Title
Linking Human Evolutionary History to Phenotypic Variation

Author
Durvasula, Arun Kumar

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Linking Human Evolutionary History to Phenotypic Variation

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Human Genetics

by

Arun Kumar Durvasula

2021

ABSTRACT OF THE DISSERTATION

Linking Human Evolutionary History to

Phenotypic Variation

by

Arun Kumar Durvasula

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2021

Professor Kirk Edward Lohmueller, Co-chair

Professor Sriram Sankararaman, Co-chair

A central question in genetics asks how genetic variation influences phenotypic variation. The distribution of genetic variation in a population is reflective of the evolutionary forces that shape and maintain genetic diversity such as mutation, natural selection, and genetic drift. In turn, this genetic variation affects molecular phenotypes like gene expression and eventually leads to variation in complex traits. In my dissertation, I develop statistical methods and apply computational approaches to understand these dynamics in human populations. In the first chapter, I describe a statistical model for detecting the presence of archaic haplotypes in modern human populations without having access to a reference archaic genome. I apply this method to the genomes of individuals from Europe and find that I can recover segments of DNA inherited from Neanderthals as a result of archaic admixture. In the second chapter, I apply this method to the genomes of individuals from several African populations and find that approximately 7% of the genomes are inherited from an archaic species. Modeling of the site frequency spectrum suggests that the presence of these haplotypes is best explained by admixture with an unknown archaic hominin species. In the final chapter, I focus on the more recent history of humans and the genetic architecture of complex traits. In particular, I find that a substantial portion of the genetic architecture is population specific, which limits our ability to transfer phenotype predictions across populations.

The dissertation of Arun Kumar Durvasula is approved.

Jason Ernst

Nelson B Freimer

Kirk Edward Lohmueller, Committee Co-chair

Sriram Sankararaman, Committee Co-chair

University of California, Los Angeles

2021

This work is dedicated to my mother, my father, and my brother.

Table of contents

Acknowledgments vi

Vita ix

Introduction 1

Chapter 1: A statistical model for reference-free inference of archaic local ancestry 7

Chapter 2: Recovering signals of ghost archaic introgression in African populations 26

Chapter 3: Negative selection on complex traits limits genetic risk prediction accuracy between populations 36

List of Figures

Figure 1.1 11

Figure 1.2 13

Figure 1.3 14

Figure 1.4 15

Figure 1.5 17

Figure 2.1 28

Figure 2.2 29

Figure 2.3 30

Figure 3.1 40

Figure 3.2 41

Figure 3.3 42

Figure 3.4 44

Figure 3.5 45

List of Tables

Table 1.1 11

Table 2.1 30

Table 3.1 43

ACKNOWLEDGMENTS

I would like to acknowledge help and support from many people including friends, family, coworkers, and advisors during the course of my PhD and the time that led up to it. First, I would like to thank my parents, my brother, Anand, and my cousins Ramesh, Nitya, and Apu for being constant sources of encouragement. I have learned a lot from my labmates throughout the years, including Alec Chiu, Ali Pazokitoroudi, Andrea Fulgione, Annabel Beichman, April Wei, Ariel Wu, Bernard Kim, Boyang Fu, Chris Kyriazis, Chris Robles, Christian Huber, Clare Marsden, Eduardo Amorim, Erin Molloy, Gustavo Valadares Barroso, Jacqueline Robinson, Jazlyn Mooney, Jesse Garcia, Pádraic Flood, Rob Brown, Ruth Johnson, Tanya Phung, Tim Beissinger, Tina Del Carpio, Tyler Kent, Xinjun Zhang, and Ying Zhen. In addition, I am grateful to friends and colleagues I have met outside of UCLA including Aaron Stern, Ariel Gewirtz, Arjun Biddanda, Chloe Robbins, Joseph Marcus, John Novembre, Joshua Schraiber, Kelsey Johnson, Robin Burns, and Sohini Ramachandran. I am also grateful to the genetics community at UCLA, especially Malika Kumar Freund, Kathy Burch, Juan De La Hoz, Bogdan Pasaniuc. I would like to thank Katherine Deck for her constant support over the past 5 years. I am grateful to Alex Chubick, James Boocock, and Colin Farrell for their friendship. I would like to thank my committee members, Jason Ernst and Nelson Freimer for their encouragement. My undergraduate advisor, Jeff Ross-Ibarra introduced me to the wonderful world of scientific research. My advisor Angela Hancock provided an excellent environment to explore my interests after my undergraduate degree as well as the opportunity to spend time in Vienna and Cologne. Finally, my PhD advisors Kirk Lohmueller and Sriram Sankararaman have provided endless amounts of support, advice, and guidance over the past five years.

Permissions: Chapter 1 is a version of Durvasula A, Sankararaman S (2019) A statistical model for reference-free inference of archaic local ancestry. PLoS Genet 15(5): e1008175.

Chapter 2 is a version of Durvasula A, Sankararaman S (2020) Recovering signals of ghost archaic introgression in African populations. Science Advances 6, eaax5097.

Chapter 3 is a version of Durvasula A, Lohmueller KE (2021) Negative selection on complex traits limits genetic risk prediction accuracy between populations. American Journal of Human Genetics 108 (4), 620-631.

All reprints are used here with permission from the authors and under the terms of their respective distribution licenses. Chapters 1 and 2 were supervised by Sriram Sankararaman and chapter 3 was supervised by Kirk Edward Lohmueller.

Funding: This work was supported a Graduate Research Fellowship from the National Science Foundation to AD (DGE-1650604). Additional support was provided by the NIH (R35GM119856 to KEL, R00GM111744 to SS, and R35GM125055 to SS). In addition, SS was supported by a grant from the Alfred P. Sloan Foundation and a gift from the Okawa Foundation.

VITA

EDUCATION

BS Biotechnology, University of California, Davis 2015

PUBLICATIONS

Boocock J, Sadhu MJ, **Durvasula A**, Bloom JS, Kruglyak L. Ancient balancing selection maintains incompatible versions of the galactose pathway in yeast. *Science*.

Durvasula A, Sankararaman S. Recovering signals of ghost archaic admixture in the genomes of present-day Africans (2020). *Science Advances*.

Durvasula A, Sankararaman S. A statistical model for reference-free inference of local archaic ancestry (2019). *PLoS Genetics*.

Huber CD*, **Durvasula A***, Hancock AM, Lohmueller KE. Gene expression drives the evolution of dominance (2018). *Nature Communications*.

Schumer M, Xu C, Powell D, **Durvasula A**, Skov L, Holland C, Sankararaman S, Andolfatto P, Rosenthal G, Przeworski M. Natural selection interacts with the local recombination rate to shape the evolution of hybrid genomes (2018). *Science*.

Schweizer RM, **Durvasula A**, Smith J, Vohr SH, Stahler DR, Galaverni M, Thalmann O, Smith D, Randi E, Green RE, Lohmueller KE, Novembre J, Wayne RK. The evolutionary history of a selectively swept coat color and immunity locus in North American wolves (2018). *Molecular Biology and Evolution*.

Durvasula A*, Fulgione A*, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Alonso-

Blanco C, Burbano HA, Pico FX, Tsuchimatsu T, Hancock AM. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana* (2017). *Proceedings of the National Academy of Sciences*.

Durvasula A*, Hoffman PJ*, Kent TV, Liu C, Kono TJY, Morrell PL, Ross-Ibarra J. (2016), angsd-wrapper: utilities for analysing next-generation sequencing data. *Mol Ecol Resour*.

Beissinger T, Wang L, Crosby K, **Durvasula A**, Hufford M, Ross-Ibarra J. Recent demography drives changes in linked selection across the maize genome (2016). *Nature Plants*.

Introduction

Substantial genetic variation is created every generation because of errors in the replication of DNA by polymerase enzymes. Some of these mutations will affect the organism's phenotype, but most will be evolutionarily neutral [1]. That is, they will have no effect on a measurable phenotype or on the organism's fitness. Those mutations that have the strongest phenotypic effect will tend to be removed by natural selection if they are deleterious or driven to fixation if they are beneficial. Mutations that are not selected out of the population will leave a trace in the genome as DNA is continually passed on from one generation to the next. A major goal of population genetics is to learn about the history of a species or population by studying these mutations.

On the other hand, those mutations that are not neutral will have a biological impact on the organism. A major goal of statistical genetics is to predict the phenotype of an individual given the genetic mutations that an individual has. This work has implications for plant and animal breeding, where breeders can choose the individuals to cross after predicting the yield or other economically valuable traits for their progeny[2]. This work also has implications for human health, where susceptibility to disease can be predicted before the onset of the disease and lifestyle or other interventions can be applied to prevent the disease from occurring [3].

In this dissertation, I develop statistical tools to study both of these topics with humans as the focal species. As large genomic datasets have become available from individuals from populations around the world, our evolutionary history has been revealed to be much more complex than originally thought[4, 5]. In particular, admixture between archaic hominins and modern humans has been pervasive throughout our history[6, 7, 8, 9]. Understanding our evolutionary history requires the development of statistical methods that can robustly infer evolutionary phenomena. In Chapter 1, I develop a statistical model that is concerned with the problem of *local ancestry inference*. The goal of such a method is to find the segments of modern human genomes that originate from archaic admixture. The method I develop in Chapter 1, ArchIE, is able to do so without the use of an archaic reference

genome. Rather, it relies on a demographic model relating the archaic and modern human populations and takes a discriminative modeling approach to model the probability that a segment in the modern human genome is archaic in origin. An application of ArchIE to genomic data from European populations reveals segments that come from the Neanderthal population, consistent with a model of archaic admixture from Neanderthals into modern humans.

Given that archaic admixture has occurred between modern humans and two separate archaic species (Neaderthals and Denisovans), an open question is whether archaic admixture occurred between modern humans and other, as-yet-unsampled archaic populations[10, 11, 12, 13]. In Chapter 2, I develop a statistical framework that is able to answer this question focusing on populations in sub-Saharan Africa. I find that there is indeed evidence for archaic introgression from an unsampled 'ghost' population that diverged from the modern human lineage prior to the Neanderthal divergence. Using ArchIE, I was able to discover the segments of ancestry that come from this archaic population. Using this genomic map of archaic introgression, I find that there are several regions where archaic ancestry is present more frequently than expected under a model of strict neutral evolution, suggesting these archaic segments may have been adaptive.

In Chapter 3, I turn my attention to predicting the phenotypes of individuals. In particular, evolutionary forces such as mutation, genetic drift, and natural selection can affect the genetic architecture of complex traits[14, 15]. Given that human populations have split from an ancestral population, it is possible that the genetic architecture of complex traits may be different in different populations [16]. This has implications for predicting phenotypes, which rely on an inferred genetic architecture. This impact is magnified in real data, where most of the inference has occurred in populations of European ancestry. An open question is to what extent understanding the genetics of complex traits in European populations tells us about the genetics of complex traits in other, non-European populations[17, 18]. In Chapter 3, using population genetics models of complex traits, I show that natural selection will tend

to keep the frequency of the largest effect alleles low, decreasing the probability that these alleles are shared between populations. Further, I show that this has implications for transferring models for genomic prediction of complex traits between populations. I conclude by discussing the need for comprhensive studies of genomic variation in diverse populations.

References

[1] Motoo Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, February 1968.

[2] Naomi R. Wray, Kathryn E. Kemper, Benjamin J. Hayes, Michael E. Goddard, and Peter M. Visscher. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics*, 211(4):1131–1141, April 2019.

[3] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, September 2018.

[4] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015.

[5] Rasmus Nielsen, Joshua M. Akey, Mattias Jakobsson, Jonathan K. Pritchard, Sarah Tishkoff, and Eske Willerslev. Tracing the peopling of the world through genomics. *Nature*, 541(7637):302–310, January 2017.

[6] David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques

Hublin, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, December 2010.

[7] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, Hélène Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, January 2014.

[8] Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, March 2014.

[9] Sriram Sankararaman, Swapan Mallick, Nick Patterson, and David Reich. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Current Biology*, 26(9):1241–1247, May 2016.

[10] Vincent Plagnol and Jeffrey D. Wall. Possible Ancestral Structure in Human Populations. *PLOS Genetics*, 2(7):e105, July 2006.

[11] Michael F. Hammer, August E. Woerner, Fernando L. Mendez, Joseph C. Watkins, and Jeffrey D. Wall. Genetic evidence for archaic admixture in Africa. *Proceedings of the National Academy of Sciences*, 108(37):15123–15128, September 2011.

[12] Joseph Lachance, Benjamin Vernot, Clara C. Elbers, Bart Ferwerda, Alain Froment, Jean-Marie Bodo, Godfrey Lema, Wenqing Fu, Thomas B. Nyambo, Timothy R. Rebbeck, Kun Zhang, Joshua M. Akey, and Sarah A. Tishkoff. Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. *Cell*, 150(3):457–469, August 2012.

[13] Aaron P. Ragsdale and Simon Gravel. Models of archaic admixture and recent history from two-locus statistics. *PLOS Genetics*, 15(6):e1008204, June 2019.

[14] Kirk E. Lohmueller. The Impact of Population Demography and Selection on the Genetic Architecture of Complex Traits. *PLOS Genetics*, 10(5):e1004379, May 2014.

[15] Lawrence H. Uricchio, Hugo C. Kitano, Alexander Gusev, and Noah A. Zaitlen. An evolutionary compass for detecting signals of polygenic selection and mutational bias. *Evolution Letters*, 3(1):69–79, 2019.

[16] Lawrence H. Uricchio, Noah A. Zaitlen, Chun Jimmie Ye, John S. Witte, and Ryan D. Hernandez. Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. *Genome Research*, May 2016. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[17] Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear E. Kenny. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *The American Journal of Human Genetics*, 100(4):635–649, April 2017.

[18] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics*, 51(4):584, April 2019.

Chapter 1: A statistical model for reference-free inference of archaic local ancestry

PLOS GENETICS

RESEARCH ARTICLE

A statistical model for reference-free inference of archaic local ancestry

Arun Durvasula[1], **Sriram Sankararaman**[1,2,3,4]*

1 Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, **2** Department of Computer Science, University of California, Los Angeles, Los Angeles, California, **3** Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, California, **4** Department of Computational Medicine, University of California, Los Angeles, Los Angeles, California

* sriram@cs.ucla.edu

Citation: Durvasula A, Sankararaman S (2019) A statistical model for reference-free inference of archaic local ancestry. PLoS Genet 15(5): e1008175. https://doi.org/10.1371/journal.pgen.1008175

Editor: Sharon R. Browning, University of Washington, UNITED STATES

Received: July 31, 2018

Accepted: May 3, 2019

Published: May 28, 2019

Data Availability Statement: Genotype data are available from the 1000 Genome project (http://www.internationalgenome.org/data). Code is available at https://github.com/sriramlab/ArchIE. All other relevant data are available from the manuscript and its Supporting Information files.

Funding: AD is supported by NSF GRFP DGE-1650604. SS is supported in part by NIH grants R00GM111744, R35GM125055, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation. The funders had no role in study

Abstract

Statistical analyses of genomic data from diverse human populations have demonstrated that archaic hominins, such as Neanderthals and Denisovans, interbred or admixed with the ancestors of present-day humans. Central to these analyses are methods for inferring archaic ancestry along the genomes of present-day individuals (*archaic local ancestry*). Methods for archaic local ancestry inference rely on the availability of reference genomes from the ancestral archaic populations for accurate inference. However, several instances of archaic admixture lack reference archaic genomes, making it difficult to characterize these events. We present a statistical method that combines diverse population genetic summary statistics to infer archaic local ancestry without access to an archaic reference genome. We validate the accuracy and robustness of our method in simulations. When applied to genomes of European individuals, our method recovers segments that are substantially enriched for Neanderthal ancestry, even though our method did not have access to any Neanderthal reference genomes.

Author summary

Recent analyses of modern human genomes have shown that archaic hominins like Neanderthals and Denisovans contribute a few percentage of ancestry to many populations. These analyses rely on having accurate reference genomes from these archaic populations. Due to the difficulty in sequencing these genomes, we lack a complete collection of reference genomes with which to identify archaic ancestry. Here, we develop a method that identifies segments of archaic ancestry in modern human genomes without the need for archaic reference genomes. We systematically evaluate the accuracy and robustness of our method and apply it to modern European genomes to uncover signals of introgression which we confirm to be from a population related to Neanderthals.

design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Admixture, the exchange of genes among previously isolated populations, is increasingly being recognized as an important force in shaping genetic variation in natural populations. Analyses of large collections of genome sequences have shown that admixture events have been prevalent throughout human history [1]. These studies have shown that modern human populations outside of Africa trace a small percentage of their ancestry to admixture events from populations related to archaic hominins like Neanderthals and Denisovans [1, 2, 3]. Further, studies of the functional impact of archaic ancestry have suggested that Neanderthal DNA contributes to phenotypic variation in modern humans [4, 5].

Central to these studies is the problem of archaic local ancestry inference—the pinpointing of segments of an individual genome that trace their ancestry to archaic hominin populations. Methods for archaic local ancestry inference leverage various summary statistics computed from modern and ancient genomes. For example, at a given genomic locus, individuals with archaic ancestry are expected to have low sequence divergence to an archaic genome [6]. A number of summary statistics [7, 8, 9] as well as statistical models that combine these statistics [2, 10, 11, 12] to infer archaic local ancestry have been proposed.

These methods are most effective in settings where reference genomes that represent genetic variation in the archaic population are available. For example, the analyses of Neanderthal [6, 10] and Denisovan admixture events [13] relied on the genome sequences from the respective archaic populations. In a number of instances, however, the archaic population is either unknown or lacks suitable reference genomes. Several recent studies have found evidence for archaic introgression in present-day African populations from an unknown archaic hominin [14, 15, 16] while analysis of the high-coverage Denisovan genome has suggested that the sequenced individual traces a small proportion of its ancestry to a highly-diverged unknown archaic hominin [10].

One of the most widely used statistics for identifying archaic ancestry is the S^*-statistic [9], which identifies highly diverged SNPs that are in high linkage disequilibrium (LD) with each other in the present-day population as likely to be introgressed. The S^*-statistic is attractive as it can be applied even where no reference genome is available. However, the power of the S^*-statistic tends to be low in the reference-free setting [3] and its accuracy depends on a number of parameters that need to be fixed in advance.

Here, we introduce a new statistical method, ARCHaic Introgression Explorer (ArchIE), that combines several population genetic summary statistics to accurately infer archaic local ancestry without the need for a reference genome. ArchIE is based on a logistic regression model that predicts the probability of archaic ancestry for each window along an individual genome. The parameters of ArchIE are estimated from training data generated using coalescent simulations. Our proposed method has several advantages. First, the model can incorporate a variety of statistics that are potentially informative of archaic ancestry. This flexibility allows the model to be applied to the reference-free setting (the setting that is the focus in this paper). However, the model can be extended to also incorporate reference genomes when available, even when these reference genomes might be from distant representatives [10] or from low-coverage samples [17, 18]. Second, our use of a statistical model allows us to efficiently estimate model parameters that optimize desired objective functions such as the likelihood. This property allows the model to be adapted to admixture events with different time depths or admixture fractions as well as to infer other population genetic parameters of interest. Indeed, recent studies have shown that statistical predictors that combine weakly-informative summary statistics can substantially improve a number of population genetic inference problems [19, 20, 21].

We show that ArchIE obtains improved accuracy in simulations over the S^*-statistic (as well as the recently proposed S' method [22]) while being robust to demographic model misspecifications that can cause the distribution of features and archaic ancestry labels in the training data to differ from the test data. We apply ArchIE to Western European (CEU) genomes from the 1000 Genomes project and show that the segments inferred to harbor archaic ancestry have an increased likelihood of being introgressed from Neanderthals even though no Neanderthal genome was used in the inference. These segments recover previously observed features of introgressed Neanderthal ancestry: we observe a decreased frequency of these segments in regions of the genome with stronger selective constraint [23] as well as elevated frequency at the BNC2 and OAS loci that have previously been reported to harbor elevated frequencies of Neanderthal ancestry [2, 3].

Results

Overview of statistical model to detect archaic local ancestry

Our method, ArchIE, aims to predict the archaic local ancestry state in a given window along an individual haploid genome. This prediction is performed using a binary logistic regression model given a set of features computed within this window. Estimating the parameters of this model requires labeled training data *i.e.*, a dataset containing pairs of features and the archaic local ancestry state for a given window along an individual genome. To obtain labeled training data, we simulate data under a demographic model that includes archaic introgression, label windows as archaic or not, compute features that are potentially informative of introgression, and estimate the parameters of our predictor on the resulting training data (Fig 1A, Methods). While our method is general enough to be applicable to non-human populations, we describe the demographic model in terms of a modern human-archaic human demographic history.

We simulate training data using a modified version of the coalescent simulator, ms [24], which allows us to track each individual's ancestry. We use the demographic model from Sankararaman *et al.* 2014 [2] (See Table 1). In this model, an ancestral population splits T_0 generations before present (B.P.) forming two populations (archaic and modern human in the case of the Neanderthal-human demography). The modern human population subsequently splits into two populations T_s generations B.P., one of which then interbreeds with the archaic population (referred to as the target population) while the other does not (the reference population). We simulate one haploid genome (haplotype) in the archaic population, 100 haplotypes in the target population and 100 haplotypes in the reference population (thus, a target population consists of 50 diploid individuals). We sample the archaic haplotype at the same time as the modern human haplotypes, but the statistics we calculate do not rely on features of the archaic genome. We simulate 10,000 replicates of 50,000 base pairs each (bp), resulting in 1,000,000 training examples. We use a window of length 50 Kb because that is the mean length of the introgressed archaic haplotype after $T_a = 2,000$ generations based on the recombination rate assumed in our simulations.

We summarize the training data using features that are likely to be informative of archaic admixture. Since we are interested in the probability of archaic ancestry for a given focal haplotype, we compute features that are specific for the focal haplotype. First, for the focal haplotype, we calculate an individual frequency spectrum (IFS), which is a vector of length n, the haploid sample size of the target population. Each entry in the vector is the number of mutations on the focal haplotype that are segregating in the target population with a specific count of derived alleles. Due to the accumulation of private mutations in the archaic population, we expect the IFS to capture the excess of alleles segregating at frequencies close to the admixture

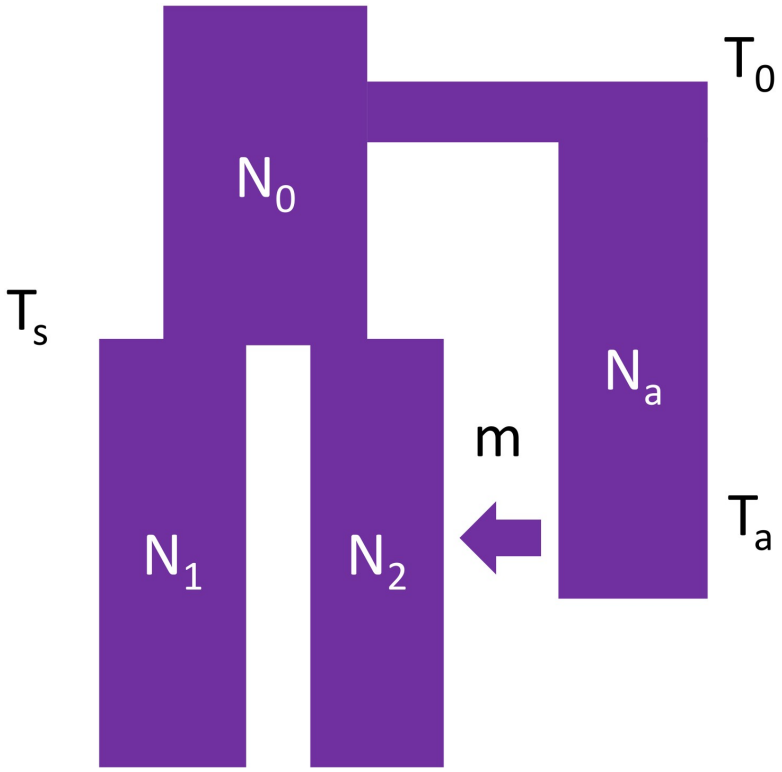

Fig 1. Outline of the demographic model used for training ArchIE. We simulate a population starting at size N_0 and splitting into archaic and modern human (MH) populations at time T_0. The MH population splits into a reference and target population of size N_1 and N_2, respectively, at time T_s. Then, at time T_a, the archaic population admixes with the target population with an associated admixture proportion m. We use data simulated from this model to train a logistic regression classifier.

https://doi.org/10.1371/journal.pgen.1008175.g001

fraction in the introgressed population. This statistic is closely related to the conditional site frequency spectrum [25].

Next, we calculate the Euclidean distance between the focal haplotype and all other haplotypes, resulting in a vector of length n. Under a scenario of archaic admixture, the distribution

Table 1. Parameters used in training simulations.

Parameter	Description	Value
N_1	Reference population size	10000
N_2	Target population size	10000
N_a	Archaic population size	10000
N_0	Ancestral population size	10000
m	Admixture fraction	2%
T_0	Archaic split time	12000
T_S	Target-Reference split time	2500
T_a	Admixture time	2000
μ	Per base pair mutation rate	1.25×10^{-8}
r	Per base pair recombination rate	1×10^{-8}

https://doi.org/10.1371/journal.pgen.1008175.t001

of pairwise differences is expected to differ when we compare two haplotypes that are both modern human or archaic versus when we compare an archaic haplotype to a modern human haplotype. We also include the first four moments of this distribution, *i.e.*, the mean, variance, skew, and kurtosis. These summaries of haplotype distance are similar to the D_1 statistic used in Hammer *et. al.* [14].

The next set of features rely on a present-day reference human population that has a different demographic history compared to the target population. The choice of the reference can alter the specific admixture events that our method is sensitive to: we expect the method to be sensitive to admixture events in the history of the target population since its divergence from the reference. While our method can also be applied in the setting where no such reference population exists, in the context of human populations where genomes from a diverse set of populations is available [1], the use of the reference can improve the accuracy and the interpretability of our predictions. Given a reference population, we compute the minimum distance of the focal haplotype to all haplotypes in the reference population. A larger distance is suggestive of admixture from a population that diverged from the ancestor of the target and reference populations before the reference and target populations split. This feature shares some similarities with the D_2 statistic from Hammer *et. al.* [14].

We also calculate the number of SNPs private to the focal haplotype, removing SNPs shared with the reference, as these SNPs are suggestive of an introgressed haplotype. Finally, we calculate S^* [9], a statistic designed for detecting archaic admixture by looking for long stretches of derived alleles in high LD.

Using these features, we train a logistic regression classifier to distinguish between archaic and non archaic segments. In our training data, we define archaic haplotypes as those for which $\geq 70\%$ of bases are truly archaic in ancestry and non-archaic as those for which $\leq 30\%$ are archaic in ancestry. We discard haplotypes that fall in-between those values in the training data resulting in 988,372 training examples.

Accuracy of estimates of archaic local ancestry

We tested the accuracy of ArchIE by simulating data under a demography reflective of the history of Neanderthals and present-day humans [2]. We evaluated the ability of ArchIE to correctly predict the archaic ancestry at each SNP along an individual haplotype. Since ArchIE predicts archaic ancestry within a window, we simulated a 1 Mb segment, applied ArchIE in a 50 Kb window that slides 10 Kb at a time, and predicted archaic ancestry at a SNP by averaging predictions across all windows that overlap the SNP (Methods). We compute Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves by varying the threshold at which we call a SNP archaic and calculating the true positive rate (TPR), false positive rate (FPR), precision, and recall (Fig 2).

We compared ArchIE to an implementation of the S^*-statistic from Vernot and Akey using their hyper parameter choices [3] and to S', a new method for reference-free inference of archaic ancestry [22] (Methods). At a 2% admixture fraction, ArchIE outperforms the S^* and S' statistics across all thresholds (Fig 2A and 2B). At a precision of 0.80, *i.e.*, false discovery rate of 20%, ArchIE obtains a recall of 0.21, S^* obtains a recall of 0.04, and S' obtains a recall of 0.09. The area under the ROC curve (AUROC) is 0.94 (±0.008) for S^*, 0.84 (±0.01) for S', and 0.97 (±0.005) for ArchIE and the area under the PR curve (AUPR) is 0.47 for S^* (±0.031), 0.28 (±0.032) for S', and 0.60 (±0.05) for ArchIE (All standard error were estimated using a block jackknife [26] using 1 Mb blocks). We also note that while the ROC curves are similar, the PR curves show a large difference, indicative of the utility of PR curves in problems where there is an imbalance in the frequencies of the two classes.

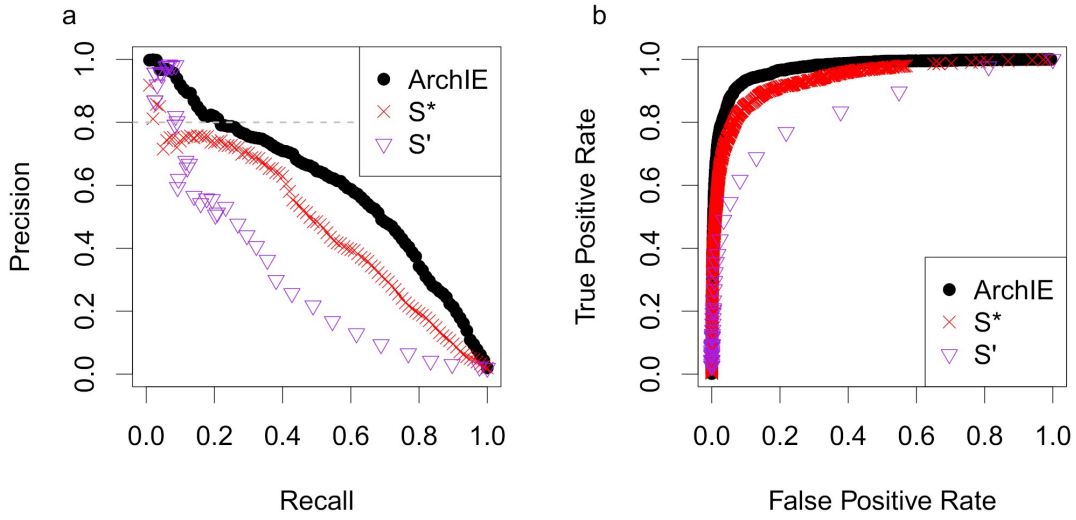

Fig 2. ArchIE obtains improved accuracy over related methods. (A) Precision-Recall (PR) and (B) Receiver Operator Characteristic (ROC) curves for ArchIE (black circles), S* (red crosses), and S' (purple triangles) in a 2% admixture scenario with a Human-Neanderthal demography. The dashed line corresponds to a false discovery rate of 20%.

https://doi.org/10.1371/journal.pgen.1008175.g002

We also evaluated the ability of ArchIE to call archaic haplotypes. Since haplotypes can range from having none of their ancestry to being entirely from the archaic population, we called haplotypes archaic if they contain $\geq$ 70% archaic ancestry or not archaic if they contain $\leq$ 30%. We see that again, ArchIE has larger AUPR (0.53 for ArchIE, 0.38 for S*) and AUROC (0.97 for ArchIE, 0.94 for S*) compared to S* (S4 Fig).

Population genetic features informative of archaic local ancestry

We examined the absolute value of the standardized weights learned by ArchIE to understand the features that contribute substantially to its predictions. Examining single features, we find that the minimum distance between the focal haplotype and each of the reference haplotypes, as well as the skew of the distance vector have the largest weights (Fig 3B). Intuitively, a larger distance to a reference population should indicate archaic ancestry. The next largest single statistic was the skew of the distance vector, which was negatively correlated with archaic ancestry. Under a simple scenario of admixture, we expect a bi-modal distribution of pairwise distances. However, when there is little archaic ancestry, the distribution will be unimodal resulting in a negative relationship between skew and archaic ancestry. The IFS contains mostly negative weights, suggesting that these features do not make a substantial contribution to the model predictions (Fig 3A).

As a further check, we wanted to determine how the performance of the model changes when trained on subsets of the features. First, since the "skew" feature has a large standardized absolute weight, we trained a model based only on this feature (S5 Fig). We find that accuracy greatly decreases, indicating that the model does best when it combines multiple features that are informative of archaic introgression. However, when we train only on the number of private SNPs or only on the minimum distance to the reference population, we see improved accuracy indicating that these features are informative of archaic ancestry independent of

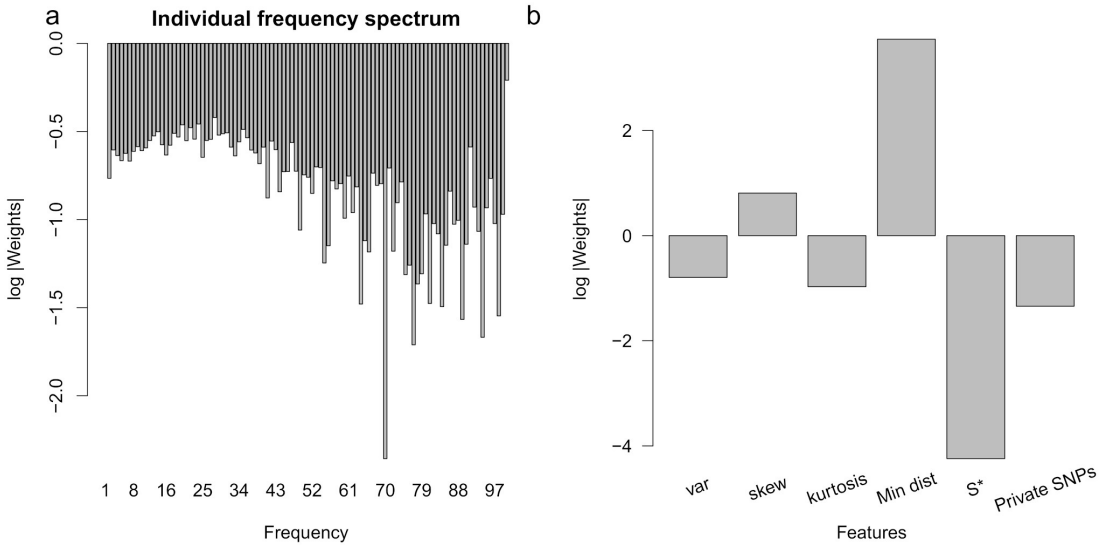

Fig 3. Relative importance of the features used as input to ArchIE. We examined the log of the absolute value of the standardized weights associated with each of the features included in the logistic regression model underlying ArchIE. Negative values indicate standardize weights with absolute values less than 1. (A) The individual frequency spectrum mostly has small weights and lower frequency entries generally have larger weights associated with them. (B) The first three entries indicate the moments of the distance vector. The minimum distance to the reference population, skew, and variance of the distance vector have the largest weights associated with them.

https://doi.org/10.1371/journal.pgen.1008175.g003

other features. When we take a combination of three features (skew, number of private SNPs, and minimum distance to the reference population), this model is still able to discern archaic from non-archaic haplotypes with slight decreased accuracy relative to the full model (S5 Fig). Finally, we tested the contribution of the reference population to the accuracy of ArchIE. We trained the logistic regression without using any features that rely on the reference and found that model still retains reasonable accuracy (AUPR = 0.36) to identify archaic ancestry (S5 Fig). This suggests that ArchIE is useful even in scenarios where a reference population is not available.

Robustness of archaic local ancestry estimates

ArchIE relies on simulating data from a model with fixed demographic and population genetic parameters. In practice, these parameters are unknown and are inferred from data with some uncertainty. Thus, we wanted to determine the sensitivity of our method to demographic uncertainty. An exhaustive exploration of demographic uncertainty is challenging given the number of parameters associated with even the simplest models. As an alternative to an exhaustive exploration, we systematically perturbed each parameter at a time, simulated data using the perturbed model, and evaluated the performance of our classifier (trained on the unperturbed parameters corresponding to the Neanderthal demographic history).

ArchIE remains accurate when many aspects of the demography are misspecified, but has reduced precision or recall under some scenarios (Fig 4, S1 Fig). The most significant decrease in accuracy (in terms of recall and precision at a fixed threshold) arises when the reference population size is decreased or the split time of the reference and the target is increased. In this setting, the reference genomes are more drifted and hence, less representative of the ancestral

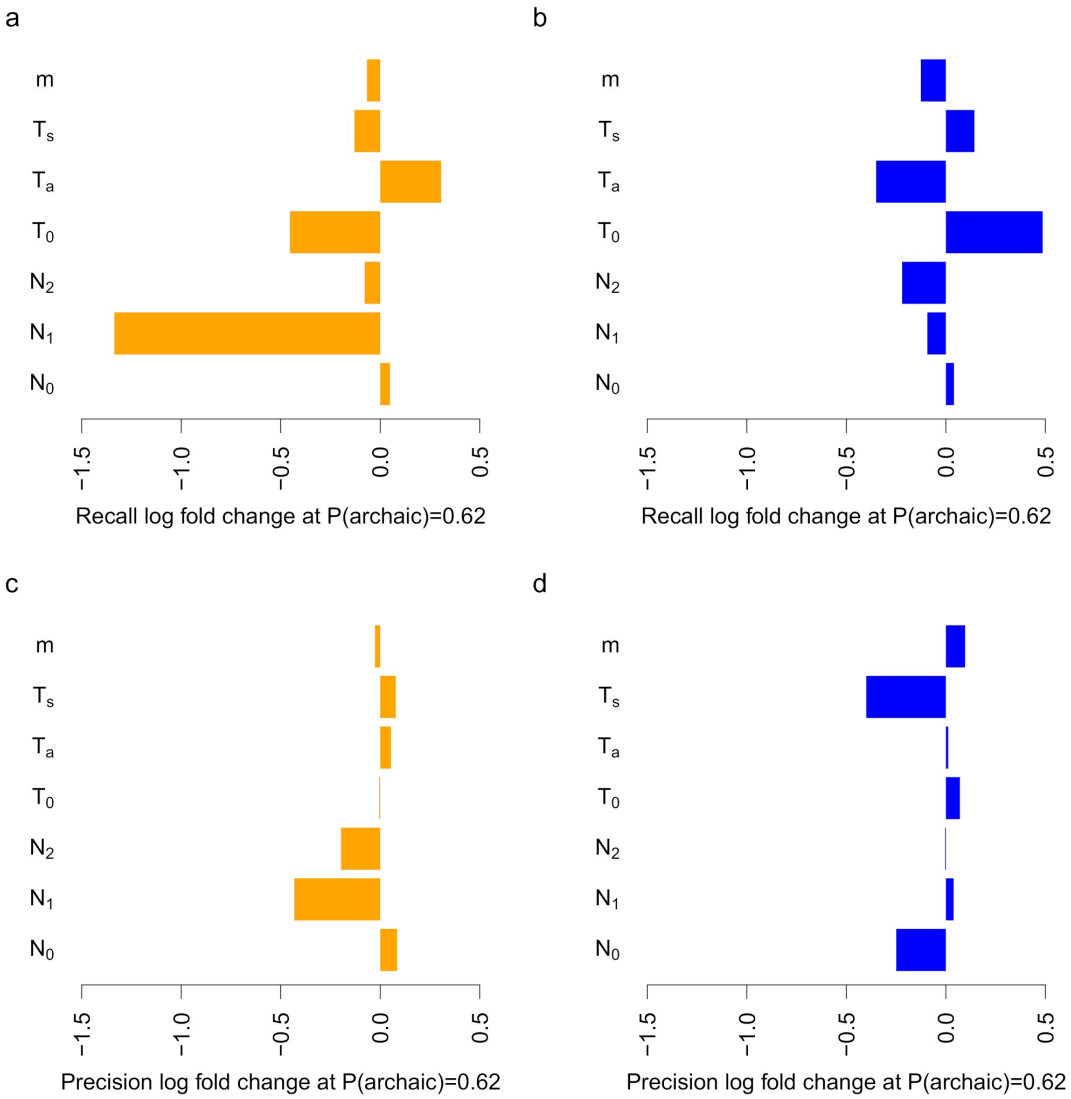

Fig 4. ArchIE is robust to misspecification in the demographic model. We tested ArchIE on data simulated after perturbing single demographic parameters lower (left, orange) and higher (right, blue) relative to their values in the training data. Values are reported as $\log_{10}$ fold changes compared to the baseline model performance. We report (a, b) recall and (c,d) precision at the threshold that gives a precision of 0.8 on the unperturbed test data (P(archaic) = 0.62).

https://doi.org/10.1371/journal.pgen.1008175.g004

population. We also compared the accuracies of ArchIE to S^* across these perturbations and found that ArchIE remains relatively accurate across these settings (S1 Table).

We also tested the effect of variation in mutation rate (μ) and recombination rate (r) since we trained our model using fixed values of these parameters ($\mu = 1.25 \times 10^{-8}$, $r = 1 \times 10^{-8}$). To evaluate how ArchIE performs on real data, we simulated test data randomly drawing pairs of μ and r from a distribution chosen to match local recombination and mutation rates along the

human genome (see Methods). The overall AUPR is reduced (0.31, S1i Fig), the $\log_{10}$ fold changes in precision and recall are −0.30 and +0.19 suggesting that ArchIE is relatively robust to variation in mutation and recombination rates.

In addition, we tested the impact of the window size and found that reasonable choices of window size do not substantially impact the performance (S2 Fig). We also assessed the impact of sample size by simulating 30 haplotypes (15 diploid individuals), representing a modestly sized genomic dataset, and found a reduction in power as expected (AUPR = 0.45) (S3 Fig).

We tested the sensitivity of ArchIE to recent and ancestral structure in the demographic model. We simulated data under two scenarios of structure, one where 25% of the target population separates immediately after the target and reference population split, 2499 generations ago, and rejoins the generation prior to the archaic admixture, 2001 generations ago (S6A Fig). We refer to this as the recent structure scenario. Additionally, we simulated data where 25% of the population in N_0 separates 12,000 generations ago and rejoins the ancestral population right before the target and reference populations split (2600 generations ago, S6B Fig). We refer to this as the ancestral structure scenario. We observe that for both scenarios, the fraction of SNPs detected as archaic is 0, suggesting that ArchIE is robust to introgression due to either recent or ancient structure at reasonable calling thresholds. We caution, however, that a more detailed exploration of structured demographic models is necessary.

Reference-free detection of Neanderthal introgression in European populations

To identify segments of archaic ancestry in modern human populations, we applied ArchIE to genomes of European individuals in the 1000 Genomes Project [27]. We used all unrelated individuals from a European (CEU) population as our target population (99 diploid individuals) and all unrelated individuals from an African (YRI) population as a reference (108 diploid individuals) and calculated the summary statistics described above. We applied ArchIE in non-overlapping 50 Kb windows. We evaluated the average percent of windows inferred as archaic as a function of the calling threshold (Fig 5A). Applying a threshold corresponding to a precision of 0.80 in simulations, we inferred 2.04% (block jackknife SE = 0.6% using 1 Mb blocks) of the genome as confidently archaic. This proportion is in line with proportion of Neanderthal ancestry from previous analyses [2, 6, 10] suggesting that the segments of archaic ancestry inferred by ArchIE likely correspond to segments of Neanderthal ancestry.

To further investigate whether the haplotypes inferred as confidently archaic by our model are enriched for introgressed Neanderthal variants, we computed a Neanderthal match statistic (NMS) defined as the number of shared variants between an individual haplotype and the Altai Neanderthal reference genome sequence [10] divided by the total number of segregating sites in that window (see Methods). We see that the archaic regions confidently inferred by ArchIE have a higher NMS suggesting that the archaic ancestry segments identified by our method are likely to represent introgressed Neanderthal sequence (we reject the null hypothesis that the difference in NMS is zero for archaic vs non-archaic haplotypes with a P value = 1.7×10^{-3} via 100 Kb block jackknife). Further, as we make the calling threshold more strict, we see an increase in the mean NMS for the archaic haplotypes (Fig 5B).

We also compared the performance of ArchIE, S', and S* on real data from CEU Europeans. For each of these methods, we computed a matching rate with the Altai Neanderthal genome, defined as the fraction of SNPs called archaic that match the Altai Neanderthal sequence divided by the total number of SNPs called archaic. At a detection rate of $\approx$ 1%, S' has a matching rate of 0.73 while ArchIE has a matching rate of 0.91 (S9 Fig; see S1 Text for details).

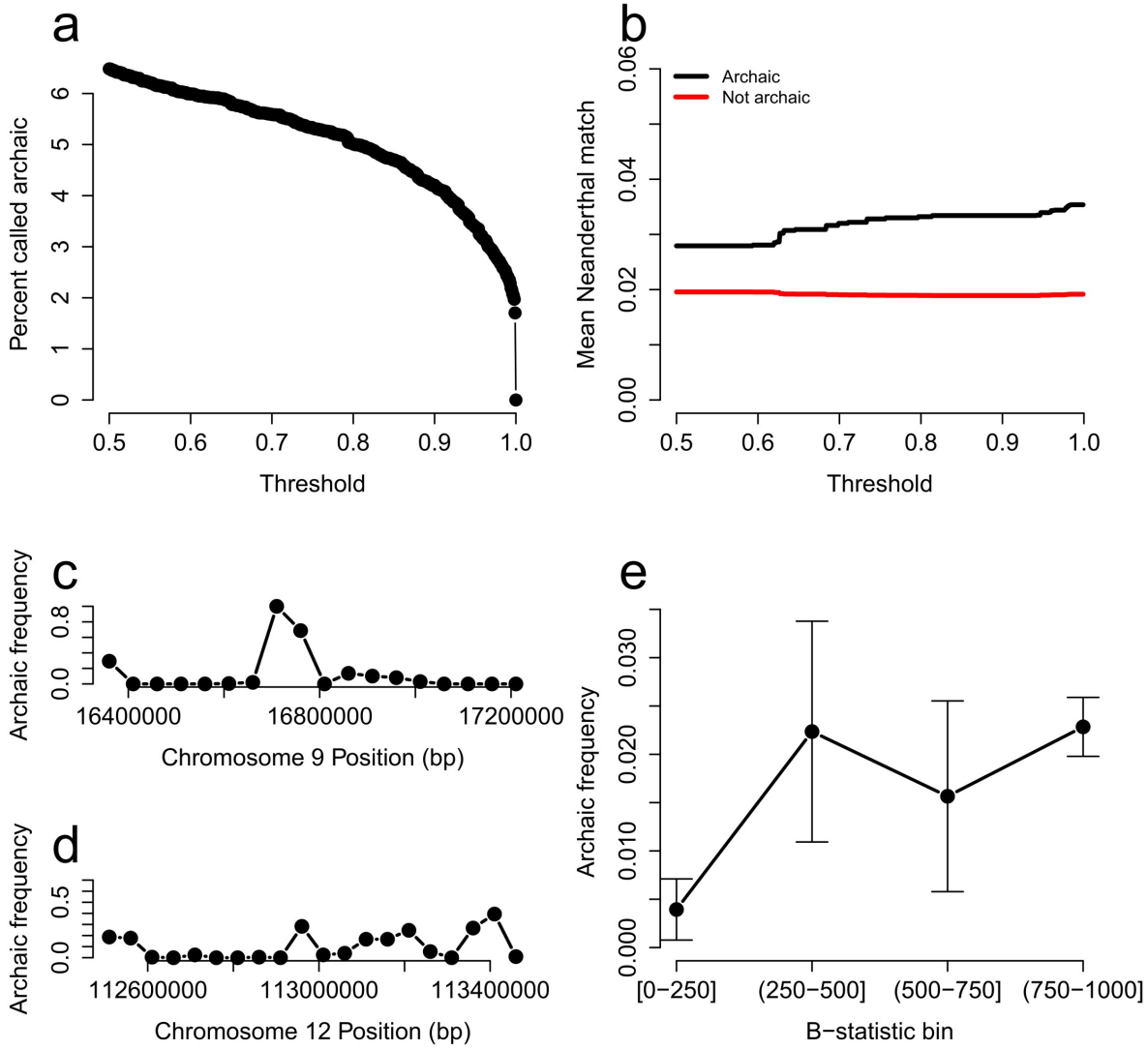

Fig 5. Application of ArchIE to 1000 Genomes European population (CEU). (A) Percentage of genome called archaic as a function of the threshold on the probability of archaic ancestry estimated by ArchIE. The dashed line refers to the threshold that yields a 20% FDR in simulations. (B) Mean Neanderthal match statistic (higher implies more similar to the sequenced Altai Neanderthal genome) for haplotypes inferred as archaic vs non-archaic as a function of the probability threshold. (C) Frequency of haplotypes confidently labeled as archaic near the *BNC2* gene and (D) the *OAS* gene cluster. (E) Mean frequency of confidently archaic segments increases with B-statistic (a measure of selective constraint). Low B-statistic denotes more selectively constrained regions (standard errors estimates are obtained using a 1 Mb block jackknife).

https://doi.org/10.1371/journal.pgen.1008175.g005

Comparing with the S* calls released from [28], we found a match rate of ≈ 50% at a detection rate of ≈ 0.5%, consistent with results reported from the authors.

We then focused on two genomic regions that have been shown to harbor introgressed Neanderthal haplotypes at elevated frequencies: the *BNC2* gene (Chromosome 9:16,409,501-

16,870,786) [2] and the *OAS* gene cluster (Chromosome 12:113,344,739-113,357,712) [7]. ArchIE detects substantially increased frequency of archaic ancestry in both these genes (Fig 5C and 5D).

Finally, we analyzed the correlation between a measure of selective constraint of a given genomic region (B-value [23]) and frequency of confidently inferred archaic segments in the CEU population in the same region. Sankararaman *et al*. 2014 [2] describe a relationship where more constrained regions (lower B-value) have a lower frequency of archaic ancestry. We observe the same trend where more neutral regions (B-value $\geq$ 750) contain more archaic ancestry than constrained regions (B-value $\leq$ 250) consistent with selection against the archaic ancestry (P value = 7.86×10^{-9} via block jackknife; Fig 5E).

These analyses suggest that ArchIE obtains results concordant with those from a previous reference-aware method [2]. We caution, however, that the observed concordance can be inflated due to any biases shared by the two methods.

Discussion

A key challenge in detecting the contribution of deeply-diverged populations (both deeply-diverged modern as well as archaic hominin populations) to the ancestry of present-day human populations arises from the lack of accurate representative genomes for these populations. Here, we present a statistical model (ArchIE) for detecting regions of archaic local ancestry without the need for an archaic reference sequence. ArchIE combines weakly informative signals computed from present-day human genomes using a logistic regression model. The parameters of the model are estimated from data simulated under a specific demographic model. Using simulations, we show that ArchIE obtains improved accuracy over other approaches for reference-free local ancestry inference. While the accuracy of ArchIE will depend on how similar the demographic model used for training is to the true demographic model, our empirical results suggest that ArchIE is relatively robust even when the true demographic model differs from the assumed model. Applying ArchIE to genomes from the CEU population in the 1000 Genomes project data, we detect 2.03 ± 0.6% archaic ancestry (at a threshold that corresponds to a false discovery rate of 0.2). We find that segments confidently labeled as archaic by ArchIE are enriched for Neanderthal ancestry.

One advantage of our approach is that the learning algorithm is general allowing it to be applied broadly to diverse inference problems as well as input summary statistics while its simplicity allows for a transparent interpretation of the features and the model.

There are several limitations of our methodology, however. First, we require some knowledge of the demographic history of the target, reference and archaic populations. We have shown that ArchIE is robust to some demographic misspecification, but it is most powerful when the simulated demography is close to the true one. Second, we rely on the data being phased. Switch-errors in phasing will reduce the power of ArchIE, which can be a problem when applying the method to less-well studied populations. In principle it is possible to use ArchIE on unphased data, calculating features on the diploid individual level rather than the haplotype level, though we do not explore that here. Third, the use of a fixed-size window ignores long-range as well as variable-length dependence among the features. Models that account for this dependency can be expected to yield improved accuracy. An example of such an approach is a recently published method that uses a hidden Markov model (HMM) that models the distribution of private variants [12]. Combining such models with the framework outlined here has the potential to yield improved accuracies. Fourth, the use of a linear model is likely to underfit the true function between features and outputs. It is possible to train more expressive models like deep neural networks, which can learn and capture non-linear

relationships between features and tend not to suffer from the curse of dimensionality [19]. These methods have been used to great success in tasks such as image classification [29] and we anticipate their use in population genetics could improve predictive power. Preliminary results applying deep learning to this problem with the features used here are promising, motivating future work (S1 Text, S7 and S8 Figs). ArchIE relies on a careful choice of features as input. These hand crafted features are informed by population genetics theory, similar to other methods that have been proposed in population genetics [19, 20, 30, 31, 14]. Automatically learning features from genetic data is direction of high interest. Finally, while several methods [9, 12, 22] have been proposed to infer aspects of archaic ancestry without access to reference genomes, these methods are typically evaluated using simulations. Assessing the accuracy of these methods on real data remains challenging. Extrapolating simulation results to accuracy on real data depends on choices of the inference problem, population genetic models, parameters used for training and testing, genomic features used as input, and accuracy metrics of interest. A comprehensive comparison of these methods across a range of demographic histories and evolutionary forces is an important topic for future work.

In conclusion, our method improves on previous methods for reference-free inference of archaic ancestry by combining informative summary statistics in a statistical learning framework. We anticipate that this method will be informative not only in human populations where questions about admixture with other hominins abound, but also in other species and systems where pervasive admixture has shaped the distribution of genetic variation.

Methods

Simulating training data

We simulated training and test data sets using a modified version of ms [24] that tracks the ancestry of each site in each individual genome. Using a previously proposed demographic model relating modern humans and Neanderthals [2], we sampled 100 haplotypes from the target, and 100 haplotypes from the reference over a region of length 50 Kb. We use a constant mutation rate $\mu = 1.25 \times 10^{-8}$ and a recombination rate $r = 1 \times 10^{-8}$.

The general demography is as follows: an archaic population of size N_a splits from a population of size N_0, T_0 generations before present (B.P.). Then, at T_S, two populations split off from the ancestral population that then have effective population sizes N_1 (termed the reference) and N_2 (termed the target) respectively. Then, at time T_A, the archaic population migrates into the target with an admixture fraction m. See Fig 1 for a graphical outline.

Feature calculation

Each simulation at a given locus generates 100 haplotypes in the target. For each haplotype, we calculate the following classes of summary statistics: individual frequency spectrum, distance vector to all haplotypes within the test population as well as the first four moments of this vector, minimum distance to haplotypes in the reference population, the number of private SNPs, and the S^*-statistic.

The individual frequency spectrum is created as follows: given a sample of n haplotypes, for each haplotype j, we construct a vector X of length n where entry X_i counts the number of derived alleles carried on the focal haplotype j whose derived allele frequency is i. For example, the first entry counts the number of singletons present in haplotype j, the second entry counts the number of doubletons and so on until n.

The distance vector is a vector of length n where entry i is the Euclidean distance from haplotype j to haplotype i over all sites, where j is the focal haplotype and i is the haplotype being compared.

The minimum distance to haplotypes in the reference population is computed as the minimum Euclidean distance from the focal haplotype to all haplotypes in the reference population.

The number of private SNPs is calculated as the number of SNPs the focal haplotype contains that are not present in the reference population.

This results in 208 features per example (a 50 Kb window for a single haploid genome), with 100 examples per locus and 10,000 loci resulting in 1,000,000 examples for training before filtering haplotypes with intermediate levels of admixture.

Learning algorithm

We used the "glm" function in R to construct a logistic regression model using the family = binomial("logit") option. We used the predict function to obtain a prediction and converted it to a probability using the "plogis" function.

Due to the process of recombination, the ancestry of a haplotype may vary along its length. On the other hand, ArchIE predicts a single ancestry state for a haplotype across a specified window. We evaluate the ability of ArchIE to predict the ancestry at each SNP along a haplotype by simulating sequences of length 1 Mb and applying ArchIE in 50 Kb windows, sliding by 10 Kb at a time. We average the predictions that each SNP on a haplotype receives across all windows that overlap the SNP to obtain the predicted archaic ancestry. We compare the predicted and the true ancestry state at each SNP along a haplotype.

We evaluated the performance using Precision-Recall (PR) curves as well as receiver operator characteristic (ROC) curves. We calculated precision (equivalently 1– the false discovery rate), recall (equivalently sensitivity) and false positive rates as:

$$Recall(t) = \frac{TP(t)}{TP(t) + FN(t)}$$

$$Sensitivity(t) = Precision(t) = \frac{TP(t)}{TP(t) + FP(t)}$$

$$False \quad positive \quad rate(t) = \frac{FP(t)}{FP(t) + TN(t)}$$

Here $TP(t)$ is the number of true positives at threshold t, $FN(t)$ is the number of false negatives at threshold t, $FP(t)$ is the number of false positives at threshold t and $TN(t)$ is the number of true negatives at threshold t. We summarize these results by reporting the recall at a fixed value of precision as well as by computing the area under the precision recall curve (AUPR) and the area under the ROC curve (AUROC). We compute the AUPR using the method of Davis and Goadrich [32]. We compute standard errors of the AUPR and AUROC using a block jackknife [26] where we drop a single 1 Mb region and recompute the statistics.

Comparisons

We compared ArchIE to the S^* [9] and S' [22] statistics. We calculate S^* in a cohort of 100 haplotypes from the target population. Then, we convert the S^* scores into a rank between [0-1] using the empirical cumulative distribution. We use a 50 Kb sliding window (10 Kb stride) across the 1 Mb region, averaging the score for a SNP.

We use a similar strategy for S'. However, since S' predicts archaic ancestry in a sample of individuals rather than on the haploid genome level, we use an algorithm to convert sample

predictions to haploid genome predictions. We run S' on the sample. Then, at some S' score threshold, we find the longest stretch of SNPs at that score or higher and interpolate the scores across genotypes, building haplotypes when individuals have the archaic allele. Then, for each SNP, we evaluate whether the SNP is archaic or not and calculate the number of true positives, false positive, true negatives, and false negatives. We repeat this procedure across thresholds and calculate the precision, recall, and false positive rates.

Robustness

We examined the robustness of ArchIE to a specified demographic model by systematically perturbing one parameter at a time, simulating a dataset, and evaluating ArchIE's performance. We doubled and halved the parameters, except when doing so would produce a demographic model that is not sensible.

We evaluated the robustness of ArchIE to mutation and recombination rate variation by calculating local rates at 50 Kb windows and then randomly drawing combinations of the rates and simulating data. Mutation rates were calculated by estimating Watterson's θ [33] from the number of segregating sites within 50 Kb windows across 50 randomly sampled west African Yoruba genomes from the 1000 Genomes Project Phase 3 release and calculating the mutation rate: $\mu = \theta_w/4N_eL$ where we set $N_e = 10,000$. Recombination rates were estimated from the combined, sex-averaged HapMap recombination map [34].

Neanderthal introgression

We validated our method using the Neanderthal introgression scenario as a test case. We downloaded phased CEU genomes from the 1000 Genomes Phase 3 dataset [27] and calculated the features mentioned above in 50 Kb windows. For each individual haplotype, we inferred the probability that the window is archaic. We then intersected our calls with the 1000 Genomes strict mask using BEDtools v2.26.0 [35], removing regions that are difficult to map to, measured as having less than 90% of sites in the callability mask.

We calculated a Neanderthal match statistic (NMS) for focal haplotype i in a window as the fraction of alleles at which the the focal haplotype matches the Altai Neanderthal [10] genome:

$$NMS_i = \frac{S_i}{N_i + H_i}$$

Here S_i denotes the number of alleles that match between the focal haplotype and the Neanderthal genome within the window. Since the Neanderthal genome is not phased, we count sites as matching if it contained at least one single matching allele or more. N_i denotes the number of Neanderthal mutations, including both homozygous and heterozygous sites. H_i denotes the number of human mutations within the window.

In order to test whether there is more Neanderthal matching in archaic haplotypes compared to non-archaic haplotypes, we computed the difference in NMS between the two classes of haplotypes at each window and test the hypothesis that the mean of this statistic averaged across the genome is zero. Specifically:

$$\Delta_{NMS,i} = \frac{\overline{NMS}_{arch,i} - \overline{NMS}_{non-arch,i}}{\overline{NMS}_i}$$

For each window i, we compute $\Delta_{NMS,i}$, defined as the difference between the mean NMS for archaic ($\overline{NMS}_{arch,i}$) and non-archaic ($\overline{NMS}_{non-arch,i}$) haplotypes divided by the mean NMS of all haplotypes ($\overline{NMS}_i$) to control for mutation rate heterogeneity. We require a minimum of

90% callable sites within the window. We compute the mean of $\Delta_{NMS,i}$ over all windows i as the genome-wide estimate and test if this estimate is significantly different from zero. To compute significance, we use a block jackknife and drop non-overlapping 100 Kb windows and recalculate the genome wide difference in means.

Background selection

In order to assess the relationship between background selection and inferred archaic ancestry, we use the B-values from McVicker *et al.* 2009 [23] and intersected them with our calls. For visualization, we binned the B-values into 4 bins, [0-250], (250-500], (500-750], and (750-1000].

We tested for significant differences in allele frequency between the lowest and highest bins using a block jackknife using a 50 Kb block size.

Supporting information

S1 Fig. Precision-Recall curves when the distribution of the test data differs from the training data used for estimating the parameters of ArchIE. We perturbed a single parameter associated with the simulations used for generating training data. m is the admixture fraction from the archaic into the target population. N_0 is the ancestral population size. N_1 is the size of the reference population and N_2 is the size of the target population. T_0 refers to the split time of the archaic and modern human population. T_s is the split time of the reference and target populations. T_a is the admixture time and mu rho refers to the experiment that uses realistic recombination and mutation rates, estimated from the human genome (see Methods for more details).
(PDF)

S2 Fig. Robustness to changing window size. ArchIE obtained similar accuracies when applied with window sizes of 100 Kb and 25 Kb relative to the 50 Kb case ('Unperturbed').
(PDF)

S3 Fig. Robustness to smaller sample sizes. We evaluated how ArchIE performs with 30 haplotypes (15 diploid individuals). We see that ArchIE loses power when the sample size is greatly reduced.
(PDF)

S4 Fig. Precision-Recall and Receiver Operating Characteristic curves for haplotype-level predictions. We evaluated ArchIE's ability to predict entire haplotypes as archaic (as opposed to archaic ancestry at each SNP in Fig 2). A haplotype is labeled as truly archaic if $\geq$ 70% of its bases are archaic in ancestry and not archaic if $\leq$ 30 is labeled archaic. We ignore haplotypes with intermediate values of archaic ancestry from our comparisons. We used haplotypes of length 50 Kb.
(PDF)

S5 Fig. Precision-Recall curves for different sets of input features. In 'No MH Ref', we removed the features that rely on the reference population. The resulting predictor has reasonable albeit reduced accuracy relative to ArchIE (labeled "Full"). We evaluated the predictive accuracy of a logistic regression model trained with only a single feature where we considered the skew feature ("skew only"), the private SNPs feature ("P SNPs only"), and the minimum distance to the reference ("Min. D only"). Accuracy is substantially decreased for "skew only" while using only the private SNPs feature ('P SNPs only') or the minimum distance to the reference ('Min. D only') results in good performance, especially at the high precision regime. In

'3 feat.', we use skew, minimum distance, and private SNPs as the only features. While this set achieves good performance, adding the full set of features still outperforms this set of three features. Area under the PR curve (AUPR) is shown in parenthesis.
(PDF)

S6 Fig. Demographic models for (A) recent structure and (B) ancient structure.
(PDF)

S7 Fig. Neural network architecture and training procedure.
(PDF)

S8 Fig. Neural network performance. Precision-recall curves for a 2% admixture scenario. Performance of the neural network is shown in blue.
(PDF)

S9 Fig. Comparison of ArchIE, S', and S* in 1000G CEU individuals.
(PDF)

S1 Table. Robustness to demographic misspecification. We simulated data under misspecified demographies, perturbing each parameter separately and evaluated the performance of S* and ArchIE. We present precision and recall at a threshold that corresponds to a precision of 0.8 (20% FDR) in the unperturbed setting. Bold denotes settings where ArchIE is higher precision as well as recall over S*.
(XLSX)

S1 Text. Neural network model description and comparison of ArchIE with S' and S* in 1000G data.
(PDF)

Acknowledgments

We would like to thank Emilia Huerta Sánchez and Benjamin Vernot for help with S*, members of the Sankararaman and Lohmueller labs, the UCLA Medical and Population Genetics group for helpful discussions, and Alec Chiu for comments on a draft of the paper and for testing code. We thank Molly Schumer and Priya Moorjani for comments on a preprint of this work. Code is available at https://github.com/sriramlab/ArchIE.

Author Contributions

Conceptualization: Arun Durvasula, Sriram Sankararaman.

Formal analysis: Arun Durvasula, Sriram Sankararaman.

Funding acquisition: Sriram Sankararaman.

Investigation: Sriram Sankararaman.

Methodology: Arun Durvasula, Sriram Sankararaman.

Software: Arun Durvasula.

Supervision: Sriram Sankararaman.

Writing – original draft: Arun Durvasula, Sriram Sankararaman.

Writing – review & editing: Arun Durvasula, Sriram Sankararaman.

References

1. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016; 538(7624):201. https://doi.org/10.1038/nature18964 PMID: 27654912
2. Sankararaman S, Mallick S, Dannemann M, Prüfer K, Kelso J, Pääbo S, et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature. 2014; 507(7492):354–357. https://doi.org/10.1038/nature12961 PMID: 24476815
3. Vernot B, Akey JM. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. Science. 2014; 343(6174):1017–1021. https://doi.org/10.1126/science.1245938 PMID: 24476670
4. Simonti CN, Vernot B, Bastarache L, Bottinger E, Carrell DS, Chisholm RL, et al. The phenotypic legacy of admixture between modern humans and Neandertals. Science. 2016; 351(6274):737–741. https://doi.org/10.1126/science.aad2149 PMID: 26912863
5. McCoy RC, Wakefield J, Akey JM. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. Cell. 2017; 168(5):916–927.e12. https://doi.org/10.1016/j.cell.2017.01.038 PMID: 28235201
6. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010; 328(5979):710–722. https://doi.org/10.1126/science.1188021 PMID: 20448178
7. Mendez FL, Watkins JC, Hammer MF. Neandertal origin of genetic variation at the cluster of OAS immunity genes. Molecular Biology and Evolution. 2013; 30(4):798–801. https://doi.org/10.1093/molbev/mst004 PMID: 23315957
8. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012; 192(3):1065–1093. https://doi.org/10.1534/genetics.112.145037 PMID: 22960212
9. Plagnol V, Wall JD. Possible Ancestral Structure in Human Populations. PLOS Genetics. 2006; 2(7):e105. https://doi.org/10.1371/journal.pgen.0020105 PMID: 16895447
10. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 2014; 505(7481):43–49. https://doi.org/10.1038/nature12886 PMID: 24352235
11. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas AS, Manica A, Moltke I, et al. Genomic structure in Europeans dating back at least 36,200 years. Science. 2014; 346(6213):1113–1118. https://doi.org/10.1126/science.aaa0114 PMID: 25378462
12. Skov L, Hui R, Shchur V, Hobolth A, Scally A, Schierup MH, et al. Detecting archaic introgression using an unadmixed outgroup. PLOS Genetics. 2018; 14(9):e1007641. https://doi.org/10.1371/journal.pgen.1007641 PMID: 30226838
13. Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature. 2010; 468(7327):1053–1060. https://doi.org/10.1038/nature09710 PMID: 21179161
14. Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences. 2011; 108(37):15123–15128. https://doi.org/10.1073/pnas.1109300108
15. Lachance J, Vernot B, Elbers CC, Ferwerda B, Froment A, Bodo JM, et al. Evolutionary History and Adaptation from High-Coverage Whole-Genome Sequences of Diverse African Hunter-Gatherers. Cell. 2012; 150(3):457–469. https://doi.org/10.1016/j.cell.2012.07.009 PMID: 22840920
16. Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, et al. Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. Genome Research. 2016. https://doi.org/10.1101/gr.196634.115
17. Hajdinjak M, Fu Q, Hübner A, Petr M, Mafessoni F, Grote S, et al. Reconstructing the genetic history of late Neanderthals. Nature. 2018; 555(7698):652–656. https://doi.org/10.1038/nature26151 PMID: 29562232
18. Slon V, Viola B, Renaud G, Gansauge MT, Benazzi S, Sawyer S, et al. A fourth Denisovan individual. Science Advances. 2017; 3(7):e1700186. https://doi.org/10.1126/sciadv.1700186 PMID: 28695206
19. Sheehan S, Song YS. Deep Learning for Population Genetic Inference. PLOS Computational Biology. 2016; 12(3):e1004845. https://doi.org/10.1371/journal.pcbi.1004845 PMID: 27018908
20. Schrider DR, Kern AD. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. PLOS Genetics. 2016; 12(3):e1005928. https://doi.org/10.1371/journal.pgen.1005928 PMID: 26977894

21. Schrider D, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. bioRxiv. 2017; p. 170670.

22. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. Cell. 2018; 173(1):53–61.e9. https://doi.org/10.1016/j.cell.2018.02.031 PMID: 29551270

23. McVicker G, Gordon D, Davis C, Green P. Widespread Genomic Signatures of Natural Selection in Hominid Evolution. PLOS Genetics. 2009; 5(5):e1000471. https://doi.org/10.1371/journal.pgen.1000471 PMID: 19424416

24. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002; 18(2):337–338. https://doi.org/10.1093/bioinformatics/18.2.337 PMID: 11847089

25. Chen H, Green RE, Pääbo S, Slatkin M. The Joint Allele-Frequency Spectrum in Closely Related Species. Genetics. 2007; 177(1):387–398. https://doi.org/10.1534/genetics.107.070730 PMID: 17603120

26. Kunsch HR. The Jackknife and the Bootstrap for General Stationary Observations. The Annals of Statistics. 1989; 17(3):1217–1241. https://doi.org/10.1214/aos/1176347265

27. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526(7571):68–74. https://doi.org/10.1038/nature15393 PMID: 26432245

28. Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. Science. 2016; p. aad9416. https://doi.org/10.1126/science.aad9416

29. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521(7553):436–444. https://doi.org/10.1038/nature14539 PMID: 26017442

30. Schrider DR, Kern AD. Machine Learning for Population Genetics: A New Paradigm. bioRxiv. 2017; p. 206482.

31. Chan J, Perrone V, Spence JP, Jenkins PA, Mathieson S, Song YS. A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks. bioRxiv. 2018; p. 267211.

32. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. ICML'06. New York, NY, USA: ACM; 2006. p. 233–240. Available from: http://doi.acm.org/10.1145/1143844.1143874.

33. Watterson GA. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology. 1975; 7(2):256–276. https://doi.org/10.1016/0040-5809(75)90020-9 PMID: 1145509

34. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449(7164):851–861. https://doi.org/10.1038/nature06258 PMID: 17943122

35. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

Chapter 2: Recovering signals of ghost archaic introgression in African populations

HUMAN GENETICS

Recovering signals of ghost archaic introgression in African populations

Arun Durvasula[1] and Sriram Sankararaman[1,2,3,4]*

While introgression from Neanderthals and Denisovans has been documented in modern humans outside Africa, the contribution of archaic hominins to the genetic variation of present-day Africans remains poorly understood. We provide complementary lines of evidence for archaic introgression into four West African populations. Our analyses of site frequency spectra indicate that these populations derive 2 to 19% of their genetic ancestry from an archaic population that diverged before the split of Neanderthals and modern humans. Using a method that can identify segments of archaic ancestry without the need for reference archaic genomes, we built genome-wide maps of archaic ancestry in the Yoruba and the Mende populations. Analyses of these maps reveal segments of archaic ancestry at high frequency in these populations that represent potential targets of adaptive introgression. Our results reveal the substantial contribution of archaic ancestry in shaping the gene pool of present-day West African populations.

INTRODUCTION

Admixture has been a dominant force in shaping patterns of genetic variation in human populations (*1*). Comparisons of genome sequences from archaic hominins to those from present-day humans have documented multiple interbreeding events, including gene flow from Neanderthals into the ancestors of all non-Africans (*2*), from Denisovans into Oceanians (*3*) and eastern non-Africans (*4*, *5*), as well as from early modern humans into the Neanderthals (*6*). However, the sparse fossil record and the difficulty in obtaining ancient DNA have made it challenging to dissect the contribution of archaic hominins to genetic diversity within Africa. While several studies have revealed contributions from deep lineages to the ancestry of present-day Africans (*7*–*12*), the nature of these contributions remains poorly understood.

RESULTS

We leveraged whole-genome sequence data from present-day West African populations and archaic hominins to compute statistics that are sensitive to introgression in the history of these populations. Specifically, we tabulated the distribution of the frequencies of derived alleles (where a derived allele is determined relative to an inferred human ancestor) in the analyzed African populations at single-nucleotide polymorphisms (SNPs) for which a randomly sampled allele from an archaic individual was observed to also be derived. Theory predicts that this conditional site frequency spectrum (CSFS) is expected to be uniformly distributed when alleles are neutrally evolving under a demographic model in which the ancestor of modern and archaic humans, assumed to be at mutation-drift equilibrium, split with no subsequent gene flow between the two groups (*13*, *14*). This expectation is robust to assumptions about changes in population sizes in the history of modern human or archaic populations. Further, we show that this expectation holds even when there is population structure or gene flow in the history of the archaic population (see Materials and Methods).

We computed $CSFS_{YRI,N}$: the CSFS in the Yoruba from Ibadan (YRI) while restricting to SNPs where a randomly sampled allele from the high-coverage Vindija Neanderthal (N) genome was observed to be derived (*15*). In contrast to the uniform spectrum expected from theory, we observe that the $CSFS_{YRI,N}$ has a U-shape with an elevated proportion of SNPs with low- and high-frequency–derived alleles relative to those at intermediate frequencies (Fig. 1 and fig. S4). The CSFS is nearly identical when we replace the Vindija Neanderthal genome with the high-coverage Denisova genome (Fig. 1 and fig. S4) (*4*). We observed a similar U-shaped CSFS in each of three additional West African populations [Esan in Nigeria (ESN), Gambian in Western Divisions in the Gambia (GWD), and Mende in Sierra Leone (MSL)] included in the 1000 Genomes Phase 3 dataset (fig. S4).

Mutational biases, errors in determining either the ancestral or the archaic allele, or recurrent mutation could produce the observed CSFS. We confirmed that the shape of the $CSFS_{YRI,N}$ was robust to the inclusion of only transition mutations, only transversion mutations, to the exclusion of hypermutable CpG sites (fig. S7), as well as when we computed the spectrum on the Yoruba genomes separately sequenced in the 1000 Genomes Phase 1 dataset (fig. S7).

We verified that this signal was robust to changes in recombination rate and background selection by restricting to regions that are likely to be evolving neutrally (by restricting to sites with estimates of background selection, *B* statistic, >800). We also assessed the effect of biased gene conversion by excluding weak-to-strong and strong-to-weak polymorphisms. We found that the U-shaped signal is robust to variation in recombination rate, background selection, and biased gene conversion (fig. S10). Errors in determining the ancestral allele could make low-frequency ancestral alleles appear to be high-frequency–derived alleles and vice versa and thus could potentially lead to a U-shaped CSFS. However, the shape of the CSFS remains qualitatively unchanged when we used either the chimpanzee genome or the consensus across the orangutan and chimpanzee genomes to determine the ancestral allele (fig. S9). We simulated both ancestral allele misidentification and errors in genotype calling in the high-coverage archaic genome. A fit to the data required both a 15% ancestral misidentification rate and a 3% genotyping error rate in the archaic genome, substantially larger than previous estimates of these error rates [1% for ancestral

[1]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. [2]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA. [3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA.
*Corresponding author. Email: sriram@cs.ucla.edu

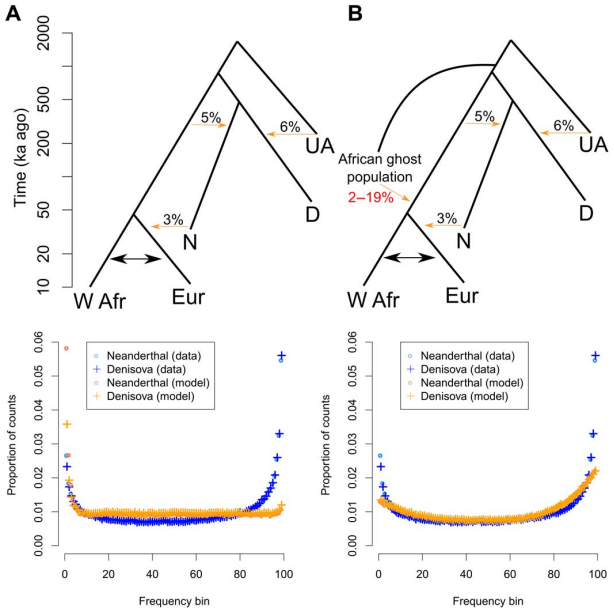

Fig. 1. Demography relating known and proposed archaic lineages to modern human populations. (**A**) Basic demographic model with CSFS fit. W Afr, West Africans; Eur, European; N, Neanderthal; D, Denisovan; UA, unknown archaic [see (*18*)]. Below, we show the CSFS in the West African YRI when restricting to SNPs where a randomly sampled allele from the high-coverage Vindija Neanderthal was observed to be derived [Neanderthal (data)], as well as where a randomly sampled allele from the high-coverage Denisovan genome was observed to be derived [Denisovan (data)]. We also show the CSFS under the proposed model [Neanderthal (model) and Denisova (model)]. Migration between Europe and West Africa introduces an excess of low-frequency variants but does not capture the decrease in intermediate frequency variants and increase in high-frequency variants. (**B**) Newly proposed model involving introgression into the modern human ancestor from an unknown hominin that separated from the human ancestor before the split of modern humans and the ancestors of Neanderthals and Denisovans. Below, we show the CSFS fit from the proposed model, which captures the U-shape observed in the data.

misidentification rate in the Enredo-Pecan-Ortheus (EPO) ancestral sequence (*16*) and 0.6% for the modern human contamination in the Vindija Neanderthal (*15*)] (section S1.1 and fig. S11). To explore the contribution of recurrent mutations, we used forward-in-time simulations that allow for recurrent mutations: The simulated CSFS does not resemble the U-shaped CSFS that we see in data (fig. S43). Together, these results indicate that the U-shaped CSFS observed in the African populations is not an artifact.

To determine whether realistic models of human history can explain the CSFS, we compared the CSFS estimated from coalescent simulations to the observed $CSFS_{YRI,N}$ using a goodness-of-fit test (see Materials and Methods and section S2). We augmented a model of the demographic history of present-day Africans (*17*) with a model of the history of Neanderthals and Denisovans inferred by Prüfer *et al.* (*15*) (Fig. 1 and figs. S1 and S16). This model includes key interbreeding events between Neanderthals, Denisovans, and modern human populations such as the introgression from Neanderthals into non-Africans, from early modern humans into Neanderthals (*6*), and into the Denisovans from an unknown archaic population (*18*). The resulting model fails to fit the observed $CSFS_{YRI,N}$ [*P* value of a Kolmogorov-Smirnov (KS) test on the residuals being normally distributed $P < 2 \times 10^{-16}$]. Extensions of this model to include realistic variation in mutation and recombination rates along the genome (KS $P < 2 \times 10^{-16}$; fig. S12 and section S1) and low levels of Neanderthal DNA introduced into African populations via migration between Europeans and Africans do not provide an adequate fit (KS $P < 2 \times 10^{-16}$; Fig. 1 and section S1) nor does a model of gene flow between YRI and pygmy populations that has been proposed previously (KS $P < 2 \times 10^{-16}$; fig. S12 and section S1) (*19*). The expectation that the CSFS is uniformly distributed across allele frequencies relies on an assumption of mutation-drift equilibrium in the population ancestral to modern humans, Neanderthals, and Denisovans. We confirmed that violations of this assumption (due to bottlenecks, expansions, and population structure in the ancestral population) were also unable to fit the data (KS $P < 2 \times 10^{-16}$ for all models; section S2, table S3, and fig. S17).

Given that none of the current demographic models are able to fit the observed CSFS, we explored models where present-day West Africans trace part of their ancestry to (A) a population that split from their ancestors after the split between Neanderthals and modern humans, (B) a population that split from the ancestor of Neanderthals after the split between Neanderthals and modern humans, or (C) a population that diverged from the ancestors of modern humans and Neanderthals before the ancestors of Neanderthals and modern humans split from each other (fig. S2 and section S3). Each of these models of admixture (which we refer to as models A, B, and C, respectively) can yield a U-shaped CSFS. The increase in the counts of low derived allele frequency SNPs is largely due to the introduction of the derived allele from the introgressing population at sites that are fixed for the ancestral allele. The increase in the counts of the high-frequency SNPs is largely due to the introduction of the ancestral alleles at sites that are fixed for the derived allele.

A search for the parameters for models A and B that produce the best fit to the CSFS results in a trifurcation, i.e., models in which the introgressing population splits off from the modern human population at the same time as the modern human–Neanderthal. Models A and B fail to fit the observed CSFS even at their most likely parameter estimates (KS $P = 3.3 \times 10^{-15}$ and $P = 5.6 \times 10^{-6}$, respectively; section S3) because of insufficient genetic drift in the African population since the split from the introgressing population (section S4.2). In addition, we show in appendix B that the spectrum for model A is expected to be symmetric, which is not observed in the data (Fig. 1). Model C, on the other hand, is consistent with the data (KS $P = 0.09$), suggesting that part of the ancestry of present-day West Africans must derive from a population that diverged before the split time of Neanderthals and modern humans. In addition to the goodness-of-fit tests, we examined the likelihood of the best-fit parameters for each of the models and found that model C provides a significantly better fit than other models (model C having a higher composite log likelihood than the next best model $\Delta\mathcal{LL} = \mathcal{LL}_{Nextbestmodel} - \mathcal{LL}_C = -6806$ when we condition on the Vindija Neanderthal genome and $\Delta\mathcal{LL} = -6240$ when we condition on the Denisovan genome; table S4 and Materials and Methods). Our analyses provide support for a contribution to the genetic ancestry of present-day West African populations from an archaic ghost population whose divergence from the ancestors of modern humans predates the split of Neanderthals and modern humans.

We applied approximate Bayesian computation (ABC) to the CSFS to refine the parameters of our most likely demographic model (model C) (section S5). Given the large number of parameters in this demographic

model, we fixed parameters that had previously been estimated (*15*) and jointly estimated the split time of the introgressing archaic population from the ancestors of Neanderthals and modern humans, the time of introgression, the fraction of ancestry contributed by the introgressing population, and its effective population size. We determined the posterior mean for the split time to be 625,000 years before the present (B.P.) [95% highest posterior density interval (HPD): 360,000 to 975,000], the admixture time to be 43,000 years B.P. (95% HPD: 6000 to 124,000), and the admixture fraction to be 0.11 (95% HPD: 0.045 to 0.19). Analyses of three other West African populations (ESN, GWD, and MSL) yielded concordant estimates for these parameters (Fig. 2 and table S7). Combining our results across the West African populations, we estimate that the archaic population split from the ancestor of Neanderthals and modern humans 360 thousand years (ka) to 1.02 million years (Ma) B.P. and subsequently introgressed into the ancestors of present-day Africans 0 to 124 ka B.P. contributing 2 to 19% of their ancestry. We caution that the true underlying demographic model is likely to be more complex. To explore aspects of this complexity, we examined the possibility that the archaic population diverged at the same time as the split time of modern humans and Neanderthals and found that this model can also produce a U-shaped CSFS with a likelihood that is relatively high, although lower than that of our best-fit model ($\Delta\mathcal{LL}$ = −2713 for the Neanderthal CSFS and $\Delta\mathcal{LL}$ = −2597 for the Denisovan CSFS, KS $P \leq 2.9 \times 10^{-6}$). Our estimates of a large effective population size in the introgressing lineage (posterior mean of 25,000; 95% HPD: 23,000 to 27,000) could indicate additional structure. We find that the N_e of the introgressing lineage in YRI and MSL is larger than that in the other African populations, possibly due to a differential contribution from a basal West African branch (*20*).

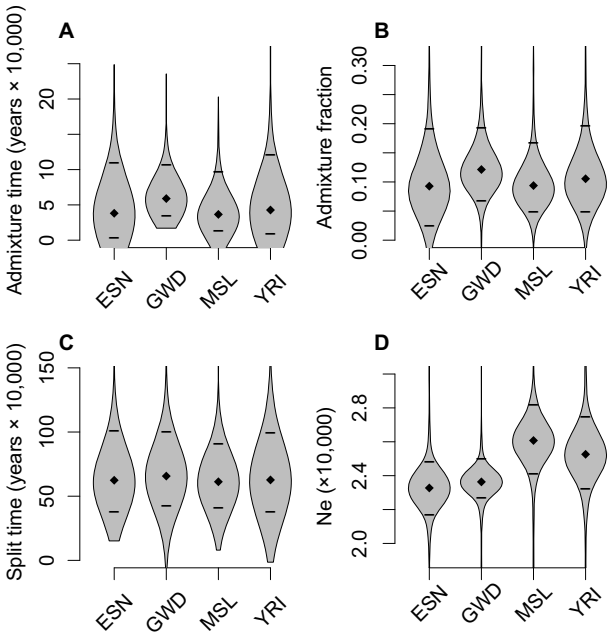

Fig. 2. ABC estimates of the demographic parameters of the archaic ghost population across four West African populations (YRI, ESN, GWD, and MSL). Posterior means are denoted by diamonds, and 95% credible intervals are denoted by lines. (**A**) The admixture time t_a, (**B**) the admixture fraction α, (**C**) the split time of the introgressing population t_s, and (**D**) the effective population size of the introgressing population N_e are shown. The parameter estimates are largely consistent across the African populations: We estimate split times of 360 ka to 1.02 Ma B.P., admixture times of 0 to 124 ka B.P., admixture fractions that range from 0.02 to 0.19, and effective population sizes that range from 22,000 to 28,000.

While we have chosen to represent the genetic contribution of the African ghost population as a single discrete interbreeding event, a more realistic model could include low levels of gene flow in a structured population over an extended period of time. Previously proposed models of ancestral structure in Africa do not fit the CSFS [KS $P < 2 \times 10^{-16}$ for the model described in (*21*) and KS $P < 2 \times 10^{-16}$ for the model proposed in (*14*); fig. S18], although we observe that the model of ancestral structure proposed by Yang *et al.* does produce a slight U-shape. We explored additional models of population structure in Africa (*22*) in which a lineage split from the ancestor of the modern humans with split times ranging from 100 to 550 ka B.P. and continued to exchange genes with the modern human population until the present with migration rates ranging from 2.5×10^{-5} to 2×10^{-2} migrants per generation. While these models of continuous gene flow produce a U-shaped CSFS for low migration rates and deep splits, they do not provide an adequate fit to the empirical CSFS over the range of parameters considered (KS $P \leq 2.3 \times 10^{-5}$; section S6 and figs. S14 and S15). We used our ABC framework to explore a more detailed model of continuous migration in which we varied split time, migration rate, and effective population size of the introgressing lineage. Simulations under the best fitting model produce a CSFS that does not adequately fit the data (KS $P = 1.83 \times 10^{-6}$). A possible reason why the continuous migration models that we have explored do not fit the data is that these models can be considered as extensions of model A with multiple admixture events. We have shown that these models can only produce symmetric CSFS, unlike the CSFS that we observe in the data (appendix B). Thus, deep population structure within Africa alone cannot not explain the data (section S6).

Given the uncertainty in our estimates of the time of introgression, we wondered whether jointly analyzing the CSFS from both the CEU (Utah residents with Northern and Western European ancestry) and YRI genomes could provide additional resolution. Under model C, we simulated introgression before and after the split between African and non-African populations and observed qualitative differences between the two models in the high-frequency–derived allele bins of the CSFS in African and non-African populations (fig. S40). Using ABC to jointly fit the high-frequency–derived allele bins of the CSFS in CEU and YRI (defined as greater than 50% frequency), we find that the lower limit on the 95% credible interval of the introgression time is older than the simulated split between CEU and YRI (2800 versus 2155 generations B.P.), indicating that at least part of the archaic lineages seen in the YRI are also shared with the CEU (section S9.2).

We then attempted to understand the fine-scale distribution of archaic ghost ancestry along the genomes of present-day Africans. We used a recently developed statistical method (ArchIE) that combines multiple population genetic statistics to identify segments of diverged ancestry in 50 YRI and 50 MSL genomes without the need for an archaic reference genome (section S7) (*23*). Briefly, the method uses summary statistics computed from present-day genome sequences as input to a logistic regression model to estimate the probability that a haploid segment of an individual genome (defined as a contiguous region of length 50 kilobases) is archaic. While the parameters of the model are estimated by simulating data under a model that closely matches the demographic history relating Neanderthals and non-Africans, we

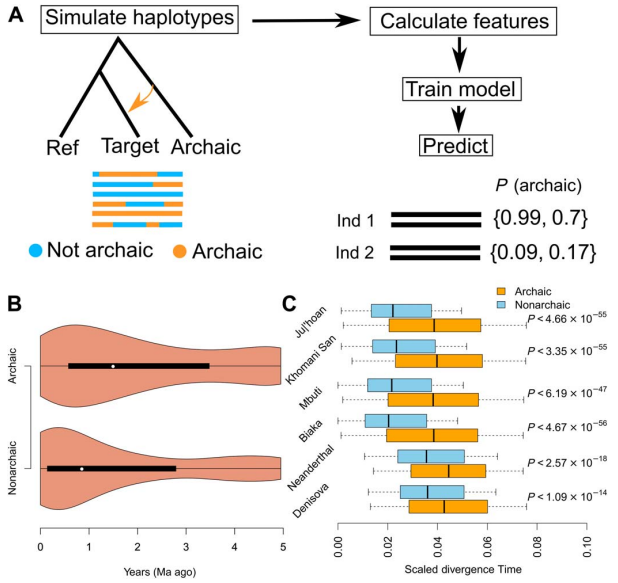

Fig. 3. Analysis of segments of archaic ghost ancestry found in the Yoruba and Mende populations. (**A**) Inference of segments of archaic ancestry was performed with ArchIE. ArchIE proceeds by simulating data under a model of archaic introgression, calculating population genetic summary statistics, and training a model to predict the probability that a 50-kb window in an individual comes from an archaic population. We apply the resulting predictor to genome sequences from the Yoruba and Mende populations. (**B**) Comparison of TMRCA between inferred archaic and nonarchaic segments to the TMRCA of a pair of nonarchaic segments in the Yoruba. On average, archaic segments are 1.69× older than nonarchaic segments. (**C**) Estimates of the divergence times of archaic segments inferred in Yoruba from KhoeSan, Jul'hoan, two modern human pygmy genomes (Mbuti and Biaka), and Neanderthal and Denisovan genomes compared to divergence times of nonarchaic segments. *P* values are computed via block jackknife. Archaic segments are more diverged from all six genomes than nonarchaic segments.

Table 1. Genes harboring a high frequency of archaic segments in the Yoruba and Mende populations. Genes were selected by ranking the union of the set of putative archaic segments by frequency in either the Mende or Yoruba population and selecting the top 10 genes. Genes in bold denote frequencies greater than 50% in the respective population.

Chromosome	Gene name	Frequency (Yoruba)	Frequency (Mende)	Gene type
chr1	RP11-286M16.1	**0.84**	**0.81**	lincRNA
chr4	KCNIP4	**0.73**	**0.69**	Protein coding
chr6	MTFR2	**0.67**	**0.78**	Protein coding
chr8	TRPS1	**0.71**	**0.75**	Protein coding
chr12	RP11-125N22.2	0.12	**0.88**	Pseudogene
chr16	HSD17B2	**0.74**	**0.68**	Protein coding
chr17	NF1	**0.83**	**0.85**	Protein coding
chr17	KRT18P61	**0.84**	0.36	Pseudogene
chr21	MIR125B2	**0.76**	**0.64**	MicroRNA

found that ArchIE has 68% power to detect archaic segments at a false discovery rate of about 7% under our best-fit demographic model, confirming that its inferences are robust and sensitive to archaic introgression in Africa.

On average, ≃6.6 and ≃7.0% of the genome sequences in YRI and MSL were labeled as putatively archaic in ancestry. We sought to test whether the putatively archaic segments identified in YRI and MSL traced their primary ancestry to other African populations (*8*–*10*) or to known archaic hominins such as the Neanderthals or Denisovans. We computed the divergence of these segments to a genome sequence from each of six populations: southern African KhoeSan, Jul'hoan; two Central African pygmy populations (Biaka and Mbuti); and two archaic hominin populations (Neanderthal and Denisovan). We expect segments introgressed from any of these populations to be less diverged relative to nonarchaic segments. On the contrary, the putatively archaic segments are more diverged, consistent with their source not being any of these populations (Fig. 3C and section S7.1). Merging the putatively archaic segments across individual genomes, we obtained a total of 482 and 502 Mb of archaic genome sequence in the YRI and MSL, respectively. We estimated the distribution of the time to the most recent common ancestor (TMRCA) between segments labeled archaic and those labeled nonarchaic using the pairwise mode of multiple sequentially Markovian coalescent (MSMC) (Fig. 3B and section S7.2) (*24*) and observed that the TMRCA is larger for the putatively archaic class of segments. Specifically, we find that the median nonarchaic segment coalescent time is 0.865 Ma ago for both populations, while the median archaic segment coalescent time is 1.51 Ma ago for YRI and 1.15 Ma ago for MSL (1.69- and 1.23-fold increases in age for YRI and MSL, respectively).

We examined the frequencies of archaic segments to investigate whether natural selection could have shaped the distribution of archaic alleles (fig. S40). We found 33 loci with an archaic segment frequency of ≥50% in the YRI (a cutoff chosen to be larger than the 99.9th percentile of introgressed archaic allele frequencies based on a neutral simulation of archaic introgression with parameters related to the time of introgression and admixture fraction chosen conservatively to maximize the drift since introgression; section S7.3 and fig. S40) and 37 loci in the MSL. Some of these genes are at high frequency across both the YRI and MSL, including *NF1*, a tumor suppressor gene (83% in YRI, 85% in MSL), *MTFR2*, a gene involved with mitochondrial aerobic respiration in the testis (67% in YRI, 78% in MSL), *HSD17B2*, a gene involved with hormone regulation (74% in YRI, 68% in MSL), *KCNIP4*, which is a gene involved with potassium channels (73% in YRI, 69% in MSL), and *TRPS1*, a gene associated with trichorhinophalangeal syndrome (71% in YRI, 75% in MSL; Table 1). Three of these genes have been found in previous scans for positive selection in the YRI: *NF1* (*25*, *26*), *KCNIP4* (*27*), and *TRPS1* (*28*). On the other hand, we do not find elevated frequencies at *MUC7*, a gene previously found to harbor signatures of archaic introgression (*29*).

DISCUSSION

Our analyses document introgression in four present-day West African populations from an archaic population that likely diverged before the split of modern humans and the ancestors of Neanderthals and Denisovans. A number of previous studies have found evidence for

deeply diverged lineages contributing genetic ancestry to the Pygmy (*8*, *9*) and Yoruba (*7*, *30*) populations. Analyses of ancient African genomes have revealed that stone-age hunter-gatherers from South Africa diverged from other modern-day populations >260,000 years (*31*) B.P. and that present-day West African populations trace part of their ancestry to a basal lineage that diverged before the split of the southern African San (*20*) (although an alternative model consistent with their data includes a complex pattern of isolation by distance between western, eastern, and southern African populations). Placing our results within the context of the complex patterns of deep divergences in the African populations will require the analysis of a diverse set of African populations that include the southern African San populations, as well as the inclusion of ancient African genomes that lack signals of recent admixture that are present in the present-day San populations (*32*).

One interpretation of the recent time of introgression that we document is that archaic forms persisted in Africa until fairly recently (*33*). Alternately, the archaic population could have introgressed earlier into a modern human population, which then subsequently interbred with the ancestors of the populations that we have analyzed here. The models that we have explored here are not mutually exclusive, and it is plausible that the history of African populations includes genetic contributions from multiple divergent populations, as evidenced by the large effective population size associated with the introgressing archaic population. Relatively, recent fossils with archaic features (or combinations of archaic and modern human features) have been found in the fossil record in Africa and the Middle East. While anatomically modern humans appear in the fossil record around 200,000 years ago, fossils with a combination of archaic and modern features can be found across sub-Saharan Africa and the Middle East until as recently as 35,000 years ago (*34*). Examples of these fossils include a cranium from Iwo Eleru (*33*) and human remains from Ishango (*35*) that have been interpreted as being consistent with deep structure and representing a complex history of interaction between modern and archaic hominins in Africa.

The signals of introgression in the West African populations that we have analyzed raise questions regarding the identity of the archaic hominin and its interactions with the modern human populations in Africa. Analysis of the CSFS in the Luhya from Webuye, Kenya (LWK) also reveals signals of archaic introgression, although our interpretation is complicated by recent admixture in the LWK that involves populations related to western Africans and eastern African hunter-gatherers (section S8) (*20*). Non-African populations (Han Chinese in Beijing and Utah residents with northern and western European ancestry) also show analogous patterns in the CSFS, suggesting that a component of archaic ancestry was shared before the split of African and non-African populations. A detailed understanding of archaic introgression and its role in adapting to diverse environmental conditions will require analysis of genomes from extant and ancient genomes across the geographic range of Africa.

MATERIALS AND METHODS

Conditional site frequency spectrum

We define the CSFS, $\mathrm{CSFS}_{\mathrm{YRI,N}}$, as the histogram of the counts of derived alleles in population pop_1 conditional on observing a derived allele in a related outgroup pop_2 (*13*). We define c_k as the number of SNPs at which the derived allele is present on k chromosomes in a sample of n total chromosomes in pop_1, while a single chromosome in the outgroup pop_2 carries a derived allele. $\mathrm{CSFS}_{\mathrm{YRI,N}}$ is the vector of counts c_k for $k \in \{1 \ldots n-1\}$.

Chen *et al.* (*13*) showed that if the ancestor of populations pop_1 and pop_2 is at mutation-drift equilibrium (i.e., the site frequency spectrum in the ancestor is $f(x) \propto \frac{1}{x}$, where $0 < x < 1$ is the derived allele frequency at a polymorphic SNP) and the two populations pop_1 and pop_2 split with no subsequent admixture, then the $\mathrm{CSFS}_{\mathrm{YRI,N}}$ is expected to be uniform, i.e., $\mathrm{CSFS}_{\mathrm{YRI,N}}(k) = \text{constant}$. This result does not depend on any additional aspects of the demographic history of either populations pop_1 or pop_2, except that they are randomly mating. We used the CSFS to study introgression in present-day Africans where we set pop_1 to present-day Africans and pop_2 to an archaic population, i.e., Neanderthal or Denisovan.

One of the complications in applying the CSFS to learn about the history of present-day Africans arises from known departures from a simple model of isolation with no subsequent admixture. However, we considered the possibility of structure in the archaic population. This structure could have several forms that include the ancestral Neanderthal population being structured or it could involve gene flow from early modern humans into Neanderthals (*6*), or as in the case of Denisovans, this could include gene flow from a highly diverged archaic population (*18*). We performed extensive simulations to show that structure in the archaic population continues and also leads to a uniform CSFS (section S1). Further, in appendix A, we show that the CSFS is uniform even if there is structure in the archaic population. However, structure within population the African population (pop_1) since its split from the archaic population (pop_2), e.g., due to admixture, is expected to produce deviations from the uniform CSFS.

Data processing

For our primary analyses of the CSFS, we used the 1000 Genomes Phase 3 dataset (release 20130502) (*36*), the high-coverage Vindija Neanderthal genome (*15*), and the high-coverage Denisovan genome (*4*). We used the annotated ancestral alleles provided by the 1000 Genomes consortium and analyzed only autosomal SNPs. Archaic genotypes (Vindija and Denisovan) come from the pipeline described in (*15*), which used snpAD for SNP calling [see S3 in (*15*)], and required a mapping quality of ≥25 and a mappability filter of 100. We did not apply an additional genotype quality filter for the data presented in fig. S4. However, we tested the sensitivity of the spectrum to the choice of genotype quality filters in the archaic when using a GQ (Genotype Quality) filter of ≥30 and ≥50 and see very little difference in the shape of the spectrum (fig. S8).

In addition, we also computed the CSFS using the chimpanzee genome to polarize the ancestral alleles (fig. S9A) (*37*). We dropped sites in cases where the chimpanzee allele did not match either human allele. As a further check, we also repeated the analysis restricting only to sites where the chimpanzee and orangutan genomes have matching alleles (*38*). These results are reported in fig. S9B. Last, we repeated our analysis filtering out CpG hypermutable sites using the CpG annotations from (*18*).

CSFS from the 1000 Genomes data

We computed $\mathrm{CSFS}_{\mathrm{YRI,N}}$ where pop_1 is a modern human population and pop_2 is an archaic population. Specifically, we chose pop_1, in turn, to be the Yoruba from Nigeria (YRI), MSL, ESN, and GWD, while we chose pop_2 to be either the high-coverage Vindija Neanderthal or the high-coverage Denisovan genome (fig. S4).

We computed the CSFS from the 1000 Genomes phase 3 data (*36*) for each of the four African populations mentioned above (fig. S4), as well as for the CEPH CEU and Han Chinese from Beijing (CHB) (fig. S6).

For all populations, we observed a U-shaped spectrum with an excess of derived alleles at low and high frequencies. In the African populations, we observed that the CSFS from conditioning on the Denisovan is nearly identical to the Vindija Neanderthal except at the lowest-frequency bins, where there is an excess of counts for the Neanderthal CSFS. We interpreted this difference as suggestive of low levels of Neanderthal-related ancestry in these populations consistent with previous studies (*18*). In CEU and CHB, we also observed a U-shaped spectrum for both the Vindija Neanderthal and Denisovan, but with a more pronounced difference between the Neanderthal and Denisovan spectra, i.e., an excess of counts in the low-frequency–derived sites when conditioned on the Vindija Neanderthal relative to the Denisovan. This difference is likely reflective of the Neanderthal introgression event experience by populations outside of Africa around 50,000 years ago (*21*, *39*). Section S8 explores the implication of observing a U-shaped CSFS in African and non-African populations.

To determine the robustness of the shape of the CSFS, we recomputed the CSFS in YRI using only transitions, transversions, and after removing CpG sites. We found very similar U-shaped CSFS across these mutation classes (fig. S7). In addition, we checked whether biased gene conversion could cause this signal by removing weak-to-strong and strong-to-weak polymorphisms. We found that the shape of the CSFS remains without these mutations (fig. S10A). Last, we checked whether the shape of the CSFS was driven by selection or low recombination rates. We used *B* values from (*40*), which estimate how much background selection has reduced diversity. We restricted to regions of the genome in the top quintile of *B* values (that is, the top one-fifth of neutral sites; $B \geq 800$) and recomputed the spectrum using YRI individuals. We found that the shape remains the same after this filtering (fig. S10B).

Model comparison

We used coalescent simulations to assess whether a demographic model produces a CSFS that matches the empirical CSFS. To assess the fit of a given demographic model $\mathcal{M}$ to the data, we compared the CSFS computed on the data simulated under $\mathcal{M}$ to that computed on the empirical data. We considered a model in which the empirical CSFS was obtained by sampling from the CSFS computed on the simulated data. For these fits, we modeled the proportion of SNPs that contain a given number k of derived alleles rather than the number of SNPs. To assess the fit of the simulated CSFS under $\mathcal{M}$ ($\boldsymbol{S}_{\mathcal{M}}$) to the observed CSFS ($\boldsymbol{O}$), we used a multinomial composite likelihood

$$L(\mathcal{M}) = P(\boldsymbol{O} \mid \boldsymbol{S}_{\mathcal{M}}) = \prod_{k=1}^{n-1} \left(\frac{S_k}{\sum_k S_k} \right)^{O_k}$$

Here, k indexes the derived allele count, S_k denotes the number of SNPs with k-derived alleles observed in the simulated CSFS, while O_k denotes the number of SNPs with k-derived alleles observed in the empirical CSFS. We caution that L is a composite likelihood that ignores the dependence among SNPs so that comparisons of L must be interpreted with caution. In the results presented here, we reported the log likelihood ($\mathcal{LL}$).

Goodness of fit

We defined a goodness-of-fit statistic that we used to assess whether the CSFS computed under a demographic model explains the major patterns of the empirical CSFS. The goodness-of-fit statistic was defined from the residuals obtained by trying to fit the simulated CSFS to the empirical CSFS. We assumed that the counts of SNPs in each derived allele frequency bin of the empirical CSFS follow a binomial distribution with a mean given by the proportion of SNPs that have the same derived allele frequency in the simulated CSFS. One complication is that the counts across bins of derived allele frequencies are not independent because of linkage disequilibrium. To account for this complication, we attempted to estimate the effective number of independent observations in the observed CSFS (rather than assume that each SNP is an independent observation). We define the residual for bin k as

$$r_k = \sqrt{m_{\text{eff}}} \frac{o_k - s_k}{\sqrt{s_k(1 - s_k)}}$$

Here, m_{eff} is the effective number of independent SNPs, o_k represents the proportion of SNPs with derived allele count k in the empirical CSFS, s_k is the proportion of SNPs with derived allele count k in the simulated CSFS, and k indexes the count of derived allele. These residuals are expected to be approximately normally distributed when the number of observations is large (as is the case with the CSFS where each bin has >1000 observations). m_{eff} is a scaling factor to ensure that the residuals are standardized.

To calculate m_{eff}, we used two replicate whole-genome simulations (3 GB) under the same demographic model and set one as the observed data and one as the simulation. We divided the number of bins n by the sum of the squared residuals

$$m_{\text{eff}} = \frac{n}{\sum_{k=1}^{n} \left(\frac{o_k - s_k}{\sqrt{s_k(1-s_k)}} \right)^2}$$

A good fit will result in approximately normally distributed residuals, while poor fits will deviate significantly from a normal distribution. To obtain a formal test of fit, we used a KS test comparing the distribution of the residuals to a normal distribution. *P* values that reject the null hypothesis suggest that the model is a poor fit to the data. We used bins of allele counts ranging from 11 to 90, excluding the lowest- and highest-frequency bins as the counts from these bins are more likely to be affected by unmodeled genotyping errors, leading to false rejections of the null hypothesis. To assess the fit of a class of models (e.g., models A, B, and C), we report the *P* value of the model with parameter estimates obtained via ABC (sections S3.1 to S3.6).

Last, we expanded the range of derived allele counts in our goodness-of-fit computation from [11, 90] to [6, 95] (table S8). While none of the models fit adequately, model C has substantially higher *P* values than the other models, indicating that it continues to explain the CSFS better across this range of allele counts. The lack of fit across the expanded range of derived allele counts is likely due to unmodeled complexities in the underlying demographic history, as well as error processes that affect the low- and high-frequency SNPs.

Model fitting

We used ABC to fit a demographic model to the CSFS of each African population using the R package abc (*41*). Using a model relating African and non-African populations with the Neanderthal and Denisovan

lineages as a base, we fit the split time, admixture time, admixture fraction, and effective population size of an introgressing lineage (section S5.2). We drew values for each of the parameters from a previous distribution, simulated 300 Mb using ms (*42*), and computed the CSFS for the resulting simulation. We repeated this procedure 75,000 times. We used the "neuralnet" setting in the R package abc to compute posterior distributions over each of the four parameters with a tolerance of 0.005. For the admixture time and split time, we report the posterior distributions in units of years by convolving the posterior generation time with a uniform distribution over [25, 33] to incorporate uncertainty in the generation time.

Local ancestry inference

We used ArchIE (*23*) to infer the segments of the genomes in 50 YRI and 50 MSL individuals who likely trace their ancestry to an archaic population. We trained ArchIE on a model where an archaic population splits 12,000 generations B.P. and introgressed 2000 generations B.P. at a 2% admixture fraction (section S7). We computed the coalescent time for segments we classified as archaic and segments we classified as nonarchaic using the posterior decoding from MSMC using a representative individual from both YRI and MSL (*24*). We also computed the scaled divergence time between archaic and nonarchaic segments with test genomes from hunter-gatherer populations, Central African Pygmy populations, and archaic populations. This scaled divergence was computed as the number of mutations specific to the segment subtracted from the number of mutations shared between the segment and the test genome. We divided this number by the number of segregating sites in the segment to normalize by the local mutation rate.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/7/eaax5097/DC1

Section S1. Current demographic models cannot explain the CSFS
Section S2. The CSFS cannot be explained by departures from panmixia in the ancestor of archaics and modern humans
Section S3. Exploration of models of introgression into the ancestors of present-day Africans
Section S4. Parameter exploration of model A
Section S5. Estimating parameters for the best-fit model of archaic introgression
Section S6. Continuous migration versus a single pulse
Section S7. Local ancestry inference
Section S8. Extended discussion
Section S9. ms command lines
Fig. S1. Demographic model from Prüfer *et al.* (*15*) (see section S1 for details).
Fig. S2. Demographic model topologies for introgression into the ancestors of present-day Africans simulations in figs. S20, S22, S24, S26, S28, and S30.
Fig. S3. Demographic model topologies for mathematical results.
Fig. S4. CSFS from 1000 Genomes Phase 3 data across all African populations included in the dataset.
Fig. S5. CSFS from 1000 Genomes Phase 3 data in the Luhya population.
Fig. S6. CSFS from 1000 Genomes Phase 3 data in the CEU and CHB.
Fig. S7. Robustness of CSFS in YRI across mutation types and the Phase 1 1000 Genomes dataset.
Fig. S8. Robustness of CSFS in YRI to genotype quality thresholds in archaic genomes.
Fig. S9. CSFS in YRI when using alternate sources for the ancestral allele.
Fig. S10. CSFS in YRI when controlling for biased gene conversion and background selection.
Fig. S11. Simulations of the baseline model (section S1) with both ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S12. Mutation rate and recombination rate variation.
Fig. S13. Simulations of the demographic model inferred from Hsieh *et al.* (*19*) relating the Yoruba, Baka, and Biaka populations.
Fig. S14. Simulations of a demographic model with structure and gene flow in Africa.
Fig. S15. Models with continuous migration (*m* in units of migrants per generation) since the introgressing lineages lineage splits.
Fig. S16. Current demographic models from the literature cannot explain the shape of the CSFS observed in fig. S4.
Fig. S17. Models involving structure in the ancestor of modern humans and archaics cannot explain the observed CSFS.
Fig. S18. Models involving ancestral structure from the literature cannot explain the observed CSFS.
Fig. S19. Model A.1: Gene flow from the modern human ancestor branch back into the modern human ancestor before the out-of-Africa event.
Fig. S20. Model sA.1: Simplified model of gene flow from the modern human ancestor branch back into the modern human ancestor before the out-of-Africa event.
Fig. S21. Model A.2: Gene flow from the modern human ancestor branch into the African branch after the out of Africa event.
Fig. S22. Model sA.2: Simplified model of gene flow from the modern human ancestor branch into the African branch after the out of Africa event.
Fig. S23. Model B.1: Gene flow from the archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S24. Model sB.1: Simplified model of gene flow from the archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S25. Model B.2: Gene flow from the archaic branch into the African branch after the out-of-Africa event.
Fig. S26. Model sB.2: Simplified model of gene flow from the archaic branch into the African branch after the out-of-Africa event.
Fig. S27. Model C.1: Gene flow from an unknown archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S28. Model sC.1: Simplified model of gene flow from an unknown archaic branch into the modern human ancestor before the out-of-Africa event.
Fig. S29. Model C.2: Gene flow from an unknown archaic branch into the African branch after the out-of-Africa event.
Fig. S30. Model sC.2: Simplified model of gene flow from an unknown archaic branch into the African branch after the out-of-Africa event.
Fig. S31. Simulations of the best-fitting parameters for models A, B, C (section S3).
Fig. S32. Model A.2 with a population size of 0:01 N_a in the introgressing population.
Fig. S33. Model A.2 with a population size of 1×10^{-4} N_a in the introgressing population.
Fig. S34. Model A.2 with a population size of 1×10^{-4} N_a in the introgressing population and migration between CEU and YRI over the last 20 ka B.P.
Fig. S35. Model A.2 with a population size of 1×10^{-5} N_a in the introgressing population, which branches off 200 ka B.P.
Fig. S36. Model A.2 where the introgressing population splits at the same time as the archaic population (550 ka B.P.) with a population size of 0.01 N_a.
Fig. S37. Model A.2 where the introgressing population splits at the same time as the archaic population, 765 ka B.P.
Fig. S38. Parameter estimates using ABC for model A.1 including ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S39. Parameter estimates using ABC for model A.2 including ancestral misidentification (e1) and genotyping error in the archaic (e2).
Fig. S40. Marginalized joint CSFS of YRI and CEU from simulations.
Fig. S41. Distribution of allele frequencies for neutral archaic SNPs from model C with 13% introgression and an introgression time of 42 ka B.P.
Fig. S42. Archaic segment frequency map for MSL and YRI.
Fig. S43. CSFS from the baseline model allowing for recurrent mutations.
Table S1. Description of the models examined in this work.
Table S2. We simulated data from the Prüfer *et al.* (*15*) model and added in ancestral misidentification error and genotyping error in the archaic.
Table S3. Model fits for null models including structure and departures from panmixia in the Modern Human (MH) ancestor.
Table S4. Model fits for alternate models including admixture from other lineages.
Table S5. Model fits for alternate models using a simplified demography.
Table S6. Model fits for variations of model A.
Table S7. Best-fitting parameter values for all populations using ABC.
Table S8. *P* values of a test of goodness of fit for the best-fitting parameters for each class of demographic models.
Appendix A. The CSFS is uniform under structure in the archaic population.
Appendix B. The CSFS is symmetric under model A.
References (*43*–*55*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky,

S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. S. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Pääbo, J. Kelso, N. Patterson, D. Reich, The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

2. R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prüfer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Höber, B. Höffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de la Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, A draft sequence of the neandertal genome. *Science* **328**, 710–722 (2010).
3. D. Reich, R. E. Green, M. Kircher, J. Krause, N. Patterson, E. Y. Durand, B. Viola, A. W. Briggs, U. Stenzel, P. L. F. Johnson, T. Maricic, J. M. Good, T. Marques-Bonet, C. Alkan, Q. Fu, S. Mallick, H. Li, M. Meyer, E. E. Eichler, M. Stoneking, M. Richards, S. Talamo, M. V. Shunkov, A. P. Derevianko, J.-J. Hublin, J. Kelso, M. Slatkin, S. Pääbo, Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
4. M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, A high-coverage genome sequence from an archaic denisovan individual. *Science* **338**, 222–226 (2012).
5. S. R. Browning, B. L. Browning, Y. Zhou, S. Tucci, J. M. Akey, Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* **173**, 53–61.e9 (2018).
6. M. Kuhlwilm, I. Gronau, M. J. Hubisz, C. de Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, H. A. Burbano, C. Lalueza-Fox, M. de la Rasilla, A. Rosas, P. Rudan, D. Brajkovic, Ž. Kucan, I. Gušic, T. Marques-Bonet, A. M. Andrés, B. Viola, S. Pääbo, M. Meyer, A. Siepel, S. Castellano, Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433 (2016).
7. V. Plagnol, J. D. Wall, Possible ancestral structure in human populations. *PLOS Genet.* **2**, e105 (2006).
8. M. F. Hammer, A. E. Woerner, F. L. Mendez, J. C. Watkins, J. D. Wall, Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15123–15128 (2011).
9. J. Lachance, B. Vernot, C. C. Elbers, B. Ferwerda, A. Froment, J.-M. Bodo, G. Lema, W. Fu, T. B. Nyambo, T. R. Rebbeck, K. Zhang, J. M. Akey, S. A. Tishkoff, Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. *Cell* **150**, 457–469 (2012).
10. P. H. Hsieh, A. E. Woerner, J. D. Wall, J. Lachance, S. A. Tishkoff, R. N. Gutenkunst, M. F. Hammer, Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies. *Genome Res.* **26**, 279–290 (2016).
11. J. Hey, Y. Chung, A. Sethuraman, J. Lachance, S. Tishkoff, V. C. Sousa, Y. Wang, Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.* **35**, 2805–2818 (2018).
12. A. P. Ragsdale, S. Gravel, Models of archaic admixture and recent history from two-locus statistics. *PLOS Genet.* **15**, e1008204 (2019).
13. H. Chen, R. E. Green, S. Pääbo, M. Slatkin, The joint allele-frequency spectrum in closely related species. *Genetics* **177**, 387–398 (2007).
14. M. A. Yang, A.-S. Malaspinas, E. Y. Durand, M. Slatkin, Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. *Mol. Biol. Evol.* **29**, 2987–2995 (2012).
15. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, eaao1887 (2017).
16. B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, E. Birney, Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008).
17. S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs; The 1000 Genomes Project, C. D. Bustamante, Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11983–11988 (2011).
18. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
19. P. H. Hsieh, K. R. Veeramah, J. Lachance, S. A. Tishkoff, J. D. Wall, M. F. Hammer, R. N. Gutenkunst, Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. *Genome Res.* 10.1101/gr.192971.115, (2016).
20. P. Skoglund, J. C. Thompson, M. E. Prendergast, A. Mittnik, K. Sirak, M. Hajdinjak, T. Salie, N. Rohland, S. Mallick, A. Peltzer, A. Heinze, I. Olalde, M. Ferry, E. Harney, M. Michel, K. Stewardson, J. I. Cerezo-Román, C. Chiumia, A. Crowther, E. Gomani-Chindebvu, A. O. Gidna, K. M. Grillo, I. T. Helenius, G. Hellenthal, R. Helm, M. Horton, S. López, A. Z. P. Mabulla, J. Parkington, C. Shipton, M. G. Thomas, R. Tibesasa, M. Welling, V. M. Hayes, D. J. Kennett, R. Ramesar, M. Meyer, S. Pääbo, N. Patterson, A. G. Morris, N. Boivin, R. Pinhasi, J. Krause, D. Reich, Reconstructing prehistoric african population structure. *Cell* **171**, 59–71.e21 (2017).
21. S. Sankararaman, N. Patterson, H. Li, S. Pääbo, D. Reich, The date of interbreeding between neandertals and modern humans. *PLOS Genet.* **8**, e1002947 (2012).
22. B. M. Henn, T. E. Steele, T. D. Weaver, Clarifying distinct models of modern human origins in Africa. *Curr. Opin. Genet. Dev.* **53**, 148–156 (2018).
23. A. Durvasula, S. Sankararaman, A statistical model for reference-free inference of archaic local ancestry. *PLOS Genet.* **15**, e1008175 (2019).
24. S. Schiffels, R. Durbin, Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
25. S. Kudaravalli, J.-B. Veyrieras, B. E. Stranger, E. T. Dermitzakis, J. K. Pritchard, Gene expression levels are a target of recent natural selection in the human genome. *Mol. Biol. Evol.* **26**, 649–658 (2009).
26. S. R. Grossman, K. G. Andersen, I. Shlyakhter, S. Tabrizi, S. Winnicki, A. Yen, D. J. Park, D. Griesemer, E. K. Karlsson, S. H. Wong, M. Cabili, R. A. Adegbola, R. N. K. Bamezai, A. V. S. Hill, F. O. Vannberg, J. L. Rinn; 1000 Genomes Project, E. S. Lander, S. F. Schaffner, P. C. Sabeti, Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
27. International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallée, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archevêque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, J. Stewart, A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
28. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**, 340–345 (2008).
29. D. Xu, P. Pavlidis, R. O. Taskent, N. Alachiotis, C. Flanagan, M. DeGiorgio, R. Blekhman, S. Ruhl, O. Gokcumen, Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. *Mol. Biol. Evol.* **34**, 2704–2715 (2017).

30. J. D. Wall, K. E. Lohmueller, V. Plagnol, Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–1827 (2009).
31. C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, M. Jakobsson, Southern african ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
32. J. K. Pickrell, N. Patterson, P.-R. Loh, M. Lipson, B. Berger, M. Stoneking, B. Pakendorf, D. Reich, Ancient west eurasian ancestry in southern and eastern africa. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 2632–2637 (2014).
33. K. Harvati, C. Stringer, R. Grün, M. Aubert, P. Allsworth-Jones, C. A. Folorunso, The later stone age calvaria from Iwo Eleru, Nigeria: Morphology and chronology. *PLOS ONE* **6**, e24024 (2011).
34. G. P. Rightmire, Middle and later pleistocene hominins in Africa and Southwest Asia. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 16046–16050 (2009).
35. I. Crevecoeur, A. Brooks, I. Ribot, E. Cornelissen, P. Semal, Late stone age human remains from ishango (democratic republic of congo): New insights on late pleistocene modern human diversity in africa. *J. Hum. Evol.* **96**, 35–57 (2016).
36. The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
37. The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
38. D. P. Locke, L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darré, D. Farré, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. D. Valle, S. Purgato, M. Rocchi, M. K. Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. A. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordoñez, C. López-Otín, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. RoyChoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs, R. K. Wilson, Comparative and demographic analysis of orang-utan genomes. *Nature* **469**, 529–533 (2011).
39. Q. Fu, M. Hajdinjak, O. T. Moldovan, S. Constantin, S. Mallick, P. Skoglund, N. Patterson, N. Rohland, I. Lazaridis, B. Nickel, B. Viola, K. Prüfer, M. Meyer, J. Kelso, D. Reich, S. Pääbo, An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
40. G. McVicker, D. Gordon, C. Davis, P. Green, Widespread genomic signatures of natural selection in hominid evolution. *PLOS Genet.* **5**, e1000471 (2009).
41. K. Csilléry, O. François, M. G. B. Blum, abc: An R package for approximate Bayesian computation (ABC). *Methods Ecol. Evol.* **3**, 475–479 (2012).
42. R. R. Hudson, Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
43. G. A. Watterson, On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
44. K. Harris, R. Nielsen, Inferring demographic history from a spectrum of shared haplotype lengths. *PLOS Genet.* **9**, e1003521 (2013).
45. M. Petr, S. Pääbo, J. Kelso, B. Vernot, Limits of long-term selection against Neandertal introgression. *Proc. Natl. Acad. Sci. U.S.A.*, 1639–1644 (2019).
46. K. Prüfer, snpAD: An ancient DNA genotype caller. *Bioinformatics* **34**, 4165–4171 (2018).
47. B. C. Haller, P. W. Messer, SLiM 2: Flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).
48. J. N. Fenner, Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423 (2005).
49. H. R. Kunsch, The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17**, 1217–1241 (1989).
50. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. C. Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. García Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
52. J. D. Wall, M. A. Yang, F. Jay, S. K. Kim, E. Y. Durand, L. S. Stevison, C. Gignoux, A. Woerner, M. F. Hammer, M. Slatkin, Higher levels of neanderthal ancestry in East Asians than in Europeans. *Genetics* **194**, 199–209 (2013).
53. L. Skov, R. Hui, V. Shchur, A. Hobolth, A. Scally, M. H. Schierup, R. Durbin, Detecting archaic introgression using an unadmixed outgroup. *PLOS Genet.* **14**, e1007641 (2018).
54. M. Kimura, Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. U.S.A.* **41**, 144–150 (1955).
55. R. C. Griffiths, The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Popul. Biol.* **64**, 241–251 (2003).

Acknowledgments: We thank K. Lohmueller, N. Patterson, M. Lipson, M. Schumer, P. Moorjani, T. V. Kent, P. Skoglund, and members of the Sankararaman and Lohmueller labs for helpful comments and discussions. **Funding:** A.D. is supported by the NSF Graduate Research Fellowship DGE-1650604, and S.S. is supported, in part, by NIH grants R00GM111744 and R35GM125055, an Alfred P. Sloan Research Fellowship, and a gift from the Okawa Foundation. **Author contributions:** A.D. and S.S. designed and carried out the study. A.D. and S.S. wrote the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Sequencing data are available from the 1000 Genomes project website www.internationalgenome.org/data. Local ancestry calls are available at https://sriramlab.cass.idre.ucla.edu/public/. Additional data related to this paper may be requested from the authors.

Submitted 29 March 2019
Accepted 3 December 2019
Published 12 February 2020
10.1126/sciadv.aax5097

Citation: A. Durvasula, S. Sankararaman, Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* **6**, eaax5097 (2020).

Chapter 3: Negative selection on complex traits limits genetic risk prediction accuracy between populations

Negative selection on complex traits limits phenotype prediction accuracy between populations

Arun Durvasula[1] and Kirk E. Lohmueller[1,2,3,*]

Summary

Phenotype prediction is a key goal for medical genetics. Unfortunately, most genome-wide association studies are done in European populations, which reduces the accuracy of predictions via polygenic scores in non-European populations. Here, we use population genetic models to show that human demographic history and negative selection on complex traits can result in population-specific genetic architectures. For traits where alleles with the largest effect on the trait are under the strongest negative selection, approximately half of the heritability can be accounted for by variants in Europe that are absent from Africa, leading to poor performance in phenotype prediction across these populations. Further, under such a model, individuals in the tails of the genetic risk distribution may not be identified via polygenic scores generated in another population. We empirically test these predictions by building a model to stratify heritability between European-specific and shared variants and applied it to 37 traits and diseases in the UK Biobank. Across these phenotypes, ~30% of the heritability comes from European-specific variants. We conclude that genetic association studies need to include more diverse populations to enable the utility of phenotype prediction in all populations.

Introduction

The past decade of genome wide association studies (GWASs) has uncovered a plethora of trait-associated loci scattered across the genome.[1–4] Geneticists have devoted many resources to turning these associations into phenotype prediction models that aggregate variants across the genome into a polygenic score. Such scores can be used to guide healthcare decisions for a variety of traits and diseases,[5] and recent work has suggested these polygenic scores may be ready for clinical use.[6,7] While individuals with high polygenic risk for diseases have been found via these scores, for example in atherosclerosis[8] and breast cancer,[9] challenges remain in applying these polygenic scores uniformly across populations. Recent analyses have suggested that because many of the largest studies are concentrated on European populations, polygenic scores may be biased and less informative in non-European populations.[10–15] There are several reasons why polygenic scores may not transfer well across populations. One possibility is that alleles have different effect sizes in different populations, owing to differences in interactions with the environment.[16] Another possibility is that differences in linkage disequilibrium (LD) between variants across populations means that causal variants may be tagged differently in non-European populations, leading to differences in effect sizes.[11,17] Finally, the original polygenic score performance in Europeans may be inflated because of population stratification.[18,19]

Here, we propose that an additional reason for the lack of transferability of polygenic scores is that each population has its own genetic architecture, owing to the evolutionary processes that give rise to traits. Under this reasoning, a population's demographic history influences the number of causal variants and their frequencies, resulting in some phenotypic variance coming from causal variants that are population specific. For example, work on the genetic architecture of skin color in African populations has uncovered distinct loci affecting the trait in each population, suggesting that populations with independent demographic histories can end up with different genetic architectures and causal variants for the same traits.[20] Indeed, modeling work suggests that genetic architecture is an outcome of the evolutionary process rather than a trait-specific property.[21]

Recent exponential growth in human populations has created an excess of new variants that tend to be low frequency and population specific (private variation[22–24]). Population genetic models of genetic architecture that include negative selection suggest that, in aggregate, low-frequency variants could contribute substantially to traits.[25–27] Application of these models to large-scale genetic datasets has discovered that many traits are under apparent negative selection, ranging from anthropometric traits to molecular phenotypes.[28–33] Depending on the interplay between allele frequency and effect size, these variants could make up a large portion of the heritability for many traits, as demonstrated by a recent GWAS on height and BMI using whole-genome sequencing data.[34,35] Because narrow-sense heritability is the proportion of variance explained by additive genetic factors, it is directly related to the accuracy of phenotypic prediction as the variance explained by the polygenic score.[36] If these private variants contribute substantially to heritability, it

[1]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90095, USA; [2]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA; [3]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, CA 90095, USA
*Correspondence: klohmueller@ucla.edu
https://doi.org/10.1016/j.ajhg.2021.02.013.

follows that the variants will not be useful for phenotype prediction between populations because they are not present in other populations. The proportion of narrow-sense heritability that private variants explain places an upper bound on the accuracy of polygenic scores between populations.

In this study, we use simulations under demographic scenarios of recent explosive population growth with varying amounts of negative selection as well as analyses of empirical data to test the role of private variants in complex traits.

Material and methods

Population genetic modeling and simulations

We performed forward simulations by using SLiM v.3.[37] We simulated a demographic history for a European and an African population according to the demographic model fit by Gravel et al.[38] (including migration). The African population size expanded to 14,474 individuals and the European population began at a size of 1,032 individuals after splitting from Africa and grew exponentially at a rate of 0.38% per generation for 920 generations. We simulated a mutational target size of 5 Mb with a mutation rate of 1.2×10^{-8} per base pair (bp) and a recombination rate of 1×10^{-8} per bp. To simulate selection across the entire region, we drew selection coefficients for new mutations from a gamma distribution with parameters fit by Kim et al.[39] (mean $= -0.01026$, $\alpha = 0.186$). We sampled 10,000 haploid genomes from each population. To simulate a quantitative trait, we followed the model described by Eyre-Walker[25] and the framework set by Lohmueller[21] where a SNP's effect on a trait, β, is given by

$$\beta = \delta s^{\tau}(1+\varepsilon)C,$$

where $\delta \in \{-1,\ 1\}$ with equal probability, $\varepsilon \sim N(0,\ 0.5)$, and s is the selection coefficient of a variant segregating in the population at the end of the simulation. C is a scaling factor for effect sizes and controls the heritability for a given mutational target size. In these simulations, C was set to obtain a heritability of ~0.4 (see Table S1). Finally, τ reflects the relationship between a SNP's effect on fitness and the trait. $\tau = 0$ indicates no relationship between fitness and the trait, while $\tau > 0$ indicates that mutations that are more evolutionarily deleterious are those that have larger effects on the trait. In this model, when $\tau > 0$, the trait itself may be under direct selection or it may be correlated with a trait under selection. We call variants private and shared on the basis of their allele frequency in a sample of 10,000 chromosomes from both populations (see below).

To compare our simulation results to the empirical data from the Exome Aggregation Consortium (ExAC), which includes African American individuals, we computed the expected allele frequency for SNP i in simulated admixed African American individuals ($p_{AA,\,i}$) as

$$p_{AA,i} = \alpha_i p_{EUR,i} + (1-\alpha_i)p_{AFR,i}$$

where $p_{EUR,\,i}$ and $p_{AFR,\,i}$ denote the allele frequencies in Europe and Africa, respectively. For each SNP, we drew an admixture proportion $\alpha_i \sim Beta(2,\ 8)$ in order to incorporate variance in the admixture proportion along the genome. The parameters of the beta distribution were chosen to match the observed variation in admixture proportion in African American individuals[40] and result in a mean proportion of African ancestry of 80%.

Defining the proportion of heritability from private variants: h^2_{private}

We begin by describing a model in which an individual, i, in a population, ϕ, has a phenotype, y_i, that is a linear combination of genotypes ($\boldsymbol{x_i}$, $x_{ij} \in \{x_{i1}, \ldots, x_{iM}\}$), effect sizes ($\boldsymbol{\beta}$, $\beta_j \in \{\beta_1, \ldots, \beta_M\}$), and a normally distributed term describing the effect of the environment, $e_i \sim N(0,\ V_E)$:

$$y_i = \boldsymbol{x_i^T}\boldsymbol{\beta} + e_i.$$

The narrow-sense heritability, h^2, of the phenotype, $\boldsymbol{y}$, in the population is given by

$$h^2 = \frac{V_A}{Var(\boldsymbol{y})}$$

where the variance of the phenotype can be decomposed into additive, dominance, interacting, and environmental terms: $Var(\boldsymbol{y}) = V_A + V_D + V_I + V_E$. The additive genetic variance is $V_A = 2\sum_{j=1}^{M} p_j(1-p_j)\beta_j^2$ when there are M variants, where p_j is the allele frequency for variant j and β_j is the effect size of variant j.

We wish to examine the proportion of heritability that comes from a particular class of variants. Consider a sister population, ψ, that diverged from the population described above (ϕ). Variants in population ϕ can be partitioned into those that appear only in ϕ (private variants) or those that appear in both populations (shared variants). The total number of variants is the sum of the number of shared and number of private variants, $M = M_p + M_s$. We wish to partition the heritability into these two classes, h_p^2 and h_s^2, which make up the total heritability: $h^2 = h_p^2 + h_s^2$. Define h^2_{private} to be the proportion of the heritability accounted for by the private variants.

The quantity of interest, then, is

$$h^2_{\text{private}} = \frac{h_p^2}{h^2} = \frac{V_{A,p}}{V_A}.$$

The additive genetic variance from private variants is $V_{A,p} = 2\sum_{j=1}^{M} p_j(1-p_j)\beta_j^2 z_j$, where z_j is an indicator function that is 1 when the variant j is private (with probability $P(\omega_j)$) to the population and 0 otherwise. We describe how z_j is estimated below when analyzing empirical data (see model to identify private variants).

Polygenic score calculation

We compute three sets of polygenic scores on the simulated individuals: (1) using all variants, (2) using variants private to the simulated population of interest, and (3) using variants shared between the simulated European and African populations. For each haploid genome, we sum the effect sizes, β, for each class of variants, resulting in three scores for each genome. We standardize the scores by subtracting the mean of the true polygenic score (class 1) and dividing by the standard deviation of the true polygenic score (class 1). We compute the Pearson correlation between classes 1 and 2 as well as classes 1 and 3 and report the r^2 value as a percentage.

Model to identify private variants

When analyzing the empirical UK Biobank data, it is challenging to assess whether a particular variant is private or shared. If a variant is seen only in one population, it is possible that it is truly private to that population, or instead, it is shared but at too low a frequency to have been discovered with the number of individuals

samples from the other population. To address this issue, we built a probabilistic model to evaluate the probability that a variant is private to a population given the number of copies of the allele in that population (that is, the allele frequency).

We begin with the intuition that rare alleles tend to be private and common alleles tend to be shared between populations, even in the presence of migration. Migration can be thought of as sampling alleles from one population and placing them in the other population. Under this model, rare alleles will tend to stay within a population and not transfer between populations. This suggests that allele frequency is informative in determining whether an allele is private or not.

Wakeley and Hey[41] use coalescent theory to determine the frequency spectrum of private variants. An application of Bayes' rule allows us to calculate the following probability:

$$P(\omega \Big| i) = \frac{P(i|\omega)P(\omega)}{P(i)},$$

where $i \in \{1, \ldots, n\}$ is the number of copies of the allele in the sample (n) and $\omega \in \{0, 1\}$ is 1 if the allele is private and 0 if not. $P(i|\omega = 1)$ is the site frequency spectrum of private variants, and $P(i)$ is given by the full site frequency spectrum. For example, in a constant-sized equilibrium population, $P(i) = (\theta/i)/(\sum_i^n \theta/i)$. $P(\omega)$ is the probability of a variant's being private to a population.

Wakeley and Hey[41] provide expressions to obtain these quantities in a constant-sized equilibrium population without natural selection. However, here we are concerned with populations that are not in equilibrium and with variants under negative selection, so we obtain these probabilities via simulation under a particular demographic model and distribution of fitness effects.

In the results presented here, we use the demographic model from Gravel et al.[38] that relates European and African populations. We use a distribution of fitness effects from Kim et al.,[39] assuming that mutations are additive (that is, $h = 0.5$) and that selection coefficients, s, are drawn from a gamma distribution with mean $= -0.01026$ and shape $= 0.186$. Using these parameters, we simulate data for 10,000 European chromosomes by using SLiM[37] and compute (1) the proportional site frequency spectrum for private variants ($P(i|\omega)$), (2) the proportional site frequency spectrum for all variants ($P(i)$), and (3) the proportion of private variants ($P(\omega)$). We defined private variants in the simulation as those that appear in the simulated European population but not the simulated African population.

Next, we store these quantities in a lookup table and use them to compute the probability that a variant is private given the number of copies of the allele in the empirical data. In the UK Biobank dataset, alleles are present at frequency 1×10^{-6} and higher. However, in simulations, the lowest allele frequency is 1×10^{-4}. For alleles below this frequency, we set the probability equal to the probability for alleles at a frequency of 1 in 10,000.

Testing our probabilistic model to infer private variants

We evaluated the ability of our model to distinguish between private and shared variants by simulating new data and performing binary classification, calling a variant private if the $P(\omega|i)$ exceeded some threshold, t. We varied this threshold and computed the number of true positive (private variants that are truly private), false positives (private variants that are truly shared), false negatives (shared variants that are truly private), and true negatives (shared variants that are truly shared). We summarized this by using receiver operator characteristic and precision recall curves (Figure S1; Tables S2 and S3).

We also validated our model by using data from ExAC.[42] For each variant in ExAC, we used our model to compute the probability that the variant is private to the non-Finnish European population on the basis of the allele frequency in that population. Then, we checked whether variants were observed in a sample of 10,406 African and African American samples.

Partitioning heritability

We applied our Bayesian model to predict which variants are private to GWAS summary statistics from 37 traits in the UK Biobank released by the Neale lab (see web resources). We computed the additive genetic variance for variants with a high posterior probability of being private to the British cohort and divided that by the total amount of additive genetic variance explained by SNPs to obtain our estimate of h^2_{private} (Note S1). We also performed the inference by using a randomized algorithm to correct for the effects of LD and misestimated effect sizes as well as population stratification (Notes S2, S3, S4, and S5; Figures S3, S4, S5, S6, S7, S8, and S9). Finally, we also independently replicated the results on BMI by using data from the GIANT consortium[43] (Note S1). Importantly, this partitioning of the heritability into shared and private components does not make use of the τ-model[25] that relates a mutation's effect on fitness to its effect on the trait.

Results

The distribution of European-specific variants in data and models

We begin by precisely defining private variants in the datasets and models that we consider. Studies of genomic variation point to the out-of-Africa bottleneck and subsequent explosive growth in population size as a key driver of the distribution of genomic variation. We focus on a simplified model of this history (Figure 1A; Gravel et al.[38]). We define private variants as those that are found in Europe but are absent from Africa and shared variants as those that are found in both populations. Note that by our definition, private variants may be shared between other out-of-Africa populations (e.g., between Europe and East Asia) because of shared recent history.

One potential concern with this definition of whether a variant is private to Europe is that it may depend on the sample size of the African population used in the comparison. We examined this possibility by computing the probability of not observing an allele present in a sample of African individuals across a range of minor allele frequencies (MAFs) with a sample size of 10,000 chromosomes. This sample size is approximately similar to the sample size of the ExAC dataset (Lek et al.[42]). We find that variants with a frequency as low as 10^{-3} in the African population have a nearly 100% probability of being sampled in ExAC (Figure S2). Thus, we would correctly classify variants segregating at low frequency in Africa as being shared.

Next, we examined the number of private variants in European populations compared to African populations in

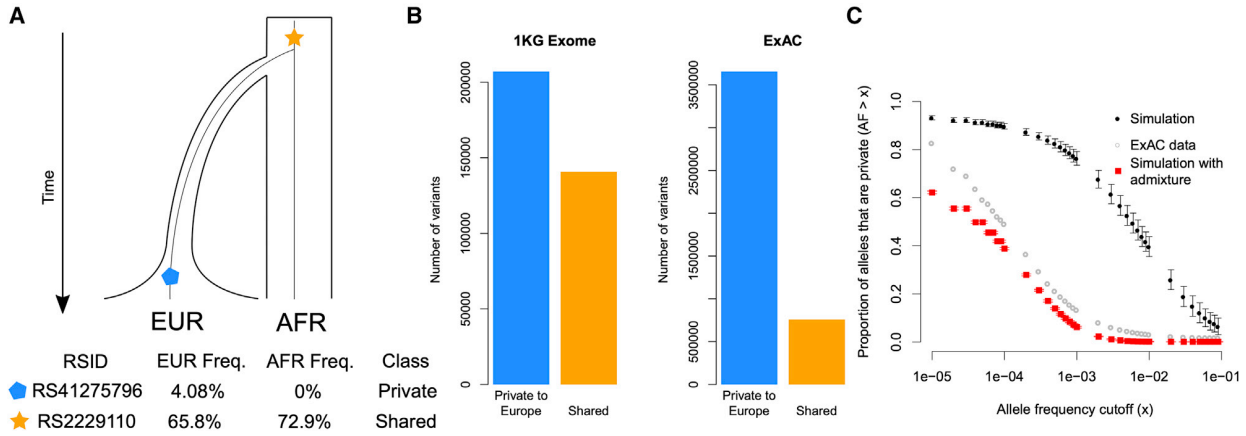

Figure 1. Human population history generates population-specific variants
(A) Model for variants that are shared (common to Europe [EUR] and Africa [AFR]) and private (occurring only in EUR and absent from AFR). Bottom, examples of private and shared variants from ExAC.[42]
(B) The number of non-synonymous variants that are private to European populations and absent from African populations (blue bars) and the number of non-synonymous variants that are shared between the two populations in the 1KG exome dataset and the ExAC dataset (orange bars).
(C) The proportion of non-synonymous alleles above a given frequency that are private to Europe and absent from Africa in the ExAC dataset and in simulations based on human history. Note that because the ExAC dataset contains admixed African American individuals, the proportion of private variants is reduced compared with the original simulation (black dots). Modeling this admixture (red dots) shows a better fit to this dataset. Error bars denote standard deviation across simulation replicates.

two datasets: the 1000 Genomes (1KG) data and the ExAC data. In order to meaningfully compare the two datasets, we focused on variants contained in the exome. For both datasets, there are many more private variants in the European population compared to shared variants (Figure 1B). This is expected under models of human history where many shared alleles were lost during the out-of-Africa bottleneck and new mutations accumulated independently in the out-of-Africa population. Because of the small population size, some of these mutations could drift to a higher frequency than they would have in a larger population.

We next conducted simulations under this model of human evolution, where an ancestral population splits into a group that underwent a genetic bottleneck out of Africa (representing a European population) and a group that stayed within Africa without a bottleneck (representing an African population; Figure 1A [38]), coupled with varying levels of negative selection on traits (including no negative selection). We include negative selection by modifying the relationship between a mutation's effect on the trait and its effect on reproductive fitness by using the model put forth by Eyre-Walker in 2010[25] (see material and methods). This model includes a parameter, τ, which ties the selection coefficient of a mutation to its effect on a trait.[25] Larger values of τ imply that more evolutionarily deleterious mutations have larger effects on the trait. Importantly, our model includes exponential growth in the out-of-Africa population, which creates an excess of private variants, as well as low levels of migration between the European and African populations, which can turn some private variants into shared variants. We compared our simulations to data from ExAC and found that our simulations predicted more higher-frequency private alleles than are observed in the data (Figure 1C). However, the ExAC data contains admixed African American individuals. Admixture can introduce variants that are private to Europe into the sample labeled "African." We simulated this admixture process (see material and methods) and found that the resulting simulation matches the data closely, suggesting that our model is a reasonable approximation of human demography and selection (Figure 1C).

Population genetic models predict population-specific variants account for heritability and impact polygenic scores

We reasoned that since there are many private causal variants in our simulations, they may account for a substantial proportion of the heritability in aggregate. We examined the contribution of private variants to heritability and found that when traits are not tied to fitness ($\tau = 0$), private variants account for ~30% of the heritability (Figure 2A). However, when the coupling between trait effects and fitness effects is moderate ($\tau = 0.25$) or strong ($\tau = 0.5$), private variants account for over half of the heritability, and there is a maximum of ~79% under strong coupling (Figures 2B and 2C). These results suggest that many causal variants, which jointly explain much of the heritability, tend to be population specific. This effect is a consequence of how the trait relates to fitness as well as the demographic history of the population.

The fact that many of the variants that affect the trait are not shared across populations may limit the applicability of polygenic scores derived from European populations to other populations. This effect would be distinct from

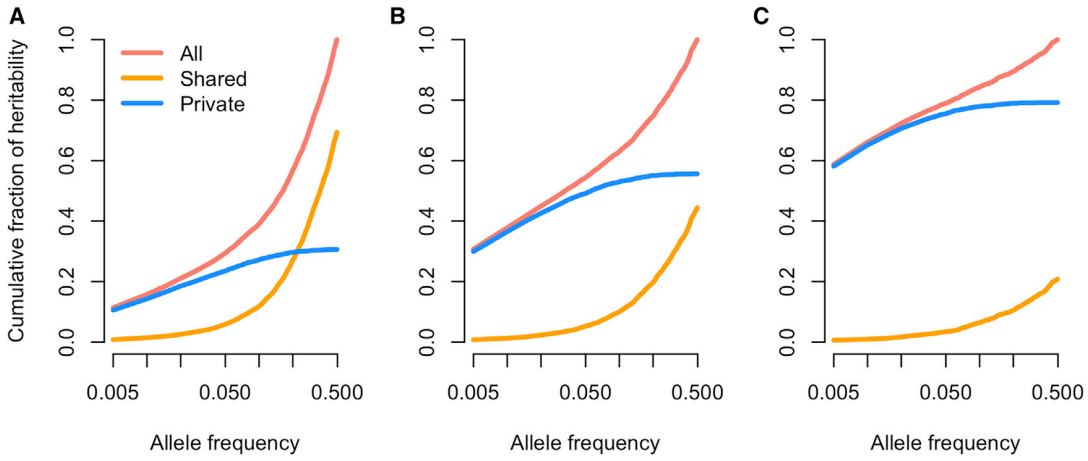

Figure 2. The effect of natural selection on the relationship between heritability and allele frequency
(A–C) Cumulative fraction of heritability explained by private and shared variants under (A) no relation between a mutation's effect on fitness and the trait ($\tau = 0$), (B) moderate coupling between a mutation's effect on fitness and the trait ($\tau = 0.25$), and (C) strong coupling between a mutation's effect on fitness and the trait($\tau = 0.5$). Note that the x axis is on a log scale. As τ increases, a greater fraction of heritability comes from variation that is found only within Europe.

imperfect tagging of causal variants due to differences in LD patterns between populations. To test for this effect in simulated data, we calculated true polygenic scores for individuals in the simulated European and African populations and asked how well polygenic scores derived from only private variants and only shared variants correlated with the true polygenic scores. Polygenic scores derived from only shared variants represent the case where a polygenic score can be transferred from Europe to another population. If shared variant effect sizes correlate well between populations, despite not contributing to a majority of additive genetic variance, polygenic scores may still be accurate across populations. We note that these simulations include identification of the true causal SNPs and, as such, are much higher than polygenic score accuracies reported elsewhere.[13] These simulations represent the best-case scenario for polygenic scores. We found that when traits are independent of fitness, the shared polygenic score has a 91% correlation in Europe and 96% correlation in Africa with the true polygenic score, suggesting that polygenic scores can be applied between populations (Figure 3A). However, we found that when trait effects are tied to fitness effects, the correlation between shared polygenic scores and the true polygenic scores decreases (Figures 3B and 3C) and the correlation between private polygenic scores and true polygenic scores increases (Figures 3D, 3E, and 3F). Note that in the analysis with private polygenic scores, each population uses variants private to that population but not from the other population. That is, the African private polygenic score uses variants private to Africa. This suggests that the reduction in accuracy does not depend on the population's specific demography, as the same pattern is present in European and African populations. For traits with strong coupling between trait effects and fitness effects ($\tau = 0.5$), the correlation between the true polygenic scores and the polygenic scores derived from shared variants drops to 62% in Europe and 57% in Africa (Table S4). These findings suggest that polygenic scores based solely on shared variants may be substantially less accurate than polygenic scores using all variants and may not transfer between populations well when the variants with the greatest effects on the trait are those under the most negative selection.

While shared variants do not capture the full distribution of polygenic scores, we asked whether individuals in the tail of the true polygenic score distribution remained in the tail when examining shared variants only. When there is no coupling between fitness and trait effects ($\tau = 0$), shared variants capture 35% of the tail correctly in Europe and 28% of the tail correctly in Africa (Table 1). However, when there is moderate coupling ($\tau = 0.25$), this number drops to 11% in Europe and 7% in Africa. When there is strong coupling, the polygenic score based on shared variants identifies none of the individuals in the tails of the distribution. If the trait under consideration is a disease, this analysis suggests that a polygenic score based on shared variation cannot identify individuals at the highest risk for that disease. In contrast, when considering only private variants, the polygenic score correctly identifies 44%–46% of individuals who are at the extremes of the distribution. These results suggest that when using scores derived from European populations, individuals who are truly in the tails of the polygenic score distribution will not be identified via shared variants alone, corresponding to a high false-negative error rate. In addition, the low recall for both of these polygenic scores suggests many individuals that are in the tails of the distribution will be missed.

While our simulations suggest private variants may be an important component of the heritability and may limit

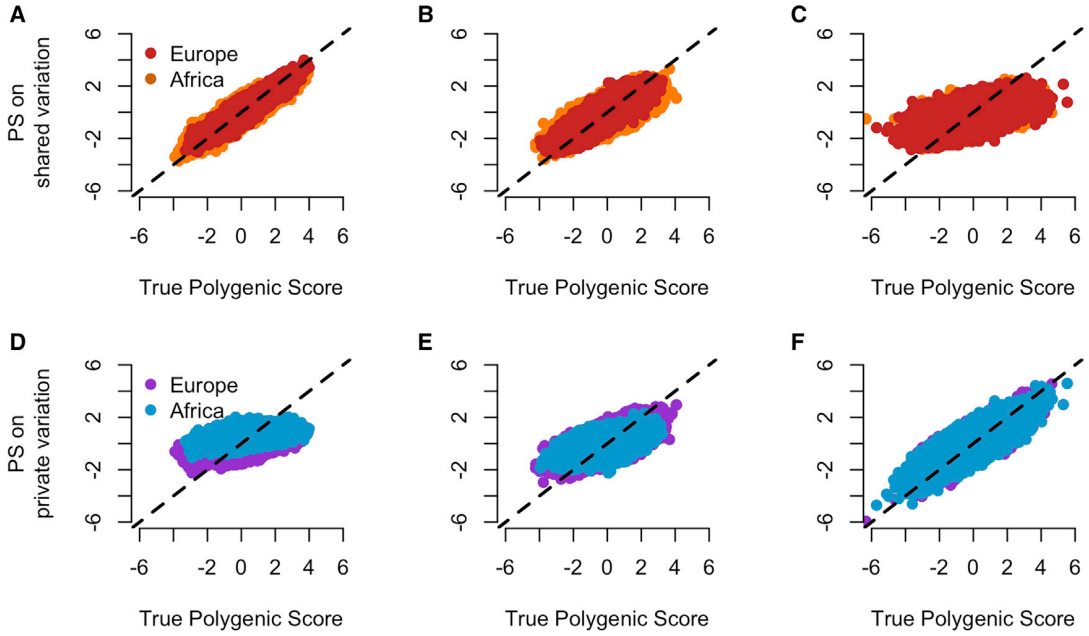

Figure 3. The relationship between polygenic scores and natural selection
(A–F) Polygenic score accuracy for shared variants only (top row) and private variants only (bottom row) in Europe and Africa on simulated data with different degrees of negative selection. In the bottom row, each score uses private variants from within the population being considered (e.g., for Africa, we use variants private to Africa) but not from the other population. The black line shows the 1:1 line. (A and D) No relationship between a mutation's effect on fitness and its effect on the trait ($\tau = 0$). (B and E) Moderate coupling between fitness and trait effects ($\tau = 0.25$). (C and F) Strong coupling ($\tau = 0.5$). As the strength of coupling increases, polygenic scores computed from shared variation become less correlated with the true polygenic score. However, at the same time, polygenic scores computed from private variation become more correlated with the true polygenic scores.

phenotype prediction across populations, their precise role depends on the extent of negative selection acting on traits (either directly or through pleiotropy), which remains an open question.[28–30,32,33] Thus, we next tested how much of the heritability private variants account for in real GWAS data in European populations, where GWAS data is abundant.

A model for private variation

We built a Bayesian model to classify variants segregating in the UK Biobank as private or shared by using the allele frequency conditional on a demographic model and distribution of fitness effects inferred for a European population (see material and methods). To validate our model, we simulated a new dataset under the same European demographic model and recorded whether each allele was observed in both populations. Then, we calculated the probability of each allele's being private to the European population. We classified variants as private if the probability $P(\omega|i) \geq t$, $\omega \in \{0, 1\}$ is 1 if the allele is private and 0 if not, $i \in \{1, \ldots, n\}$ is the number of copies of the allele in the sample, and t is some probability cutoff. For each cutoff, we calculated (1) the number of variants that we predict are private and are truly private (true positives), (2) the number of variants that we predict are private and are truly not private (false positives), (3) the number of variants that we predict are not private and are truly private (false negatives), and (4) the number of variants that we predict are not private and are truly not private (true negatives).

We summarize these numbers by using two curves: a precision-recall curve (Figure S1A) and a receiver operator characteristic (ROC) curve (Figure S1B). We find that at a precision of 94%, we have a recall of 99% and that the area under the ROC curve is 0.80, suggesting that our model is able to distinguish between private and shared variants on the basis of allele frequency alone (Table S3). We also tested the model on a simulated dataset including five times more individuals than the 10,000 individuals used in the initial simulation. Importantly, for this comparison, we used the same lookup table, based on 10,000 individuals, as before. This allows us to test how sample size affects our inferences. We find that the precision-recall curve is largely the same, but there is a decrease in the ROC curve (AUROC = 0.70).

In addition, examining $P(\omega|i)$ versus the allele frequency in the simulated independent dataset (Figure S1C), we find that alleles higher than ~10% frequency have a negligible probability of being private. This is consistent with the intuition that common alleles are unlikely to be private.

We also examined several posterior probability thresholds in detail ($t \in \{0.1, 0.23, 0.4\}$; Table S3). Across these thresholds, we find that the false discovery rate (FDR) from simulations is ~5%, suggesting that the model is relatively robust to the threshold used.

Table 1. The effect of natural selection on identifying high-risk individuals

τ	Shared (Europe)	Private (Europe)	Shared (Africa)	Private (Africa)
0	35%	22%	28%	11%
0.25	11%	20%	7%	18%
0.5	0%	46%	0%	44%

Percentage of individuals in the extreme 5% tail of the true polygenic score distribution that are recovered when using only private variants and shared variants in simulated European and African populations. Overall, the percentage of individuals correctly classified is low, suggesting that there will be many false negatives when using polygenic scores to identify individuals in the tails of the risk distribution. Further, as the degree of coupling between fitness effects and trait effects increases, shared variants correctly classify fewer individuals, while private variants classify more individuals correctly.

Next, we empirically validated the performance of our model to infer whether variants are private. Using data from ExAC,[42] we use our framework described above to calculate the probability that each variant is private by using the allele frequency in the non-Finnish Europeans (NFE). In Figure S1D, we plot this probability for a random subset of 10,000 variants. We see that variants above 10% frequency have a very low probability of being private and that variants below that frequency increase in probability of being private as their frequency decreases.

In addition, we classified variants in ExAC as private to "EUR" by using the simulation-based FDR of 5% ($P(\omega|i) \geq 0.23$) and checked whether those variants were present in the "AFR" subset of samples (Table S2). We see that 83% of the variants we call private are not observed in "AFR" in a sample of 10,406 chromosomes. This suggests that our empirical-based FDR is 17% and is higher than the simulation-based FDR. However, the "AFR" sample in ExAC is a mixture of African American and African samples. Importantly, African American samples are admixed between European and African populations.[42] This has the effect of introducing European variants into the "AFR" samples, making variants we expect to be private to "EUR" appear shared. Therefore, this estimate of the accuracy is most likely an underestimate.

Nonetheless, these simulations and empirical evaluations suggest that our model is able to distinguish between private and shared variants on the basis of allele frequency alone. Additionally, out of an abundance of caution, we utilize two different empirically based FDRs of 17% in downstream inferences as described below. Importantly, our determination of whether a variant is private or shared is expected to hold regardless of the sample size taken from either population (Table S2).

Inference of heritability accounted for by private variants: h^2_{private}

We used summary statistics for 37 different traits and diseases from the UK Biobank relating to anthropometric and blood-related traits as well as cancer-related and non-cancer related diseases (see web resources) to infer the proportion of the SNP-based heritability attributable to private variants, h^2_{private}. Using these data and our probabilistic method to determine whether a variant is private or not, we find that the average $h^2_{\text{private}} = 31\%$, and there is substantial variation across traits (standard deviation: 11%; Figure 4). Examining categories of diseases, we find that cancer-related diseases have $h^2_{\text{private}} = 12\%$, while non-cancer-related diseases have $h^2_{\text{private}} = 32\%$. Similarly, private variants account for ~30% of the heritability in blood-related and anthropometric traits. We observe substantial variability across different traits within a category. Two blood pressure-related traits have h^2_{private} of nearly 50%, while other blood-related traits have a lower proportion.

The effect of falsely identified private variants on our inference of h^2_{private}

To ensure that our results from the UK Biobank data described above were not driven by shared SNPs that we mistakenly classified as private, we adjusted for an empirically based FDR. At the threshold used for classifying variants as being private ($P(\omega|i) = 0.23$), validation in the empirical data suggest the FDR is ~17% (see above). In other words, approximately 17% of SNPs that we identify as private may actually be shared. Thus, we adjusted our estimates of h^2_{private} by randomly reclassifying 17% of the private SNPs as shared and re-computed h^2_{private} ("17% FDR correction" in Figure 4). Despite the extremely conservative nature of this correction (because the empirical FDR is based on an admixed sample), we find that a sizeable proportion of the heritability (about 22%) still comes from private variants (Figure 4).

In addition to this conservative correction, we also performed an even more stringent correction where we sorted the SNPs we call private by their heritability and removed 17% of the SNPs that explain the most heritability. As expected, the amount of heritability from private variants goes down, but for most traits, the heritability explained by private variants is still greater than 10% ("Max FDR correction" in Figure 4). This suggests that our central claim, that private variants contribute to heritability, remains true even if our classification method is imperfect.

The effect of population stratification

Recent studies have highlighted the effects of stratification on polygenic scores.[18,19] We considered whether stratification could have an effect on our analyses. To test this, we repeated our analyses by using only those SNPs showing stronger associations with the trait. Specifically, we employed p value cutoffs, using only SNPs with a p value lower than the cutoff (Figure S3). Broadly, for quantitative traits, we observe that as the p value threshold becomes stricter, the proportion of the heritability attributable to private variants decreases. This is due to the power to detect associations for private variants. The power to detect an association will be lower for private variants than shared variants because private variants tend to have lower allele frequencies. Therefore, as the p value cutoff

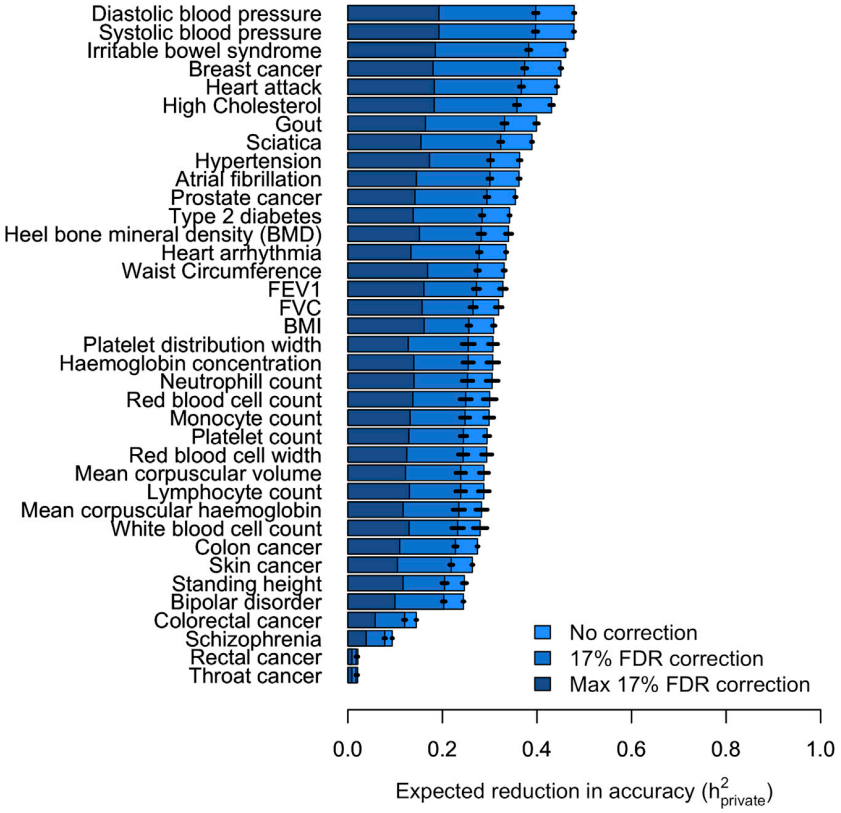

Figure 4. Estimates of the amount of heritability from private variants
The expected reduction in accuracy when transferring a polygenic score from Europe to Africa (expressed as the percentage of heritability explained by private variants) across 37 traits and diseases in the UK Biobank. We only include SNPs with an MAF $> 10^{-3}$. The mean reduction is 26.7% (SD across traits is 14.7%). "17% FDR correction" refers to randomly setting 17% of the SNPs that we call private to shared. "Max 17% FDR correction" refers to setting the 17% of the SNPs that explain the most heritability from private to shared. Lines indicate standard errors obtained via a 1 Mb block jackknife.

decreases, we expect a lower proportion of heritability to come from private variants. We found that the total variance explained by SNPs for dichotomous traits was much lower than for quantitative traits. This effect produced a statistical artifact where the heritability from private variants tended to be very high for dichotomous traits (Figure S4).

In addition to this analysis, we were also concerned with the effect of differential population structure from rare variants.[35,44] Therefore, we checked the robustness of our results to allele frequency filters. We computed h^2_{private} for atrial fibrillation, BMI, standing height, diastolic blood pressure, and type 2 diabetes with MAF cutoffs from 10^{-5}, 10^{-4}, 10^{-3}, and 10^{-2} (Figure 5). We find that although h^2_{private} decreases, it still remains substantial up to a cutoff of 10^{-2}. Although this analysis removes both real and spurious signals, it suggests that private variants do indeed explain a non-negligible proportion of heritability.

Across different p value thresholds, a non-negligible proportion of heritability comes from private variants. However, this analysis does not alleviate all concerns about population stratification, as at a large enough sample size, an association due to stratification can be arbitrarily strong. Similarly, stratification could still occur when using variants at different MAF cutoffs. While these analyses provide evidence that our results are not primarily driven by stratification, they cannot completely rule it out. Further advances in controlling for stratification of rare variants will be crucial to understand the full contribution of private variants to heritability.

The effect of unmodeled LD on our inferences

Our inferences of h^2_{private} make the assumption that the estimated effect sizes for the GWAS SNPs were the true effect sizes of the causal variants. Further, we assumed that the variants were all independent of each other. In truth, these assumptions are violated for a variety of reasons. First, because of LD, SNPs may be correlated with one another. Second, some of the non-zero effect sizes of GWAS SNPs may be due to the fact that the GWAS SNP is tagging (in LD with) an untyped causal variant and is itself not causal. Third, even if the GWAS variants analyzed in our study are the true causal variants, their effect sizes may be misestimated by the effects at nearby SNPs in LD with them. Thus, given these challenges, we carefully considered the effect that unmodeled LD may have on our inferences (see Note S2).

First, we developed an estimator of the SNP-based heritability that downsamples the number of SNPs to be independent of each other. We checked the robustness of our results to this effect by randomly selecting a single SNP in a window and computing the proportion of heritability from private variants by using these randomly selected SNPs. We select only one SNP per window to avoid counting SNPs located nearby each other that are in LD with each other. We randomly selected SNPs to avoid biases due to the fact that more sophisticated methods for fine-mapping SNPs by using LD patterns may have different performance for different allele frequencies. We find similar results via our LD-pruned estimator compared with the full data (Note S3; Figures S5, S6, and S7). We also ensure that our estimates are sensible by estimating the proportion of additive genetic variance from variants we infer to be shared (Figure S8). If the inference procedure

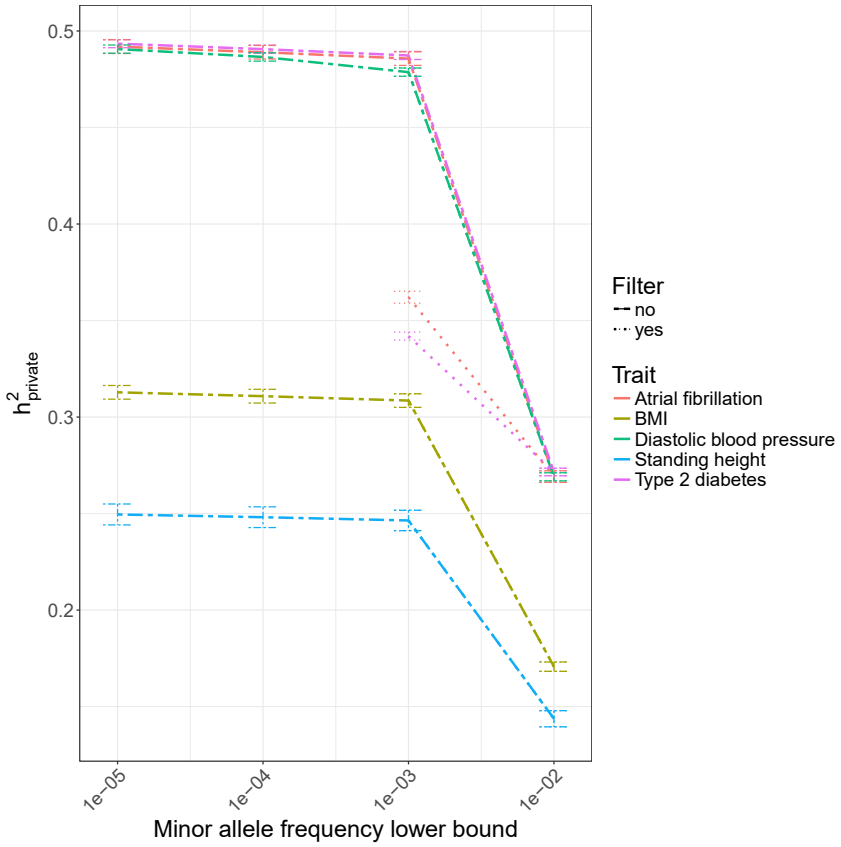

Figure 5. The effect of MAF cutoffs on heritability from private variants
"Filter" refers to a quality and allele frequency filter that removes variants with a frequency below 10^{-3}, a Hardy Weinberg equilibrium p value $< 10^{-10}$, or SNP information score < 0.8. We show results for five commonly studied traits. For BMI, diastolic blood pressure, and standing height, the "filter yes" lines are behind the "filter no" lines. This suggests that the variant filter has no effect on the estimated heritability. Lines indicate standard errors obtained via a 1 Mb block jackknife.

works correctly, this number should be $1 - \widehat{h^2_{private}}$. In Figure S8, we see that this is indeed the case.

Second, based on first principles, our estimates of h^2_{private} most likely underestimate the true proportion due to LD between tagging and causal variants (see Note S2). Because shared variants tend to be more common, they will tend to be in LD with more (and therefore tag more) variants. Because of this effect, shared variants could have inflated marginal effect sizes compared to private variants. This would lead to overestimating the heritability from shared variants compared to private variants, making our inferences conservative. We tested for this effect in real data by testing the correlation between marginal effect sizes and recombination rate for variants we predict to be private and variants we predict to be shared (Note S4). We found that variants we predict to be private have lower correlation than variants we predict to be shared, consistent with the idea that shared variants tag more variants than private variants (we also note that this reasoning is the motivation for LD score regression[45]). In addition, coalescent simulations show that our estimator of h^2_{private} is indeed slightly downwardly biased (Note S5; Figure S9).

Discussion

In this work, we have shown that recent population growth and negative selection create population-specific genetic architectures for phenotypes, which has the direct effect of reducing the accuracy of polygenic scores when applied between populations. The reduction in accuracy will depend on how differentiated populations are and accuracy decreases as populations become more differentiated. Another case to consider is admixed populations where some causal variants could be introduced and thus become shared variants. In these cases, we expect the utility of polygenic scores to be higher, but this will depend on how recent the admixture was and how many causal variants are transferred between populations, which can vary between individuals.

In our simulation results, we found that when there was no coupling between trait effects and fitness, approximately 30% of the heritability comes from private variants and that this proportion increases as the coupling increases. Although we expect this general pattern to hold, the specific values will depend on the distribution of fitness effect for causal alleles, the mutation target size, and the demographic history of the populations under study. We have used a distribution of fitness effects that was fit to non-synonymous variants[39] and note that the estimates of selection on causal alleles could be revised in future studies. In addition, our model with admixture fits the observed data better than a model without admixture (Figure 1C), but we may still be underestimating the number of private alleles, which would cause our estimates to be a lower bound. Nevertheless, our results suggest that a non-negligible proportion of the heritability comes from private alleles.

We find that phenotypes with a majority of heritability explained by private variants are not likely to be predicted well in non-European populations, even if effect sizes are accurately inferred. Our analysis of the UK Biobank data suggests that most traits examined here have at least 20% of the heritability explained by private variants ($h^2_{\text{private}} > 20\%$), indicating that cross-population polygenic scores are limited in accuracy and many

population-specific causal variants remain to be discovered. We note that our inferences on the empirical data do not make use of the Eyre-Walker τ model.[25] As such, our inferences from empirical data do not make any assumptions about the relationship between a mutation's effect on fitness and the trait.

At first glance, our result that many traits have a population-specific genetic component seems at odds with recently reported results suggesting that the genetic correlation between traits in European and East Asian populations is very high.[46,47] However, we note that both of these studies examined common variants (MAF $>$ 5%), which are more likely to be shared. Our study explicitly considers a larger range of allele frequencies, which is more likely to include population-specific variants.

Our results have several implications for users of polygenic scores. First, we show that the transferability of polygenic scores depends on the particular trait being examined. For traits with larger values of h^2_{private} (such as diastolic and systolic blood pressure), the transferability would be lower because we find these traits derive more of their heritability from variants that are more likely to be private (h^2_{private}: $\approx$48% for both). In contrast, we find that traits with lower values of h^2_{private}, such as white blood cell count, can be more easily transferred because the heritability is spread more evenly across the spectrum of MAFs (h^2_{private}: 28%). Although we include standard errors estimated via a jackknife, this procedure may not account for all the uncertainty. Therefore, specific differences across traits should be interpreted cautiously. In addition, our inferences, like those in Lam et al.[46] and Liu et al.,[47] focus on the SNP heritability rather than the total heritability of particular traits.

Several recent reviews and commentaries have pointed out the potential for misuse of polygenic scores to justify racism and white supremacy, especially when comparing polygenic scores across populations.[16,48–51] Importantly, although our study indicates that population-specific variants play a role in complex traits, it is incorrect to conclude that population-specific variants lead to differences in traits between populations. Previous simulation studies have suggested that the interplay between demography and negative selection will not lead to large differences in trait heritability between populations.[21,27] Instead, these evolutionary forces can change how the heritability is accounted for. For example, as we show here, population growth and negative selection can lead to heritability's being accounted for by lower-frequency variants that are population specific instead of common variants shared across populations. Further, non-genetic factors most likely play an important role in differences in phenotype between populations.[52]

We also highlight a crucial issue in identifying individuals in the tails of the phenotype distribution. If polygenic scores are to be used more commonly in the clinic, false-negative rates must be more closely examined across populations and phenotypes. Our work suggests that many causal variants may not be shared between populations, indicating that variants ascertained in European populations may not be informative in other populations. This could occur because, on average, more European-specific variants have been either directly included in GWASs or imputed more often than variants specific to other non-European populations. To ensure equal predictive power of polygenic scores across populations, whole-genome sequencing-based association studies must be undertaken in non-European populations. Such studies would allow for unbiased discovery of private variants accounting for much of the heritability, resulting in improved polygenic prediction in non-European populations. Finally, large imputation panels from the relevant population of interest are necessary to include variation that is not present in Europe.

Data and code availability

The scripts required to carry out the inference of heritability from private variants can be found at https://github.com/LohmuellerLab/PRS.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.02.013.

Acknowledgments

We thank Sriram Sankararaman, James Boocock, Alec Chiu, and Ruth Johnson for helpful discussions and Bogdan Pasaniuc, Nelson Freimer, and members of the Lohmueller lab for helpful comments on a draft of this manuscript. A.D. is funded by NSF Graduate Research Fellowship DGE-1650604 and K.E.L. is funded by NIH grant R35GM119856.

Declaration of interests

The authors declare no competing interests.

Received: September 15, 2020
Accepted: February 17, 2021
Published: March 9, 2021

Web resources

GWAS summary statistics, http://www.nealelab.is/uk-biobank

References

1. Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
2. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
3. Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A., and Yang, J. (2017). 10 Years of GWAS discovery: Biology, function, and translation. Am. J. Hum. Genet. *101*, 5–22.

4. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.
5. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.R., Bhatia, G., Do, R., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.
6. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
7. Khera, A.V., Chaffin, M., Wade, K.H., Zahid, S., Brancale, J., Xia, R., Distefano, M., Senol-Cosar, O., Haas, M.E., Bick, A., et al. (2019). Polygenic prediction of weight and obesity trajectories from birth to adulthood. Cell *177*, 587–596.e9.
8. Natarajan, P., Young, R., Stitziel, N.O., Padmanabhan, S., Baber, U., Mehran, R., et al. (2017). Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. Circulation *135*, 2091–2101.
9. Maas, P., Barrdahl, M., Joshi, A.D., Auer, P.L., Gaudet, M.M., Milne, R.L., Schumacher, F.R., Anderson, W.F., Check, D., Chattopadhyay, S., et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. JAMA Oncol. *2*, 1295–1302.
10. Scutari, M., Mackay, I., and Balding, D. (2016). Using genetic distance to infer the accuracy of genomic prediction. PLoS Genet. *12*, e1006288.
11. Martin, A.R., Gignoux, C.R., Walters, R.K., Wojcik, G.L., Neale, B.M., Gravel, S., Daly, M.J., Bustamante, C.D., and Kenny, E.E. (2017). Human demographic history impacts genetic risk prediction across diverse populations. Am. J. Hum. Genet. *100*, 635–649.
12. Kim, M.S., Patel, K.P., Teng, A.K., Berens, A.J., and Lachance, J. (2018). Genetic disease risks can be misestimated across global populations. Genome Biol. *19*, 179.
13. Martin, A.R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B.M., and Daly, M.J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. Nat. Genet. *51*, 584–591.
14. Mostafavi, H., Harpak, A., Agarwal, I., Conley, D., Pritchard, J.K., and Przeworski, M. (2020). Variable prediction accuracy of polygenic scores within an ancestry group. eLife *9*, e48376.
15. Ragsdale, A.P., Nelson, D., Gravel, S., and Kelleher, J. (2020). Lessons learned from bugs in models of human history. Am. J. Hum. Genet. *107*, 583–588.
16. Novembre, J., and Barton, N.H. (2018). Tread lightly interpreting polygenic tests of selection. Genetics *208*, 1351–1355.
17. Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. Nature *570*, 514–518.
18. Berg, J.J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A.M., Mostafavi, H., Field, Y., Boyle, E.A., Zhang, X., Racimo, F., Pritchard, J.K., and Coop, G. (2019). Reduced signal for polygenic adaptation of height in UK Biobank. eLife *8*, e39725.
19. Sohail, M., Maier, R.M., Ganna, A., Bloemendal, A., Martin, A.R., Turchin, M.C., Chiang, C.W., Hirschhorn, J., Daly, M.J., Patterson, N., et al. (2019). Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. eLife *8*, e39702.
20. Martin, A.R., Lin, M., Granka, J.M., Myrick, J.W., Liu, X., Sockell, A., Atkinson, E.G., Werely, C.J., Möller, M., Sandhu, M.S., et al. (2017). An unexpectedly complex architecture for skin pigmentation in Africans. Cell *171*, 1340–1353.e14.
21. Lohmueller, K.E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet. *10*, e1004379.
22. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.
23. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
24. Gao, F., and Keinan, A. (2014). High burden of private mutations due to explosive human population growth and purifying selection. BMC Genomics *15* (*Suppl 4*), S3.
25. Eyre-Walker, A. (2010). Evolution in health and medicine Sackler colloquium: Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc. Natl. Acad. Sci. USA *107* (*Suppl 1*), 1752–1756.
26. Sanjak, J.S., Long, A.D., and Thornton, K.R. (2017). A Model of compound heterozygous, loss-of-function alleles is broadly consistent with observations from complex-disease GWAS datasets. PLoS Genet. *13*, e1006573.
27. Uricchio, L.H. (2020). Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. Hum. Genet. *139*, 5–21.
28. Hernandez, R.D., Uricchio, L.H., Hartman, K., Ye, C., Dahl, A., and Zaitlen, N. (2019). Ultrarare variants drive substantial cis heritability of human gene expression. Nat. Genet. *51*, 1349–1355.
29. Gazal, S., Finucane, H.K., Furlotte, N.A., Loh, P.R., Palamara, P.F., Liu, X., Schoech, A., Bulik-Sullivan, B., Neale, B.M., Gusev, A., and Price, A.L. (2017). Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. Nat. Genet. *49*, 1421–1427.
30. Gazal, S., Loh, P.R., Finucane, H.K., Ganna, A., Schoech, A., Sunyaev, S., and Price, A.L. (2018). Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. Nat. Genet. *50*, 1600–1607.
31. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L., Yap, C.X., Xue, A., Sidorenko, J., McRae, A.F., et al. (2018). Signatures of negative selection in the genetic architecture of human complex traits. Nat. Genet. *50*, 746–753.
32. Schoech, A.P., Jordan, D.M., Loh, P.R., Gazal, S., O'Connor, L.J., Balick, D.J., Palamara, P.F., Finucane, H.K., Sunyaev, S.R., and Price, A.L. (2019). Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. Nat. Commun. *10*, 790.
33. Uricchio, L.H., Kitano, H.C., Gusev, A., and Zaitlen, N.A. (2019). An evolutionary compass for detecting signals of polygenic selection and mutational bias. Evol. Lett. *3*, 69–79.
34. Wainschtein, P., Jain, D.P., Yengo, L., Zheng, Z., Cupples, L.A., Shadyab, A.H., McKnight, B., Shoemaker, B.M., Mitchell, B.D., Psaty, B.M., et al. (2019). Recovery of trait heritability from whole genome sequence data. bioRxiv. https://doi.org/10.1101/588020.

35. Young, A.I. (2019). Solving the missing heritability problem. PLoS Genet. *15*, e1008222.
36. de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
37. Haller, B.C., and Messer, P.W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. Mol. Biol. Evol. *36*, 632–637.
38. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D.; and 1000 Genomes Project (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. USA *108*, 11983–11988.
39. Kim, B.Y., Huber, C.D., and Lohmueller, K.E. (2017). Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics *206*, 345–361.
40. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D., and Mountain, J.L. (2015). The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am. J. Hum. Genet. *96*, 37–53.
41. Wakeley, J., and Hey, J. (1997). Estimating ancestral population parameters. Genetics *145*, 847–855.
42. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
43. Turcot, V., Lu, Y., Highland, H.M., Schurmann, C., Justice, A.E., Fine, R.S., Bradfield, J.P., Esko, T., Giri, A., Graff, M., et al. (2018). Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat. Genet. *50*, 26–41.
44. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
45. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
46. Lam, M., Chen, C.Y., Li, Z., Martin, A.R., Bryois, J., Ma, X., Gaspar, H., Ikeda, M., Benyamin, B., Brown, B.C., et al. (2019). Comparative genetic architectures of schizophrenia in East Asian and European populations. Nat. Genet. *51*, 1670–1678.
47. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986.
48. Rosenberg, N.A., Edge, M.D., Pritchard, J.K., and Feldman, M.W. (2018). Interpreting polygenic scores, polygenic adaptation, and human phenotypic differences. Evol. Med. Public Health *2019*, 26–34.
49. Harmon, A. (2018). Why white supremacists are chugging milk (and why geneticists are alarmed). The New York Times, October 17, 2018. https://www.nytimes.com/2018/10/17/us/white-supremacists-science-dna.html.
50. Fuentes, A., Ackermann, R.R., Athreya, S., Bolnick, D., Lasisi, T., Lee, S.H., McLean, S.A., and Nelson, R. (2019). AAPA statement on race and racism. Am. J. Phys. Anthropol. *169*, 400–402.
51. Saini, A. (2019). Superior: The Return of Race Science (Beacon Press).
52. Coop, G. (2019). Reading tea leaves? Polygenic scores and differences in traits among groups. arXiv, 1909.00892. https://arxiv.org/abs/1909.00892.