.

Thorndike, R.L. (1953) ‘Who belongs in the family?’, *Psychometrika*, 18(4), pp. 267–276. Available at: https://doi.org/10.1007/BF02289263.

Thuesen, N.H. *et al.* (2022) *Benchmarking freely available human leukocyte antigen typing algorithms across varying genes, coverages and typing resolutions*. preprint. Bioinformatics. Available at: https://doi.org/10.1101/2022.06.28.497888.

Tian, C. *et al.* (2017) ‘Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections’, *Nature Communications*, 8(1), p. 599. Available at: https://doi.org/10.1038/s41467-017-00257-5.

Tinbergen, N. (1963) ‘On aims and methods of Ethology’, *Zeitschrift für Tierpsychologie*, 20(4), pp. 410–433. Available at: https://doi.org/10.1111/j.1439-0310.1963.tb01161.x.

Trehearne, A. (2016) ‘Genetics, lifestyle and environment: UK Biobank is an open access resource following the lives of 500,000 participants to improve the health of future generations’, *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 59(3), pp. 361–367. Available at: https://doi.org/10.1007/s00103-015-2297-0.

Trumble, B.C. and Finch, C.E. (2019) ‘The Exposome in Human Evolution: From Dust to Diesel’, *The Quarterly Review of Biology*, 94(4), pp. 333–394. Available at: https://doi.org/10.1086/706768.

Trumble, B.C. *et al.* (2017) ‘Apolipoprotein E4 is associated with improved cognitive function in Amazonian forager-horticulturalists with a high parasite burden’, *The FASEB Journal*, 31(4), pp. 1508–1515. Available at: https://doi.org/10.1096/fj.201601084R.

Tuminello, E.R. and Han, S.D. (2011) ‘The Apolipoprotein E Antagonistic Pleiotropy Hypothesis: Review and Recommendations’, *International Journal of Alzheimer’s Disease*, 2011, pp. 1–12. Available at: https://doi.org/10.4061/2011/726197.

van Eden, W. *et al.* (2002) ‘Balancing the immune system: Th1 and Th2’, *Annals of the Rheumatic Diseases*, 61(Supplement 2), pp. 25ii–2528. Available at: https://doi.org/10.1136/ard.61.suppl_2.ii25.

Vilhjálmsson, B.J. *et al.* (2015) ‘Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores’, *The American Journal of Human Genetics*, 97(4), pp. 576–592. Available at: https://doi.org/10.1016/j.ajhg.2015.09.001.

Virtanen, P. *et al.* (2020) ‘SciPy 1.0: fundamental algorithms for scientific computing in Python’, *Nature Methods*, 17(3), pp. 261–272. Available at: https://doi.org/10.1038/s41592-019-0686-2.

Visscher, P.M. *et al.* (2017) ‘10 Years of GWAS Discovery: Biology, Function, and Translation’, *The American Journal of Human Genetics*, 101(1), pp. 5–22. Available at: https://doi.org/10.1016/j.ajhg.2017.06.005.

Vitti, J.J., Grossman, S.R. and Sabeti, P.C. (2013) ‘Detecting Natural Selection in Genomic Data’, *Annual Review of Genetics*, 47(1), pp. 97–120. Available at: https://doi.org/10.1146/annurev-genet-111212-133526.

Wakeley, J. (1996) ‘The Variance of Pairwise Nucleotide Differences in Two Populations with Migration’, *Theoretical Population Biology*, 49(1), pp. 39–57. Available at: https://doi.org/10.1006/tpbi.1996.0002.

Wakeley, J. (2009) ‘Coalescent Theory: An Introduction’, 58. Available at: https://doi.org/10.1093/schbul/syp004.

Wakeley, J. (2020) ‘Developments in coalescent theory from single loci to chromosomes’, *Theoretical Population Biology*, 133, pp. 56–64. Available at: https://doi.org/10.1016/j.tpb.2020.02.002.

Wallin, M.T. *et al.* (2019) ‘The prevalence of MS in the United States: A population-based estimate using health claims data’, *Neurology*, 92(10), pp. e1029–e1040. Available at: https://doi.org/10.1212/WNL.0000000000007035.

Walton, C. *et al.* (2020) ‘Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition’, *Multiple Sclerosis Journal*, 26(14), pp. 1816–1821. Available at: https://doi.org/10.1177/1352458520970841.

Wang, J.H. *et al.* (2011) ‘Modeling the cumulative genetic risk for multiple sclerosis from genome-wide association data’, *Genome Medicine*, 3(1), p. 3. Available at: https://doi.org/10.1186/gm217.

Weissensteiner, H. *et al.* (2016) ‘HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing’, *Nucleic Acids Research*, 44(W1), pp. W58–W63. Available at: https://doi.org/10.1093/nar/gkw233.

Wendel-Haga, M. and Celius, E.G. (2017) ‘Is the hygiene hypothesis relevant for the risk of multiple sclerosis?’, *Acta Neurologica Scandinavica*, 136, pp. 26–30. Available at: https://doi.org/10.1111/ane.12844.

Wilde, S. *et al.* (2014) ‘Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y’, *Proceedings of the National Academy of Sciences*, 111(13), pp. 4832–4837. Available at: https://doi.org/10.1073/pnas.1316513111.

Wiuf, C. and Hein, J. (1999) ‘Recombination as a Point Process along Sequences’, *Theoretical Population Biology*, 55(3), pp. 248–259. Available at: https://doi.org/10.1006/tpbi.1998.1403.

Wright, S. (1931) ‘EVOLUTION IN MENDELIAN POPULATIONS’, *Genetics*, 16(2), pp. 97–159. Available at: https://doi.org/10.1093/genetics/16.2.97.

Wu, J., Liu, Y. and Zhao, Y. (2021) ‘Systematic Review on Local Ancestor Inference From a Mathematical and Algorithmic Perspective’, *Frontiers in Genetics*, 12. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2021.639877.

Wu, Y. *et al.* (2022) 'Admixture mapping of anthropometric traits in the Black Women's Health Study: evidence of a shared African ancestry component with birth weight and type 2 diabetes', *Journal of Human Genetics*, 67(6), pp. 331–338. Available at: https://doi.org/10.1038/s10038-022-01010-7.

Yan, Z.-H. (2012) 'Relationship between *HLA-DR* gene polymorphisms and outcomes of hepatitis B viral infections: A meta-analysis', *World Journal of Gastroenterology*, 18(24), p. 3119. Available at: https://doi.org/10.3748/wjg.v18.i24.3119.

Yang, M.A. and Fu, Q. (2018) 'Insights into Modern Human Prehistory Using Ancient Genomes', *Trends in Genetics*, 34(3), pp. 184–196. Available at: https://doi.org/10.1016/j.tig.2017.11.008.

Yang, Y. and Lawson, D. (2022) 'HTRX: an R package for learning non-contiguous haplotypes associated with a phenotype', *bioRxiv*, p. 2022.11.29.518395. Available at: https://doi.org/10.1101/2022.11.29.518395.

Yengo, L. *et al.* (2022) 'A saturated map of common genetic variants associated with human height', *Nature*, 610(7933), pp. 704–712. Available at: https://doi.org/10.1038/s41586-022-05275-y.

Young, A.I. (2019) 'Solving the missing heritability problem', *PLOS Genetics*, 15(6), p. e1008222. Available at: https://doi.org/10.1371/journal.pgen.1008222.

Zaykin, D.V. *et al.* (2002) 'Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals', *Human Heredity*, 53(2), pp. 79–91. Available at: https://doi.org/10.1159/000057986.

Zerboni, L. *et al.* (2014) 'Molecular mechanisms of varicella zoster virus pathogenesis', *Nature Reviews Microbiology*, 12(3), pp. 197–210. Available at: https://doi.org/10.1038/nrmicro3215.

Zhernakova, A. *et al.* (2010) 'Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection', *The American Journal of Human Genetics*, 86(6), pp. 970–977. Available at: https://doi.org/10.1016/j.ajhg.2010.05.004.

Zvelebil, M. & Rowley-Conwy, P. (1984) 'Transition to farming in Northern Europe: A hunter-gatherer perspective', Norwegian Archaeological Review 17, 104–128

†The International HapMap Consortium (2003) 'The International HapMap Project', *Nature*, 426(6968), pp. 789–796. Available at: https://doi.org/10.1038/nature02168.