# Integrating Ancient and Modern DNA To Study Human History in South Asia and the Americas

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Committee on Higher Degrees in Systems, Synthetic, and Quantitative Biology

have examined a dissertation entitled

*Integrating Ancient and Modern DNA To Study Human History in South Asia and the Americas*

presented by Nathan Nakatsuka

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

*Signature* _____

*Typed name*: Prof. Alkes Price

*Signature* Michael Desai (May 13, 2020) _____

*Typed name*: Prof. Michael Desai

*Signature* _____

*Typed name*: Prof. Sohini Ramachandran

*Date*: May 13, 2020

# Integrating Ancient and Modern DNA To Study Human History in South Asia and the Americas

A dissertation presented

by

Nathan Nakatsuka

to

The Committee on Higher Degrees in Systems, Synthetic, and Quantitative Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Systems, Synthetic, and Quantitative Biology

Harvard University

Cambridge, Massachusetts

May 2020

Dissertation Advisor: David Reich Nathan Nakatsuka

**Integrating Ancient and Modern DNA To Study Human History in South Asia and the Americas**

## Abstract

In the last two decades, advances in next-generation sequencing, genome-wide genotyping arrays, and methods to obtain DNA from ancient individuals have propelled the field of population genetics such that it is now an extraordinarily powerful tool for inferring human history. This thesis focuses on the study of DNA from both present-day and ancient individuals who existed between 50 to over 10,000 years ago. The analyses focus on South Asian, African-American, and Native American history and some applications of these historical insights for improving human health.

Chapter 1.1 provides a background to the field, including a survey of the current methods for obtaining DNA from ancient individuals, the major statistical techniques for using genetics to infer human history, and past ways that archaeological and linguistic information have been integrated with genetics to arrive at more robust, contextualized models of the past. Chapter 1.2 provides a philosophical framework for conducting ethically-responsible genetics research in non-Western cultural contexts with a focus on South Asian and Native American groups. This ethical framework forms the groundwork for the studies in this thesis, and each subsequent section is introduced with a statement on how the study's approach fits into this framework.

Chapter 2 details a study on over 260 distinct present-day South Asian groups, demonstrating that 81 of these groups, including 14 with estimated census sizes over one million, descend from founder events more extreme than those in Ashkenazi Jews and Finns, both of which have high rates of recessive disease due to founder events. These founder events are population bottlenecks that can be detected on the basis of identity-by-descent (IBD) segments shared within the last 100 generations from a common founder, and they highlight an underappreciated opportunity for recessive disease-associated gene mapping in South Asia. Chapter 3 is an admixture mapping study that examines the increased European ancestry at chromosome 1 of African-Americans with Multiple Sclerosis (MS) relative to those without the disease and determines that the signal is due to two genetic variants within the *CD58* and *FCRL3* genes that together predict a 1.44-fold greater risk for MS in European-Americans compared to African-Africans. Chapter 4 introduces a new software, *ContamLD*, for estimating contamination in autosomal ancient DNA (aDNA) by measuring the breakdown of linkage disequilibrium in a sequenced individual due to the contaminant DNA.

Chapter 5 primarily focuses on three aDNA studies of Central and South America. The first study explores the earliest migrations into Central and South America from 11,000 to 4,000 years ago, while the next two studies focus on the changes in genetic structure over time in the Andes and Patagonia, including genetic analysis of individuals from major Andean cultures, such as the Wari, Tiwanaku, and Inca, as well as ancient individuals from the regions associated with the major Patagonian groups found at the time of European contact with representation of both maritime and terrestrial diet groups. Chapter 6 ties together the insights learned in these studies and compares them to findings of other world regions, highlighting surprising similarities and differences. Overall, this thesis presents just a small snapshot of the power of population genetics for inferring human history and how these insights can sometimes be used in unique ways to advance human health.

# Table of Contents

**Acknowledgments**

I find it surreal that my PhD is finally coming to a close. The past 5 years have been some of the most enjoyable years of my life, full of exploration and emotional, academic, and spiritual growth, that has been at times challenging but ultimately immensely rewarding. My first acknowledgments go to the Reich lab, where together everyone created an incredibly supportive and friendly environment that made it a joy to come into work every day. David in particular has been an incredible mentor who has always supported both my academic and personal development. He allowed me to maintain a flexible schedule while responding to emails in a timely fashion and teaching me how to write scientific papers, make understandable presentations, work with multiple collaborators at the same time, manage a lab, and balance family life with work. He truly has been an exceptional role model for my life. I would also like to thank Nick Patterson for his guidance, mentorship, and support, particularly in developing *ContamLD* and all of the software questions I have bugged him with over the years. I am extremely grateful to Nadin Rohland and the rest of the wet lab for all of the processing of the skeletal material and the data generation. I want to thank Swapan "Shop" Mallick and Matthew Mah for bioinformatic support (and in particular Shop for all of the numerous data processing issues he has helped me with over the years). Éadaoin Harney has been a wonderful fellow grad student in the lab with invaluable contributions to the *ContamLD* project. Iosif Lazaridis, Mark Lipson, Vagheesh Narasimhan, Iñigo Olalde, Pontus Skoglund, Priya Moorjani, Arti Tandon, Heng Li, and Iain Mathieson have all helped me immensely throughout my PhD, teaching me many things and sometimes collaborating on different projects. Jakob Sedig has provided great discussions about archaeology and helpful comments on many of my write-ups, especially the ethics section in Chapter 1.2. Lastly, Michelle Lee has been a remarkable administrative assistant who has helped support me through the many administrative hurdles of PhD life.

I would next like to thank all of the incredible collaborators outside of the Reich lab involved with each of the projects in this thesis. It was wonderful working with Kumarasamy Thangaraj and Niraj Rai throughout the South Asia project as they helped me to think through the work and explain additional details about each of the groups in the study. Girisha Katta Mohan and Sudha Srinivasan helped enormously with their clinical genetics knowledge and provision of samples of Indians with genetic disease. For the African-American admixture mapping project, Nikolaos Patsopoulos, Ashley Beechan, Nicolas Altemose, and Jorge Oksenberg in particular provided great insights and analyses vital to the success of the project. For the Central and South America project of Chapter 5.1, Cosimo Posth was an amazing co-first author who was a joy to work with. André Strauss, Johannes Krause, and César Méndez were also outstanding collaborators who coordinated their teams for the Chapter 5.1 project. Doug Kennett, Brendan Culleton, and Thomas Harper were of great help in radiocarbon dating and marine calibration through all the Native American studies. Lars Fehren-Schmitz, as well as Elsa Tomasto-Cagigao, and Gustavo Politis, have been extraordinary collaborators for the studies in Chapters 5.1 and 5.2, and it has been great also working with Jacob Bongers for the Chincha Valley study of Chapter 5.2. Rodrigo Nores, Josephina Motti, Ricardo Guichón, Graciela Cabana, and Pierre Luisi were all invaluable collaborators in Chapter 5.3 who had astute archaeological insights, excellent analytical and figure-making skills (Pierre), and very thorough proof-reading of the manuscripts. The other collaborators not mentioned here all served integral roles to each

of these studies, and I am deeply grateful for all of their support and help over the years. In addition, I am deeply grateful to the Summer Internship for Indigenous Peoples in Genomics (SING) for their helpful discussions regarding ethics of genetics research in Native American communities.

Throughout my PhD years, I have been deeply privileged to have the support of the MD/PhD office, especially Loren Walensky, Amy Cohen, Marcia Goldberg, Jennifer DeAngelo, Robin Lichtenstein, and Yi Shen. The Systems, Synthetic, and Quantitative Biology program has been extraordinarily supportive during my PhD, especially Elizabeth Pomerantz, Samantha Reed, Timothy Mitchison, and Andrew Murray. My Dissertation Advisory Committee (Alkes Price, Michael Desai, and Joel Hirschhorn) provided extremely helpful advice and constructive criticism to guide my PhD journey, and I also thank Sohini Ramachandran, in addition to Alkes and Michael, for being my thesis readers.

During my PhD I have also received support on a personal level from the Native American Health Organization (NAHO), the Harvard University Native American Program (HUNAP), the Office of Recruitment and Minority Affairs (ORMA) at HMS, and the Harvard GSAS Office of Diversity and Minority Affairs office (especially Sheila Thomas and Karina Gonzalez). It has been particularly great mentoring the Four Directions summer undergraduate program, the Native American High School Summer Program, and MissionSafe, as well as doing under-represented minority recruitment through all of these organizations at many conferences and programs over the past 7 years.

I have gained deep spiritual and emotional support through the Longwood Christian Community (LCC), Reality Boston church, the Christian Medical and Dental Association (CMDA), and Boston Healthcare Fellowship (BHF) and would not have been able to make it without them. I would like to thank all of the friends I have made over the years in Boston who have also supported me on my journey. My deepest gratitude and thanks go to my family, especially my parents and brothers who have supported me every step of the way and provided the environment that helped me to get here in the first place. Lastly, I would like to thank God the Father, His son Jesus Christ, and the Holy Spirit for their work in my life, in particular preserving me and giving me the strength to endure through this long process. It has been the greatest privilege being able to study how He has put together the universe and orchestrated the events in history that leave their traces in the genomes of our species.

# Chapter 1: Background

## Chapter 1.1: Background to the Field of Population Genetics

Over the past 70 years, the field of genetics has expanded in scope a remarkable extent. Fueled by revolutions in molecular biology and DNA sequencing, genetics has been able to pervade all fields of medicine and even address many questions in the social sciences. Much of the recent success of this field has been due to the development of next generation sequencing, which has made it possible to determine the genetic code of almost any species of interest very quickly and at low cost. The relative ease of sequencing has facilitated studies of broader scope and deeper resolution, which has led to improved insight about many different organisms, especially humans.

One of the biggest applications of human genetics has been to biomedicine where essentially all organ systems can be studied with standard genetic tools. Some key aims in medical genetics are to find genetic variants and biological pathways underlying phenotypic traits, including disease, response to drugs, or the effect of the environment on different individuals. One goal of these studies is to improve our understanding of the biology underlying disease to facilitate the discovery of potential drug targets. Another goal is to elucidate the allelic architecture of different diseases, meaning the determination of whether particular diseases are due primarily to many common variants of small effect size, fewer rare variants of larger effect size, or other possible combinations. Common ways to achieve these goals are to perform large scale association studies where thousands to millions of individuals with and without the trait of interest have their genomes assessed with either genotyping chips, whole-exome sequencing, or whole-genome sequencing, to find genetic variants associated with the trait. Genotyping chips (SNP arrays) generally cover between 300,000-1.5 million single nucleotide polymorphisms (SNPs) spread throughout the genome and usually facilitate the discovery of variants more common in the population (>1% frequency) that lead to small differences in predisposition to the disease, while whole-exome and whole-genome sequencing usually is focused on the discovery of variants less common in the population which might have larger effect size on the trait of interest. Whole-exome and whole-genome sequencing have led to the successful identification of hundreds of cases where a single genetic variant causes a disease (these diseases are known as Mendelian genetic disorders) [1-3]. On the other side of the spectrum, chip-based genome wide association studies (GWAS) have led to the discovery of thousands of common SNPs associated with different traits [4]. The associations discovered with all of these methods have led to novel biological insights, therapeutic development, drug repurposing, and improved clinical care [5].

A more surprising area where genetics has made major inroads is in the fields of archaeology and history. In the past 20 years, the DNA (and sometimes amino acids [6]) of present-day and ancient species have been studied to answer questions about the

migrations (who, when, and to what extent), admixtures, and population sizes of a wide range of organisms, especially humans [7]. Human history in particular can be determined by integrating information from archaeological studies of material remains with linguistic analyses of human languages and genetic analyses of human DNA. The field of population genetics has been revolutionized in the past ten years by the new technology of ancient DNA sequencing, which allows geneticists to study the DNA of individuals who lived up to hundreds of thousands of years ago. At the beginning of this field, ancient DNA was limited to mitochondria (for which there are often thousands of copies per cell) and then PCR analyses of small fractions of the nuclear genome [8]. However, the advent of next generation sequencing finally allowed efficient sequencing of the entire genome, causing a rapid advance in the field of paleo-genomics. With this technology geneticists can now determine the genetic differences across space and time, meaning they can compare individuals and groups from different geographical regions or the same region over time to look for potential changes due to mixtures from external groups, genetic drift (random genetic changes over time), or natural selection.

**Methods for Obtaining DNA from Ancient Individuals**

DNA can be obtained from ancient human skeletal material by taking a small sample from a bone and creating a powder by drilling (Figure 1). The petrous bone, which is part of the temporal bone, has the highest DNA content, likely due to the fact that it is the hardest and most dense bone in the mammalian body [9, 10]. Tooth cementum are the second most widely used material, followed by long bone. After powder is obtained, the DNA is extracted from the bone by lysing the cells and binding the DNA to silicon dioxide (silica) particles (either spin columns or silica-coated magnetic beads), which is optimized to isolate highly degraded DNA in the form of very short molecules (~25-40 base pairs, bp) [11, 12]. Lysis buffers are used that have proteinase K to digest bone and tooth collagen and ethylenediaminetetraacetate (EDTA) to decalcify the bone and tooth matrix and release DNA. This process has been automated in protocols that use silica-coated magnetic beads [11]. The DNA is often then treated with the enzyme Uracil-DNA Glycosylase (UDG) to remove the vast majority of artifacts from cytosine-to-thymine mutations characteristic of ancient DNA [13]. Following this, DNA libraries are made from the purified DNA, which involves the ligation of adaptor oligonucleotides to the ends of the DNA fragments to facilitate unique identification of the DNA fragments and subsequent sequencing. The libraries are made using either double-stranded or single-stranded library preparation protocols; the single-stranded protocol has the advantage of higher DNA yield but is more expensive and time intensive [14]. After libraries are made, the DNA sequences are often enriched for those that overlap mitochondrial DNA [15] and about 1.24 million nuclear targets chosen to cover SNPs of interest for population genetic analyses [16-18]. The enriched products are then sequenced on a next-generation machine, such as the Illumina NextSeq500. The enrichment is useful to keep costs low and to improve DNA yield from samples with a low percentage of human DNA. However, whole-genome shotgun sequencing allows one to obtain more power for many analyses as will be described below. The sequencing data usually have their ends trimmed to remove the adaptors and residual mutations

from DNA damage, and the resulting data is aligned to the human genome using custom parameters that account for the smaller fragments of ancient DNA. Since ancient DNA data is often of relatively low coverage (~1x), it is not possible to have reliable heterozygous calls, so a random allele is taken at each SNP position to represent the genotype of the individual at that position ("pseudo-haploid" genotypes).

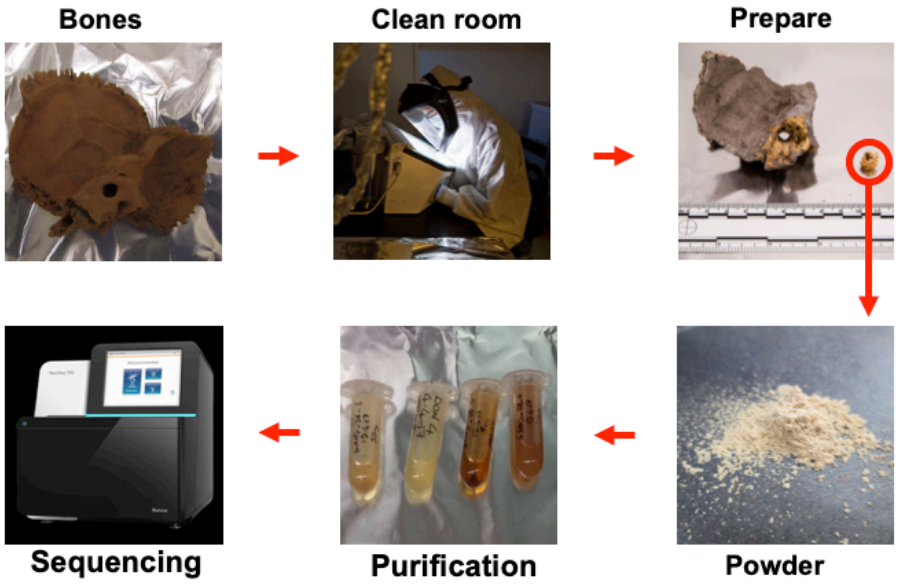


**Figure 1. Ancient DNA processing.** A summary of steps taken to obtain DNA sequencing data from skeletal bone.

**Statistical Methods for Inferring Human History Using Genetics**

*Analyses of Uni-Parental DNA*

As the molecular biology methods for generating ancient DNA data advanced, so did the computational and statistical methods for analyzing the new data. Population genetics can be divided into methods based on analysis of uni-parental data and methods based on analysis of autosomal data. Uni-parental data are mitochondria, which are inherited essentially only from one's mother, and Y-chromosomes, which are inherited only from one's father. Both mitochondria and most of the Y-chromosome are non-recombining, meaning they are passed down each generation without mixing with the material inherited from the individual's other parent. Thus, the only change that occurs over generations is due to random mutations. Uni-parental DNA is usually analyzed based on haplotypes of the mitochondria or Y-chromosomes, which are more generally defined to be groups of genetic variants that are often inherited together. In these analyses, uni-parental DNA haplotypes can be arranged into hierarchical clusters with groupings that reflect temporal relationships (i.e. groups more similar to each other share more genetic history) (Figure 2). These hierarchical relationships can reveal history, because they can show how different groups are related to each other and potential mixtures between groups (e.g. if one group has two different sets of uni-parental haplotypes, it could be a mixture of genetic ancestry related to two groups that only have one of the sets of haplotypes) [19]. They also can show evidence for

population bottlenecks, if the diversity of haplotypes drops markedly [20], or population expansions, if the diversity suddenly increases [21, 22], and this has been systematized for mitochondria in the software BEAST [23, 24], which provides Bayesian estimates of population size over time. In addition, the sex-specific nature of uni-parental DNA means that they can provide insight into sex-specific differences in demographic history. For example, if migration occurred primarily via males, then the Y-chromosome haplotypes might change in the population that received the male influx, while its mitochondrial haplotypes would stay the same [25].

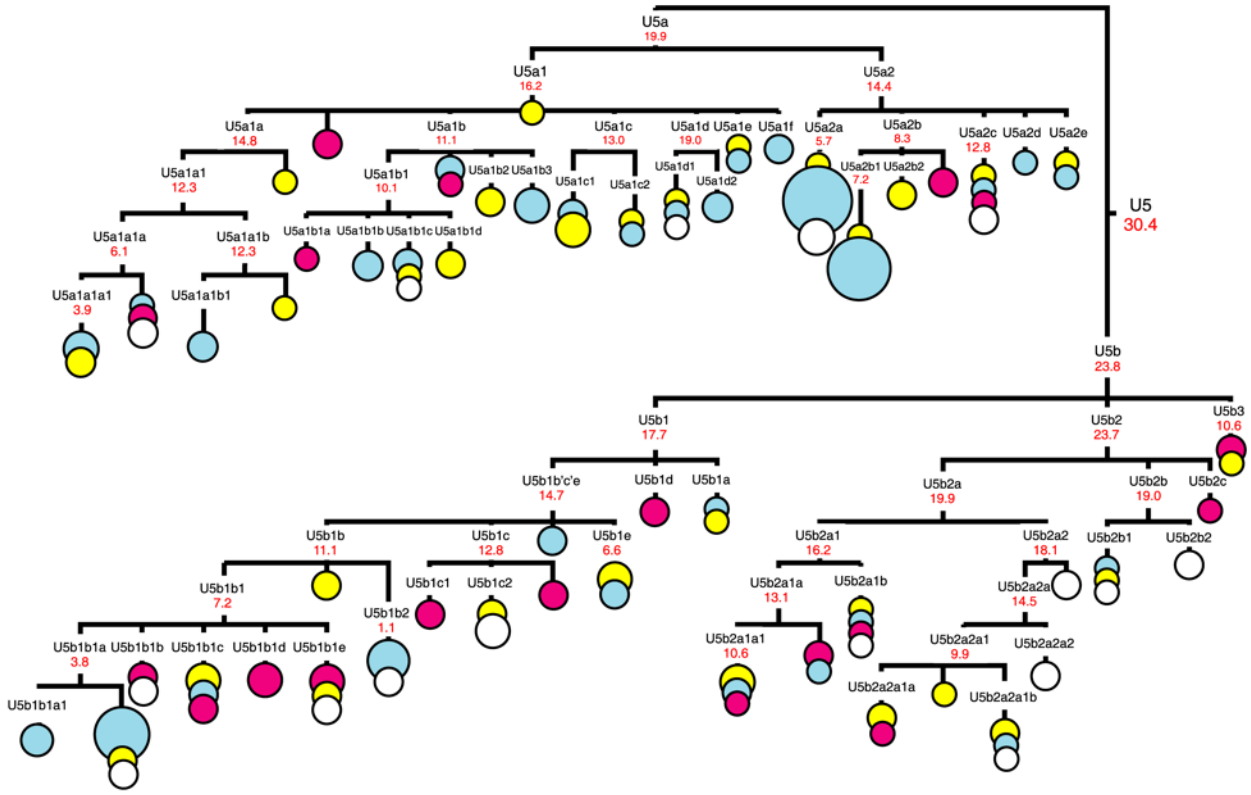


**Figure 2. An example of a mitochondrial-based phylogenetic tree**. As published in Malyarchuk *et al*. [26].

*Analyses of Autosomal DNA*

Population genetic methods for analyzing autosomal DNA in principle should be more powerful than the analysis of uni-parental DNA, because the number of independent markers is many fold higher in the autosomes. A large number of methods have been developed to study this data, exploiting various features that provide information about demographic history. Below I will briefly summarize the main methods that are currently in use in the field.

*Determination of Kinship*

One important analysis in archaeology is the determination of kinship. This can be determined with a variety of methods. One method uses mismatch rates, where more closely related individuals have lower mismatch rates. This was used, for example,

to uncover the existence of a matrilineal dynasty ~1,000 years ago in the Chaco Canyon of Southwestern United States [27]. The software PLINK [28, 29] infers the amount of the genome between pairs of individuals that are identical-by-state (IBS), meaning they happen to be genetically the same at those SNPs. This is then used to determine the amount of the genome that is identical-by-descent (IBD), meaning they are identical due to common ancestry. Closely related individuals will share a large amount of their genome IBD. For example, children will share 50% of their genome IBD with their biological parents. In most genetic analyses, closely related individuals are removed from further analyses due to their potential to bias the statistics, since the related individuals are not independent. However, the presence of close relatives can point to interesting insights into the cultures being studied (e.g. burial patterns of relatives). This is thus an important complement to other genetic analyses that assess population structure on a broader scale.

*Dimensionality Reduction*

One set of methods that allow qualitative analysis of the data for population structure are dimensionality reduction techniques. These methods show the most important aspects of variation in the data, which can point to past mixtures between different groups. The most widely used method is Principal Components Analysis (PCA), which is a transformation of a matrix to create a set of vectors that are linearly uncorrelated with each other and ordered by the amount of variance they explain in the data [30]. This can be done directly on the genotype matrix or on another matrix that shows genetic distance between the individuals. In addition, one set of individuals can be projected onto the principal components inferred by another set of individuals to see how the genetic variation in the second set of individuals is present (or not) in the projected set of individuals. Often ancient individuals are projected onto the principal components inferred by present-day individuals, in part because of the missing data in most ancient individuals (since they are usually sequenced to much lower coverage than the present-day individuals) and also because this can more clearly show past admixtures (if the ancient individuals are more “extreme” than the present-day individuals, this is often due to the present-day individuals being a mixture of genetic ancestry related to the ancient individuals) (Figure 3A). This process requires proper normalization of the principal components, because due to the missing data and the lack of shared genetic drift with more recent individuals, ancient individuals can be biased to appear closer to the origin than they should be [31]. The linear nature of PCA is a positive for population genetics, because it means that the distances in each dimension are usually more directly correlated with genetic distance and can sometimes point to past admixture [32]. One example of this is the presence of admixture clines in PCA produced when groups on the cline are variable proportions of ancestry from two different groups. An example of this is the “Indian cline” produced because many groups in India have variable proportions of mixture between a group related to present-day Europeans, called Ancestral North Indian, and a group distinct from both East Asians and Europeans, called Ancestral South Indian [30] (Figure 3B).

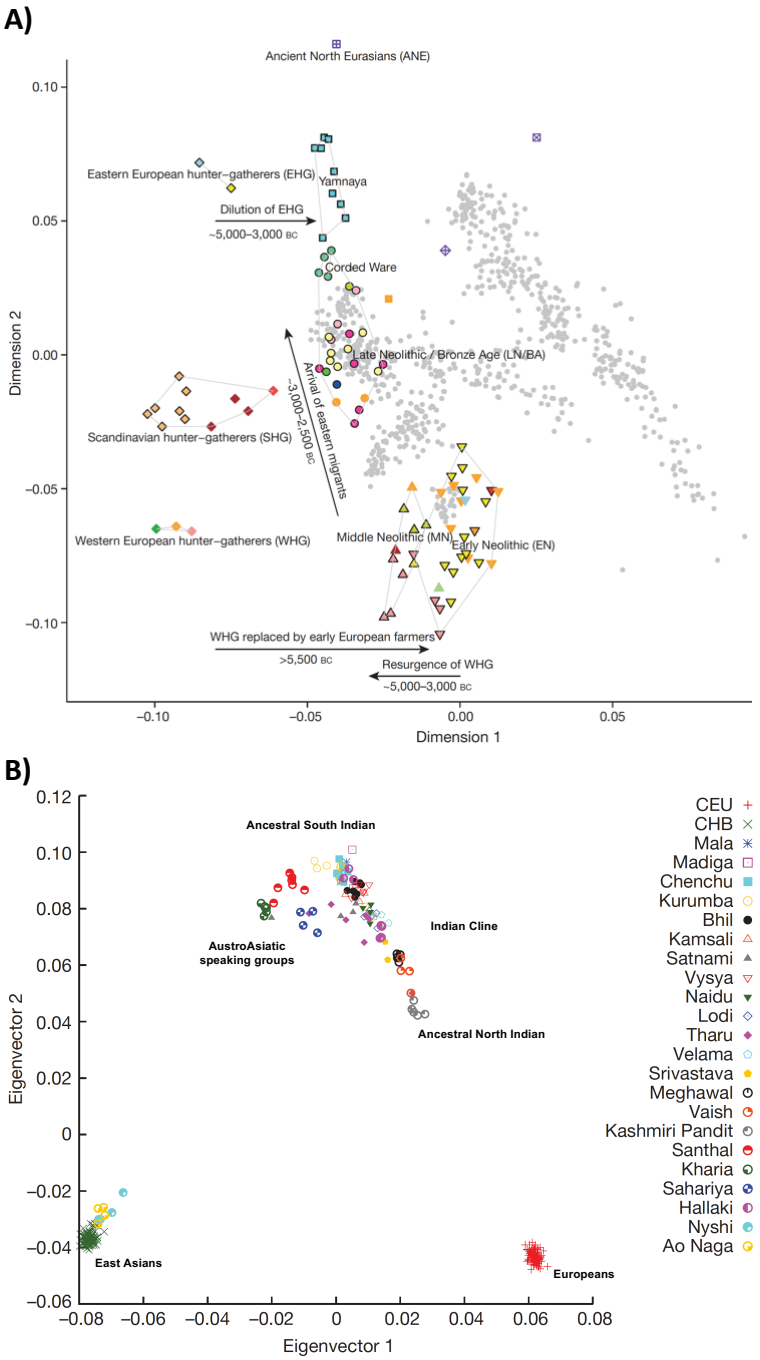


**Figure 3. Two different examples of PCAs from previously published papers. (A)** The principal components (PCs) for this PCA were built from present-day European and Near East individuals (gray dots), and the ancient individuals were projected onto these PCs [17]. The ancient hunter-gatherers and Neolithic farmers (colored shapes) are more "extreme" on the PCA, because the present-day individuals are a mixture of ancestry related to these individuals and thus fall in more intermediate positions. **(B)** PCA of present-day South Asian populations, showing the Indian cline that was produced from an admixture between a group with greater relatedness to present-day Europeans

(called Ancestral North Indian) and a group diverged from both Europeans and East Asians (called Ancestral South Indian). Adapted from Reich *et al*., 2009 [33].

The linear nature of PCA also leads to visualization difficulties, however. For example, if there are groups with large genetic divergence with other groups in the dataset, then this signal can sometimes mask the smaller genetic differences between the other groups in the dataset. Other dimensionality reduction techniques that get around this difficulty are t-SNE (t-distributed stochastic neighbor embedding) [34], which has primarily been used in single-cell RNA analyses, and UMAP (Uniform Manifold Approximation and Projection) [35], which has been used in some ancient DNA studies [36]. These techniques adapt distances so that both global and local population structure can be visualized well. Lastly, multi-dimensional scaling (MDS) is another form of non-linear dimensionality reduction that can be used for data visualization. This method reduces the data to a number of dimensions specified by the user in a way that preserves between-individual distances (PCA, by contrast, is preserving the covariance of the data) usually by minimizing a metric called "stress", which is the difference between the actual distances of the individuals and their predicted values based on the MDS model [37].

*Global Ancestry Clustering*

Related to dimensionality reduction techniques are general clustering techniques that can also provide qualitative overviews of the data. One common technique is to fit the genetic data in a maximum likelihood model such that each individual gets assigned a certain proportion of ancestry from a number of populations specified by the user. Algorithms for this include STRUCTURE [38] and ADMIXTURE [39], which iteratively place proportions of ancestries into the different populations until the best fit is found. Users usually run the algorithm and specify increasing numbers of populations starting from K=2 until the global likelihood of the data starts to decrease, and the smallest value of K that maximizes the global likelihood of the data is thought to be a more likely representative of the underlying population history. These methods can lead to false inferences of population history, however, because high genetic drift in a group can cause it to be modeled as its own population even if that is not the best model of history [40].

*Local Ancestry Inference*

These global ancestry techniques measure overall proportion of ancestry in individuals, but it is also sometimes possible to do local ancestry measurements where the genome is partitioned into different segments with ancestry from each specific population (Figure 4). This can be done in either a supervised or unsupervised manner, where the supervised case uses training data, which in this case is genetic data of individuals from the populations of interest, to obtain the genetic distributions that can then be used to label the genomes of the test individuals with these populations. In the unsupervised case, the genetic distributions of the background populations are learned directly from the test individuals. Many different statistical/computational tools exist to

perform these tests (reviewed here [41]) and often rely on either Hidden Markov Models or directly maximizing the likelihood of the data given the proposed model. Local ancestry estimation can be useful for admixture mapping (Chapter 3), where admixed groups are used to localize genomic regions underlying diseases with very different prevalence between the two different populations that admixed to form the admixed group (e.g. studying African-Americans, who are admixed between European and West African ancestry, to determine the genetic basis for the higher rates of Multiple Sclerosis in Europeans relative to West Africans) [42].

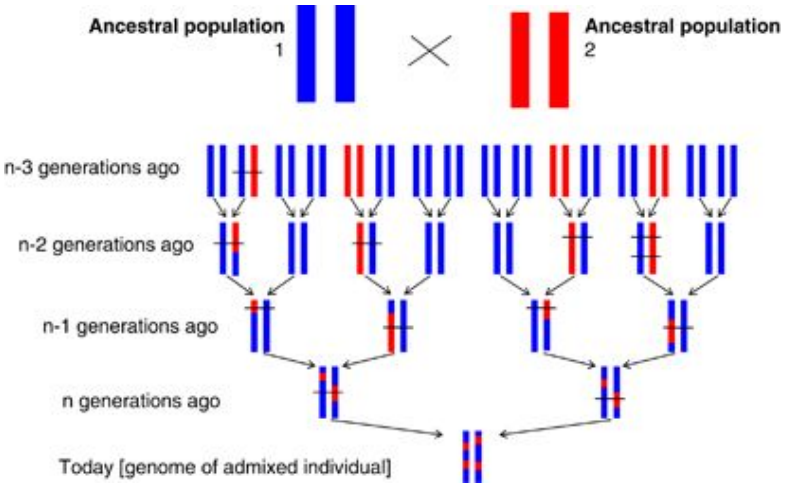


**Figure 4. Schematic of the genetic basis for local ancestry mapping.** The genomes of admixed individuals will have genetic segments from each of the parental populations, and different algorithms can be used to determine what population each part of the admixed individual came from. Image reproduced from Baye and Wilke, 2010 [43].

*Phylogenetic Trees*

Another set of methods for obtaining qualitative overviews of population structure are phylogenetic trees. These trees show the genetic relationships between different groups and can be obtained through a variety of methods. One popular method is neighbor-joining, which takes a distance matrix between the groups of interest and does an iterative procedure in which each group is joined to the group with smallest distance from it and then the procedure is repeated, comparing the previously joined groups with each other with repetition of this process until the tree is completely resolved [44]. Other methods for tree creation are maximum parsimony based trees, which create trees that minimize the total number of changes needed between the different groups [45], or UPGMA (unweighted pair group method with arithmetic mean), which is similar to neighbor-joining except that it does the group joining by measuring the average distance between the elements in each group with the elements of the other groups (relative to neighbor-joining, which creates a new node from the joined groups at each step rather than continuing to use the information from the original groups each time) [46]. Importantly, UPGMA assumes that all lineages are evolving at the same rate, while neighbor-joining trees do not make this assumption.

Some other methods for creating phylogenetic trees account for multiple sources of data in a Bayesian framework to find the most likely model [24].

In all of these phylogenetic trees, it is often useful to do bootstrapping to determine the strength of evidence for each node. Bootstrapping usually involves randomly re-sampling different subsets of the data and re-creating the trees with this data. Nodes are then labeled by the percentage of times they held up in all attempts at re-creating the trees. Nodes with higher values have more support for them representing the accurate relationships between the different groups in that particular dataset [47].

*The f-Statistic Framework*

One toolbox that has had remarkable success in population genetics is the *f*-statistic framework based on allele frequency correlations [48]. Unlike many of the methods described above, these methods allow formal testing for a history of population mixture rather than qualitative inference. These methods were inspired by $F_{ST}$ (fixation index) values, which were developed by Cavalli-Sforza and Edwards [49] to measure allele frequency differentiation between pairs of populations. $F_{ST}$ can be defined in different ways, but one way to define it is the variance in allele frequencies between pairs of populations divided by the total variance in the population [50]. If genetic differentiation is high between the two populations, then a large proportion of the total variance in allele frequencies will be between the populations, and $F_{ST}$ between the two groups will be high. Another way to define it is the value such that the allele frequency difference between the two populations has mean 0 and variance 2* $F_{ST}$ *p * (1-p), where p is the allele frequency in the population ancestral to the two populations. $F_{ST}$ values between different human populations range from ~0.20 between West African and non-African groups to ~0.002 between different European groups [50]. $F_{ST}$ can be corrected for inbreeding based on heterozygosity in the group (inbreeding will lead to very low heterozygosity), which is helpful because the $F_{ST}$ will be biased upward in these cases due to the artificial drift in the inbred group.

The *f*-statistics that form the core of a large number of modern population genetic tools are $f_2$, $f_3$, and $f_4$ (these are empirically determined from the data, while $F_2$, $F_3$, and $F_4$ are the underlying theoretical values of the assumed phylogeny corresponding to these statistics) [48]. An $f_2$-statistic of the form $f_2$(A,B) is the expectation value of the squared difference in allele frequencies of one population (A) with another population (B), which is the branch length between the two populations that is used for fitting edges in admixture graphs (described below). An $f_3$-statistic of the form $f_3$(C; A,B) is the expectation value of the product of the differences between the target population (C) and the other populations (A and B). In other words it is E[(c-a)*(c-b)], where c, a, and b are the frequencies of an allele at a particular SNP (the values are summed across all SNPs). A negative $f_3$-statistic can only occur if population C has ancestry from populations related to both A and B. Intuitively, it is because C has "intermediate" allele frequencies between A and B. Thus, the $f_3$-statistic can be used as a rigorous test for admixture. However, there are often false negatives with this test, because a large amount of genetic drift can cause a $f_3$-statistic to be positive even if C is in reality

admixed between groups related to A and B. If population C is an outgroup to A and B (often the Central African pygmy population, Mbuti, is used), then this outgroup-$f_3$-statistic can be used to measure shared genetic drift between A and B (with larger values indicating more shared drift), allowing comparison between different populations for genetic similarity.

A $f_4$-statistic is of the form $f_4(A,B; C,D) = E[(a-b)*(c-d)]$ where a, b, c, and d are the allele frequencies of the respective populations (A, B, C, and D) at a particular SNP and the statistic is summed across all SNPs. Often an outgroup to B, C, and D (e.g. Mbuti) is placed in position A, so the statistic shows whether C and D are a clade with respect to B (if the statistic is consistent with 0). On the other hand, If B has more shared history with C, then the statistic will be significantly negative (B would have significantly more shared alleles with C than with D), and if B has more shared history with D, then the statistic will be significantly positive. These statistics are not biased by more recent genetic drift in the test groups, because the extra drift would be uncorrelated to the ancestry of the other groups. If C and D are groups from the same physical location at different time periods, then if B has significant allele sharing with the more recent group, this would indicate that B shares more history with the more recent group, which could be due to admixture between the groups or that a group mixed into both B and the recent group. The D-statistic is the same as the $f_4$-statistic, except it is normalized to be between -1 and 1. The D-statistic is better for testing whether a particular phylogeny is accurate, so it was used to determine that modern humans have some Neanderthal ancestry [51]. Usually for $f_3$- and $f_4$-statistics, significance is declared at $|Z|>3$, which corresponds to $p<0.0013$ [48]. Ancestry proportions can be estimated with $f_4$-ratios. For example, with the phylogeny of Figure 5, an estimate of the proportion of ancestry related to group B that is in group X can be estimated as $\alpha=(f_4(A, O; X, C)/f_4(A,O; B, C))$. The intuition behind this is that if X has ancestry related to B, then the ancestry causes X to become more like B (and therefore more like A), which would cause the statistic $f_4(A, O; X, C)$ to become positive (due to X and A attracting). The fraction by which it is positive relative to $f_4(A, O; B, C)$ is the fraction of B-related ancestry that X has.

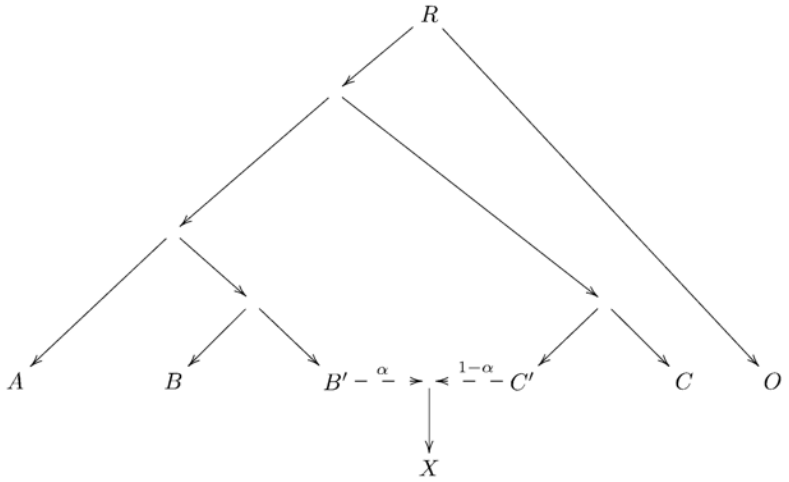


**Figure 5. An example phylogeny to explain $f_4$-ratio estimation**. From Patterson *et al*., 2012 [48].

The $f_2$-, $f_3$- and $f_4$-statistics can be used for several different population genetic analyses besides being used alone. For example, TreeMix uses $f_2$-statistics to fit a tree where admixture edges are allowed, which represent gene flow between different groups. The number of admixture edges are specified by the user and genetic drift approximated by Gaussian noise over time [52]. Similar to ADMIXTURE and STRUCTURE, the user will usually continue to add admixture edges until the addition of such an edge leads to limited change in likelihood of the model, meaning the edge is not necessary for the model. Similarly, *qpgraph* is a software that assesses user-specified admixture graphs and fits the $f_2$-, $f_3$- and $f_4$-statistics that correspond to the graph [48]. The resulting fit can be assessed by the largest Z-score (with large Z-scores corresponding to a poor fit), the total log-likelihood of the graph, or AIC (Akaike Information Criterion, which maximizes the likelihood while also penalizing additional parameters to prevent over-fitting). In this way a user can determine which population models are unlikely compared to models that fit the data well.

Another method based on *f*-statistics is *qpWave*, which assesses the number of "waves" of ancestry in a test set of groups relative to a set of outgroups [53]. For example, if admixed African-Americans, a West African group, and a Northern European group were in the test set, and the outgroup set were groups that could differentiate West Africans from Northern Europeans but could not differentiate within different West African or European groups (e.g. using East Asians as the outgroup set), then *qpWave* would show only two waves of ancestry for the 3 groups, since African-Americans generally can be modeled a linear combination of West Africans and Northern Europeans and thus would not show up as a distinct genetic population. *qpWave* does this by finding the rank of the matrix of $f_4$-statistics of the form $f_4(T_1,T_2; O_1,O_2)$ where $T_1$ and $T_2$ are all possible combinations of test populations, and $O_1$ and $O_2$ are all possible combinations of outgroup populations. The rank shows only the number of distinct ancestry sources, because groups that are admixtures of ancestry related to other groups in the test set will be a linear combination of those sources and thus will be eliminated from the matrix in the decomposition to reduced row-echelon form.

The last method based on *f*-statistics is *qpAdm*, which has the same set up as *qpWave*, but it calculates admixture proportions of a group based on potential source populations by fitting the proportions to weights of all relevant $f_4$-statistics of the same form as in *qpWave* [54]. The intuition behind this software is that the test set of populations have various relationships to the outgroups (note: they are not strict outgroups in the traditional sense of the definition, because they have differential relationships to the test populations, but they are called outgroups because they cannot have ancestry that admixed into the target group except for ancestry that came from the source groups). The relationships will be captured in $f_4$-statistics, and the target group can be modeled as a mixture of the source groups by fitting the $f_4$-statistics in a similar way to the $f_4$-ratio described above except weighted based on their values. If the proposed model is a good fit to the data, then the target group will have admixture proportions between 0 and 1, and the number of waves of ancestry will be consistent with n-1, where n is the total number of test groups, including the target group (because

the test groups will all be of distinct ancestry and the target group will be a linear combination of ancestries related to the test groups). If groups on a cline are modeled, it is possible for the admixture proportion to be negative if a more extreme group on the cline is modeled as a mixture of groups more intermediate on the cline. This method can also be used to assess potential sex-biased mixture by examining the differences in admixture proportions based on the autosomes relative to the proportions based on the X-chromosome. For example, if there were primarily male-mediated gene flow, then the proportions would be significantly higher on the autosomes than on the X chromosome, because males only transmit one X-chromosome [55].

*Linkage Disequilibrium*

Allele frequency correlation methods have the advantage that they are unbiased by population-specific drift and can use pseudo-haploid data with variable coverage, which is helpful for ancient DNA analyses [48]. On the other hand, they are missing potential power for measuring admixture as well as information about dates of the events, because they do not account for information from linkage disequilibrium (LD), which is the correlations between different SNPs near each other on a chromosome. These correlations can be induced by admixture between two populations previously separated for a long time, because an admixture event will cause the genetic variants from each ancestry to be correlated with those variants of the same ancestry in the admixed individual. LD breaks down over genetic distance due to recombination that occurs over generations, so after the initial admixture event, admixed individuals of later generations will have chromosomes that are mosaics of fragments of DNA from each ancestry. Hence, the correlations will be found only at closer and closer genetic distances with each passing generation (Figure 4).

The number of generations since an admixture event can be estimated by modeling the breakdown in LD in these admixed individuals. This is the basis for *ALDER* (Admixture-Induced Linkage Disequilibrium for Evolutionary Relationships), which uses an exponential curve to model the admixture LD breakdown and obtain estimates for the dates of the admixture event [56]. The amplitude of the curve allows one to estimate admixture proportions (larger proportions will lead to a larger amplitude). *MALDER* extends *ALDER* to situations where the admixed population had multiple admixture events in their history by fitting a sum of exponential curves rather than only one curve. This was used to determine that Khoisan groups (Southern African hunter-gatherers and pastoralists) have mixture related to Niger-Congo-speaking populations and west Eurasian ancestry indirectly through eastern Africa [57]. Other methods date admixture events by analyzing the tract length distribution after performing local ancestry inference [58, 59], but this can be difficult if there are poor reference populations or the admixing sources are very genetically similar to each other. A recently developed method, *DATES* [60], estimates the dates of admixture events by modeling ancestry covariance over genetic distance rather than admixture LD. By not directly using admixture LD, *DATES* has the advantage of not requiring high quality data (since ancestry covariance is computed separately for each individual and does not need

to be limited to the markers with complete data across all individuals as *ALDER* does) and even being able to work in a single individual.

*Haplotype Based Methods*

*ALDER* can be used to identify the most likely source for an admixture (it will give the strongest admixture LD signal). Similarly, the software *ChromoPainter* finds haplotypes in sequence data and models each group as a combination of the haplotypes of all potential groups. This information can then be inputted to *GLOBETROTTER*, which can identify all potential admixture events and date them using the same ideas as for *ALDER* [61]. The haplotype sharing amongst the groups can also be used for clustering as in the program *fineSTRUCTURE* [62], which in principle has more power than the previously described clustering algorithms, because they do not take haplotype information into account. Alternatively, the program *diCal2* uses haplotypes to determine different models of history by assessing their fit with specified phylogenetic models with admixture, because these models each have unique haplotype distributions that can be determined, for example, by using the conditional sampling distribution, which is the conditional probability of observing a new haplotype given a set of already observed haplotypes [63].

*Identical-by-Descent (IBD) DNA Fragments*

Related to haplotype analyses are analyses of IBD fragments. These were mentioned previously as useful for determining kinship, but they also are useful for inferring other aspects of population history, particularly more recent history (the past 100 generations). IBD segments can be inferred based on being unlikely to have arisen by chance and thus being likely to have been inherited by both individuals from a recent shared ancestor. One method to infer IBD segments is to search for long segments (> 2-3 centiMorgans, cM) essentially identical (with allowance for recent mutations or sequencing error) between pairs of individuals as in the software GERMLINE [64]. Another method is to model the haplotype frequencies and select haplotypes that are low in likelihood as in the software *FastIBD* or *RefinedIBD* [65, 66]. IBD can be used to infer history, because the shared ancestry amongst a set of groups can only fit certain historical models [67]. In addition, the mutations on the IBD segments and the breakdown of IBD over time (via recombination) can be used to measure the amount of time since populations split or mixed and the direction of admixture, which is whether the ancestry primarily went from one group into the other or vice versa. Lastly, IBD can be used to infer population sizes over time, because smaller population sizes will lead to higher IBD shared within the group. If there was a population bottleneck (a large decrease in population size) or founder event (when a small group of a population separates and founds a new population) within the past 100 generations, this would lead the group to have a high amount of shared within-group IBD even if the group is now large [68, 69]. The resolution of the inferences will vary based on if there is high coverage whole-genome sequencing data or SNP array data [70]. Populations with high within-group IBD, such as South Asians, Finns, and Ashkenazi Jews, often have higher

rates of certain recessive genetic diseases, because it is more likely for the recessive mutations to appear in homozygous form in these groups [71, 72] (Chapter 2). This is a related, but different phenomenon than recent inbreeding, where closely related individuals mate and their children have a high rate of homozygosity (though with much larger DNA segments than in the founder event case), such as in areas like Pakistan [73].

*Rare Variant and Site Frequency Spectrum (SFS)-Based Methods*

Another method with significant power for inferring population history makes use of the joint site frequency spectrum (SFS), which is the frequency of each allele at a particular site in the set of individuals being examined (the histogram of number of alleles at each site). The observed SFS from the data can be fit to the expected SFS from different models of human history, and the models with the highest likelihood can be chosen. This will allow one to determine not only the phylogenetic admixture graph but also the dates of the admixture events and split times as well as population size changes (e.g. exponential growth). It has been claimed that the SFS lacks identifiability (i.e. there is no injectivity; one SFS can fit multiple models of history) [74], but it has been counter-argued that under reasonable conditions and large enough sample size, the SFS can indeed be injective [75]. The mapping from demographic model to expected SFS can be calculated forward in time using Wright-Fisher diffusion equations [76, 77] or a Moran model [78], or backwards in time using the coalescent [79]. The coalescent can also be used to infer the history of population size, such as in PSMC (Pairwise Sequentially Markovian Coalescent), which uses a Hidden Markov Model on the distribution of heterozygous sites in the genome to reconstruct population size histories back to 120,000 BP (years before present, where present is defined as 1950, following standard practice) [80]. MSMC (Multiple Sequentially Markovian Coalescent) and smc++ [81] extend this and use multiple individuals to obtain both population size histories and split times.

Another SFS method, RareCoal [82, 83], restricts analysis to only rare variants (those very infrequent in the population). This software analyzes genetic variants with less than 1% allele frequency (in a relevant reference panel) that are shared between the different groups. The allele sharing is then fit to admixture graphs in a maximum likelihood model.

*Assorted Other Methods*

Approximate Bayesian Computation (ABC) is a general method that can be used for many purposes. This method obtains Bayesian probability distributions of different parameters by performing many simulations of different models of population history, analyzing the resulting summary statistics that would result from such models, and rejecting the models if they are too far off from the summary statistics of the actual data [84]. In population genetics, ABC can be used with a variety of different summary statistics including versions of the methods mentioned previously (e.g. IBD) [85]. Deep learning can be used in a similar manner, with the summary statistics fed into neural networks to learn different demographic parameters of interest, including natural selection [86].

*Integrating Archaeological and Linguistic Information to Determine History*

The analytical methods presented above are not comprehensive, but they summarize many of the major methods used by population geneticists to infer demographic history. However, this information must be integrated with archaeological context and linguistic data where possible. Genetic studies can provide clear evidence for migration, which can provide complementary data to archaeological evidence for the movement of material or cultural goods. For example, several studies documented the genetic history of Europe as beginning with early hunter-gatherers, some of whom replaced prior groups [87], followed by ancestry from Anatolian farmers, followed by the spread of ancestry related to Yamnaya pastoralists from the Pontic-Caspian Steppe (who themselves were a mixture of Eastern European hunter-gatherers and Near Eastern ancestry) [17, 88]. These insights were obtained by studying DNA of individuals from different archaeological cultures and modeling their ancestry, if possible, as a mixture of ancestry of ancient individuals from prior cultures. Moreover, their ancestry can be used to model the ancestry of individuals more recent in time. By doing this procedure and comparing to the spread of archaeological material, one can determine whether the spread of different cultural or technological traits might have been due to the spread of people themselves or just the ideas. For example, the spread of farming from Iran to the Levant has been shown to have occurred independent of the spread of people [54]. In contrast, the Bell Beaker phenomenon in Europe was accompanied by a large migration of Steppe-related ancestry [89]. Linguistic information is generally only available for cultures more recent than 5,000 BP, but the spread of different language families can be correlated to genetics, as for example in the Austronesian expansion, where groups that speak Austronesian languages all have ancestry related to aboriginal Taiwanese [90]. Similarly, the speakers of Indo-Iranian and Balto-Slavic languages both share the same ancestry from the Steppe, providing evidence that people with Steppe ancestry spread these related languages [60]. On the other hand, Melanesian ancestry mostly replaced the Austronesian ancestry of Solomon Islanders even though this group still speaks an Austronesian language [91, 92], showing the need for care in interpreting the relationships between genetics and language family. In addition, archaeological cultures are often not genetically homogeneous, so they should only be grouped together if their genetic homogeneity can be shown. Indeed, migrants with ancestral ancient South Indian ancestry, likely from the Indus Valley civilization, were found in Iran as outliers to the genetic profile in that region, showing cultural contact between the two regions [60]. Similarly, migrations can occur within an individual's lifetime, and strontium isotope data can be used to provide evidence of this if the individual's strontium profile differs from that of the region they were found in (the strontium would have been integrated into their body during their lifetimes, so if the profile matches that of another area, it would provide evidence that the individual migrated from that region) [93, 94]. Such migrations could be indicative of particular cultural practices, thus providing relevant anthropological insight into the culture. For example in Bronze Age Europe, it was shown via genetics and isotopic analyses that high-status

households (determined archaeologically by burials) stayed in the same region, while low-status, genetically unrelated individuals came into the area over time [95].

These examples mentioned all show how genetics can be integrated with archaeology and linguistics to make inferences about human history. The work in this thesis related to inference of demographic history focuses on South Asia and the Americas (particularly South America). Many of the methods described above were used, including data from archaeology and linguistics, to provide new insights into human history in these regions.

**References:**


1. Gilissen C, Hoischen A, Brunner HG, Veltman JA: **Unlocking Mendelian disease using exome sequencing.** *Genome biology* 2011, **12:**228.
2. Posey JE: **Genome sequencing and implications for rare disorders.** *Orphanet journal of rare diseases* 2019, **14:**153.
3. Frésard L, Montgomery SB: **Diagnosing rare diseases after the exome.** *Molecular Case Studies* 2018, **4:**a003392.
4. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D: **Benefits and limitations of genome-wide association studies.** *Nature Reviews Genetics* 2019, **20:**467-484.
5. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J: **10 years of GWAS discovery: biology, function, and translation.** *The American Journal of Human Genetics* 2017, **101:**5-22.
6. Cappellini E, Welker F, Pandolfi L, Ramos-Madrigal J, Samodova D, Rüther PL, Fotakis AK, Lyon D, Moreno-Mayar JV, Bukhsianidze M: **Early Pleistocene enamel proteome from Dmanisi resolves Stephanorhinus phylogeny.** *Nature* 2019, **574:**103-107.
7. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E: **Tracing the peopling of the world through genomics.** *Nature* 2017, **541:**302-310.
8. Rohland N, Hofreiter M: **Ancient DNA extraction from bones and teeth.** *Nature protocols* 2007, **2:**1756.
9. Hansen HB, Damgaard PB, Margaryan A, Stenderup J, Lynnerup N, Willerslev E, Allentoft ME: **Comparing ancient DNA preservation in petrous bone and tooth cementum.** *PloS one* 2017, **12**.
10. Frisch T, Sørensen M, Overgaard S, Lind M, Bretlau P: **Volume-referent bone turnover estimated from the interlabel area fraction after sequential labeling.** *Bone* 1998, **22:**677-682.
11. Rohland N, Glocke I, Aximu-Petri A, Meyer M: **Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing.** *Nature protocols* 2018, **13:**2447.
12. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome**

**sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.

13. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D: **Partial uracil–DNA–glycosylase treatment for screening of ancient DNA.** *Philosophical Transactions of the Royal Society B: Biological Sciences* 2015, **370:**20130624.
14. Wales N, Carøe C, Sandoval-Velasco M, Gamba C, Barnett R, Samaniego JA, Ramos Madrigal J, Orlando L, Gilbert MTP: **New insights on single-stranded versus double-stranded DNA library preparation for ancient DNA.** *Biotechniques* 2015, **59:**368-371.
15. Maricic T, Whitten M, Pääbo S: **Multiplexed DNA sequence capture of mitochondrial genomes using PCR products.** *PloS one* 2010, **5**.
16. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al: **An early modern human from Romania with a recent Neanderthal ancestor.** *Nature* 2015, **524:**216-219.
17. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al: **Massive migration from the steppe was a source for Indo-European languages in Europe.** *Nature* 2015, **522:**207-211.
18. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M: **Genome-wide patterns of selection in 230 ancient Eurasians.** *Nature* 2015, **528:**499.
19. Sarno S, Boattini A, Carta M, Ferri G, Alu M, Yao DY, Ciani G, Pettener D, Luiselli D: **An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of sicily and southern Italy.** *PLoS One* 2014, **9**.
20. Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, Rootsi S, Ilumäe A-M, Mägi R, Mitt M: **A recent bottleneck of Y chromosome diversity coincides with a global change in culture.** *Genome research* 2015, **25:**459-466.
21. Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Sayres MAW, Ayub Q, McCarthy SA, Narechania A, Kashin S: **Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences.** *Nature genetics* 2016, **48:**593-599.
22. Chiaroni J, Underhill PA, Cavalli-Sforza LL: **Y chromosome diversity, human expansion, drift, and cultural evolution.** *Proceedings of the National Academy of Sciences* 2009, **106:**20174-20179.
23. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.
24. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N: **BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis.** *PLoS computational biology* 2019, **15:**e1006650.
25. Webster TH, Sayres MAW: **Genomic signatures of sex-biased demography: progress and prospects.** *Current opinion in genetics & development* 2016, **41:**62-71.

26. Malyarchuk B, Derenko M, Grzybowski T, Perkova M, Rogalla U, Vanecek T, Tsybovsky I: **The peopling of Europe from the mitochondrial haplogroup U5 perspective.** *PloS one* 2010, **5**.
27. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, Rohland N, Mallick S, Stewardson K, Kistler L, et al: **Archaeogenomic evidence reveals prehistoric matrilineal dynasty.** *Nat Commun* 2017, **8:**14115.
28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
29. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.
30. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
31. Skoglund P, Sjödin P, Skoglund T, Lascoux M, Jakobsson M: **Investigating population history using temporal genetic differentiation.** *Molecular biology and evolution* 2014, **31:**2516-2527.
32. Novembre J, Stephens M: **Interpreting principal component analyses of spatial population genetic variation.** *Nature genetics* 2008, **40:**646-649.
33. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
34. Maaten Lvd, Hinton G: **Visualizing data using t-SNE.** *Journal of machine learning research* 2008, **9:**2579-2605.
35. McInnes L, Healy J, Melville J: **Umap: Uniform manifold approximation and projection for dimension reduction.** *arXiv preprint arXiv:180203426* 2018.
36. Margaryan A, Lawson D, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy L, Jørsboe E, Ingason A, Pedersen M: **Population genomics of the Viking world.** *bioRxiv* 2019**:**703405.
37. Mardia KV: **Some properties of clasical multi-dimesional scaling.** *Communications in Statistics-Theory and Methods* 1978, **7:**1233-1241.
38. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data.** *Genetics* 2000, **155:**945-959.
39. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
40. Lawson DJ, Van Dorp L, Falush D: **A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots.** *Nature Communications* 2018, **9:**1-11.
41. Mazandu GK, Geza E, Seuneu M, Chimusa ER: **Orienting future trends in local ancestry deconvolution models to optimally decipher admixed individual genome variations.** In *Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations.* IntechOpen; 2019
42. Reich D, Patterson N: **Will admixture mapping work to find disease genes?** *Philosophical Transactions of the Royal Society B: Biological Sciences* 2005, **360:**1605-1607.

43. Baye TM, Wilke RA: **Mapping genes that predict treatment outcome in admixed populations.** *The pharmacogenomics journal* 2010, **10:**465-477.
44. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Molecular biology and evolution* 1987, **4:**406-425.
45. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Systematic Biology* 1971, **20:**406-416.
46. Sokal RR: **A statistical method for evaluating systematic relationships.** *Univ Kansas, Sci Bull* 1958, **38:**1409-1438.
47. Efron B, Halloran E, Holmes S: **Bootstrap confidence levels for phylogenetic trees.** *Proceedings of the National Academy of Sciences* 1996, **93:**13429-13429.
48. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
49. Cavalli-Sforza LL, Edwards AW: **Phylogenetic analysis: models and estimation procedures.** *Evolution* 1967, **21:**550-570.
50. Holsinger KE, Weir BS: **Genetics in geographically structured populations: defining, estimating and interpreting F ST.** *Nature Reviews Genetics* 2009, **10:**639-650.
51. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y: **A draft sequence of the Neandertal genome.** *science* 2010, **328:**710-722.
52. Pickrell J, Pritchard J: **Inference of population splits and mixtures from genome-wide allele frequency data.** *Nature Precedings* 2012**:**1-1.
53. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al: **Reconstructing Native American population history.** *Nature* 2012, **488:**370-374.
54. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K: **Genomic insights into the origin of farming in the ancient Near East.** *Nature* 2016, **536:**419.
55. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O: **The genomic history of southeastern Europe.** *Nature* 2018, **555:**197.
56. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B: **Inferring admixture histories of human populations using linkage disequilibrium.** *Genetics* 2013, **193:**1233-1254.
57. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D: **Ancient west Eurasian ancestry in southern and eastern Africa.** *Proceedings of the National Academy of Sciences* 2014, **111:**2632-2637.
58. Gravel S: **Population genetics models of local ancestry.** *Genetics* 2012, **191:**607-619.
59. Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M: **Dating the age of admixture via wavelet transform analysis of genome-wide data.** *Genome biology* 2011, **12:**R19.

60. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M: **The formation of human populations in South and Central Asia.** *Science* 2019, **365:**eaat7487.
61. Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, Myers S: **A genetic atlas of human admixture history.** *Science* 2014, **343:**747-751.
62. Lawson DJ, Hellenthal G, Myers S, Falush D: **Inference of population structure using dense haplotype data.** *PLoS genetics* 2012, **8**.
63. Steinrücken M, Kamm J, Spence JP, Song YS: **Inference of complex population histories using whole-genome sequences from multiple populations.** *Proceedings of the National Academy of Sciences* 2019, **116:**17115-17120.
64. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: **Whole population, genome-wide mapping of hidden relatedness.** *Genome Res* 2009, **19:**318-326.
65. Browning BL, Browning SR: **A fast, powerful method for detecting identity by descent.** *The American Journal of Human Genetics* 2011, **88:**173-182.
66. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data.** *Genetics* 2013, **194:**459-471.
67. Palamara PF, Lencz T, Darvasi A, Pe'er I: **Length distributions of identity by descent reveal fine-scale demographic history.** *The American Journal of Human Genetics* 2012, **91:**809-822.
68. Bray SM, Mulle JG, Dodd AF, Pulver AE, Wooding S, Warren ST: **Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population.** *Proceedings of the National Academy of Sciences* 2010, **107:**16222-16227.
69. Gauvin H, Moreau C, Lefebvre J-F, Laprise C, Vézina H, Labuda D, Roy-Gagnon M-H: **Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population.** *European Journal of Human Genetics* 2014, **22:**814-821.
70. Browning SR, Browning BL: **Accurate non-parametric estimation of recent effective population size from segments of identity by descent.** *The American Journal of Human Genetics* 2015, **97:**404-418.
71. Guha S, Rosenfeld JA, Malhotra AK, Lee AT, Gregersen PK, Kane JM, Pe'er I, Darvasi A, Lencz T: **Implications for health and disease in the genetic signature of the Ashkenazi Jewish population.** *Genome biology* 2012, **13:**R2.
72. Slatkin M: **A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases.** *The American Journal of Human Genetics* 2004, **75:**282-293.
73. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, Won H-H, Karczewski KJ, O'Donnell-Luria AH, Samocha KE: **Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity.** *Nature* 2017, **544:**235-239.
74. Myers S, Fefferman C, Patterson N: **Can one learn history from the allelic spectrum?** *Theoretical population biology* 2008, **73:**342-348.

75. Bhaskar A, Song YS: **DESCARTES'RULE OF SIGNS AND THE IDENTIFIABILITY OF POPULATION DEMOGRAPHIC MODELS FROM GENOMIC VARIATION DATA.** *Annals of statistics* 2014, **42:**2469.
76. Lukić S, Hey J: **Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion.** *Genetics* 2012, **192:**619-639.
77. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD, Project G: **Demographic history and rare allele sharing among human populations.** *Proceedings of the National Academy of Sciences* 2011, **108:**11983-11988.
78. Kamm J, Terhorst J, Durbin R, Song YS: **Efficiently inferring the demographic history of many populations with allele count data.** *Journal of the American Statistical Association* 2019**:**1-16.
79. Kingman JFC: **The coalescent.** *Stochastic processes and their applications* 1982, **13:**235-248.
80. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences.** *Nature* 2011, **475:**493-496.
81. Terhorst J, Kamm JA, Song YS: **Robust and scalable inference of population history from hundreds of unphased whole genomes.** *Nature genetics* 2017, **49:**303.
82. Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D: **Iron age and Anglo-Saxon genomes from East England reveal British migration history.** *Nature communications* 2016, **7:**1-9.
83. Flegontov P, Altınışık NE, Changmai P, Rohland N, Mallick S, Adamski N, Bolnick DA, Broomandkhoshbacht N, Candilio F, Culleton BJ: **Palaeo-Eskimo genetic ancestry and the peopling of Chukotka and North America.** *Nature* 2019, **570:**236-240.
84. Turner BM, Van Zandt T: **A tutorial on approximate Bayesian computation.** *Journal of Mathematical Psychology* 2012, **56:**69-85.
85. Beaumont MA, Zhang W, Balding DJ: **Approximate Bayesian computation in population genetics.** *Genetics* 2002, **162:**2025-2035.
86. Sheehan S, Song YS: **Deep learning for population genetic inference.** *PLoS computational biology* 2016, **12:**e1004845.
87. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwangler A, Haak W, Meyer M, Mittnik A, et al: **The genetic history of Ice Age Europe.** *Nature* 2016, **534:**200-205.
88. Allentoft ME, Sikora M, Sjögren K-G, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlström T, Vinner L: **Population genomics of bronze age Eurasia.** *Nature* 2015, **522:**167.
89. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A: **The Beaker phenomenon and the genomic transformation of northwest Europe.** *Nature* 2018, **555:**190.

90. Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D: **Reconstructing Austronesian population history in Island Southeast Asia.** *Nat Commun* 2014, **5:**4689.
91. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, Buckley H, Phillip I, Ward GK, Mallick S: **Population turnover in Remote Oceania shortly after initial settlement.** *Current Biology* 2018, **28:**1157-1165. e1157.
92. Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, Shing R, Buckley H, Kinaston R, Walworth M: **Language continuity despite population replacement in Remote Oceania.** *Nature ecology & evolution* 2018, **2:**731-740.
93. Knipper C, Mittnik A, Massy K, Kociumaka C, Kucukkalipci I, Maus M, Wittenborn F, Metz SE, Staskiewicz A, Krause J: **Female exogamy and gene pool diversification at the transition from the Final Neolithic to the Early Bronze Age in central Europe.** *Proceedings of the National Academy of Sciences* 2017, **114:**10083-10088.
94. Haak W, Brandt G, de Jong HN, Meyer C, Ganslmeier R, Heyd V, Hawkesworth C, Pike AW, Meller H, Alt KW: **Ancient DNA, Strontium isotopes, and osteological analyses shed light on social and kinship organization of the Later Stone Age.** *Proceedings of the National Academy of Sciences* 2008, **105:**18226-18231.
95. Mittnik A, Massy K, Knipper C, Wittenborn F, Friedrich R, Pfrengle S, Burri M, Carlichi-Witjes N, Deeg H, Furtwängler A: **Kinship-based social inequality in Bronze Age Europe.** *Science* 2019, **366:**731-734.

# Chapter 1.2: The Ethics of Genetics Research on Ancient and Present-Day Non-Western Individuals

This section is based on the following papers I contributed to during my PhD:

> **Nakatsuka, N.** "The Ethics of Genetics Research on Ancient and Present-Day Non-Western Individuals." In review at *Current Anthropology*.
>
> Claw, K.; Anderson, M.; Begay, R.; Tsosie, K.; Fox, K.; Garrison, N.; **the Summer Internship for INdigenous peoples in Genomics (SING) Consortium**. "A framework for enhancing ethical genomic research with indigenous communities." *Nature Communications*. 27 July 2018. 9(1): 2957.
>
> Bardill, J.; Bader, A.; Garrison, N.; Bolnick, D.; Raff, J.; Walker, A.; Malhi, R.; **the Summer Internship for INdigenous peoples in Genomics (SING) Consortium**. "Advancing the ethics of paleogenomics." *Science: Policy Forum*. 27 Apr. 2018. 360(6387): 384-385.

### Overview

This section will focus on attempting to provide a framework for conducting ethically-responsible genetics research of ancient and present-day individuals with a particular emphasis on non-Western contexts and a focus on South Asian and Native American groups. The term "Western" here refers generally to European and European-American cultural groups, but it is left purposely imprecise due to the variation in these groups and because the ethical framework developed here is applicable to all groups as will be shown later. After providing a background of some past negative interactions of non-Western groups with human genetics researchers, I will attempt to summarize the main issues at play and then outline some principles undergirding the ethics of genetics research in non-Western communities. I will use these principles to show why there might be potential similarities and differences with ethical obligations for genetic research in Europe or European-Americans. The ethical framework detailed here will form the groundwork for the different research procedures used in the studies of this thesis where these principles were applied to the genetic study of ancient and present-day DNA from African-Americans and groups in South Asia, the South American Andes, Patagonia, and other regions in South America.

### Past Negative Experiences of Human Genetics Research in Non-Western Groups

Genetics research in non-Western contexts (as well as in some Western ones) has a long history of ethical transgressions that have made many non-Western communities distrustful of the field. For example, in 1989, scientists began a partnership initiated by the Havasupai tribe with the hopes of discovering some of the etiology underlying their very high prevalence of diabetes. The scientists from Arizona State University collected their blood and sought to find genetic markers correlated with diabetes, but they were unsuccessful. For many years the Havasupai tribe did not hear back from the scientists regarding the results. In 2003, a member of the Havasupai tribe happened to go to a lecture by a graduate student talking about Havasupai genetics. She realized that the student and her professor had used their DNA to study schizophrenia

and the migratory history of the Havasupai without the permission of the tribe. She was devastated to hear this information discussed in public, because she thought the studies about mental disease brought shame to their tribe, and the human origins research negated their traditional origin stories. The tribe filed a lawsuit against ASU, and eventually they were compensated $700,000, but the damage was already done, and the tribe remains wary of scientists to this day [1, 2].

In a similar way, blood samples from the Nuu-chah-nulth were taken to study potential genetic links to arthritis, but when no link was found, the samples were instead used for studies of human origins without the permission of the tribe [3]. The samples were eventually returned, but only after the tribe inquired about the results and realized what happened to their blood samples. In another case, several indigenous groups in South America, such as the Karitiana, accused the Human Genome Diversity Project (HGDP) of biopiracy due to the scientists taking their blood for the benefit of the scientists and not coming back to inform the group about the results or help the group in any way [4-6]. (The current HGDP samples in use, however, were not from the collection that was obtained via criticized methods; see the end of this essay for more details.) Similar concerns were raised after attempts to patent and commercialize the Native Hawaiian genome [7]. In all of these cases, the scientists were either attempting to use the genetic material for purposes not originally consented to by the communities or conducting the research in a way inconsistent with the cultural preferences of the group whose DNA was being studied.

In ancient DNA (aDNA) research, remains of ancient individuals is often housed in museums away from their original locations, sometimes due to past exploitation (e.g. being taken in the past from groups without permission). In some of these cases, the human remains are affiliated to known cultures and can be traced to living descendants, and in the case of Native American remains in the United States, the Native American Graves Protection and Repatriation Act (NAGPRA), first enacted in 1990, requires that research done on these remains involve the relevant tribes, potentially with the goal of repatriating the remains to the most closely affiliated group when one can be identified [8-10]. In other cases, the remains might be labeled “culturally unaffiliated” in which case no existing tribe is legally associated with them. For some of these “culturally unaffiliated” remains, however, Native American tribes currently live on or near the land where the remains were found, and in some cases, these tribes feel culturally connected to the remains or ancient individual. There is currently no consensus on protocol for research on culturally unaffiliated ancient individuals, though there was recently critique of a study on culturally unaffiliated remains from the Chaco Canyon in New Mexico held at the American Museum of Natural History in New York. Several tribes felt connected to the human remains and responded negatively when they were not consulted about the study. In addition, some took issue with the way it was presented [11]. On the other hand, the very large number of tribes that claimed cultural connection to the remains potentially presented a large practical barrier for consultation.

There have been fewer high-profile examples of ethically problematic genetic studies in South Asia, but there are also many areas that must be thought through when performing genetic studies in this cultural setting. For example, there are very few

ethical committees in India [12], so it is difficult to make standardized regulations and translate material into the proper language for each relevant group. Given the history of the caste system and population bottlenecks in India, there is substantial genetic homogeneity within many of the ethnic groups [13, 14], so potential discrimination and stigma can occur with genetic results specific to certain groups, similar to in Native American communities. In addition, due to the poverty in many parts of India, many individuals are not able to act on genetic studies (e.g. prenatal counseling or medical treatment for genetic diseases), which could lead to increased anxiety arising from knowledge gained about increased risk in medical genetics studies [15]. Low socioeconomic status can also contribute to lower education levels, making informed consent more difficult, and there is potential danger for coercion if the study involves access to healthcare the individuals otherwise could not afford. Lastly, if genetic studies lead to information relevant for prenatal testing (e.g. recessive monogenic genetic variants at high rates in particular groups), there are potential concerns with selective abortion, particularly with female abortions [16] but also for birth defects given the current law that bans termination of pregnancy after 20 weeks of gestation except on grounds of maternal health [17]. The genetic information also has the potential in some groups to lead to mistreatment of individuals, and particularly females, found to be a carrier of certain genetic traits (including the possibility of abandonment, divorce, or ostracization from a marriage) [18].

The myriad factors involved with this topic make it difficult to come up with a clear set of guidelines and standards for genetic research, but the principles presented in this essay will hopefully be useful for thinking through approaching genetic research of ancient and present-day individuals, particularly from non-Western contexts.

**Some Major Issues Relevant to Genetics Research in Non-Western Contexts**

There have been many positive cases of genetics research in non-Western groups, but these negative examples were chosen to highlight potential special considerations when performing genetics research in non-Western contexts compared to Western cultural contexts (though these sometimes still apply in Western contexts as well, as will be described below). Below I will present a non-comprehensive list describing some of these issues.

The first relates to **group consent compared to individual consent**. In this essay, individual consent is defined as an individual's permission for something to occur, while group consent is defined as the group providing permission for something, with the group also choosing the decision-making process governing the decision about that permission. In a strict sense, if there is not a system designated by the group for making decisions for its members, then the term group "consent" is not valid [19]. In this case, consultation can still occur [19], but there are more complexities in decision making, as will be described later in the essay. Group consent will often look different for different cultural groups. In most studies of Western cultural groups, group consent means permission given by different governmental or other institutional regulatory bodies whose representatives have been chosen by the people. However, Western cultures generally place very significant value on individual autonomy, so it is often the case that

these groups consider it only important to obtain the informed consent of the individual participating in the study (i.e. they believe the individual's wishes should not be constrained by the wishes of others in the group [20-22]). In contrast, for many non-Western cultural groups (in particular, including many Native American tribes/nations), group consent is considered more important due to the more communal culture, where decision making is often made by group consensus, and individualism is valued to a lesser extent (i.e. the individual's identity is more tied to the group than to her/himself as an individual [23-25]). In these cases, a representative body of leaders (such as a tribal council) often makes decisions for the whole group. Group consent is important in these settings, because the implications of the research can affect the group as a whole to a greater extent than for other communities (particularly in small groups like many Native American tribes). In addition, group consent can be said to have occurred in many cases if the group's views are adequately represented by the government as is the case in many Western cultural contexts. In contrast, there are some groups whose views are not adequately represented in the government often due to the histories of the region (e.g. colonialism, slavery, or oppressive practices against particular groups leading to power inequalities), so the current laws might not reflect the group's perspectives.

As a related issue, consent in many non-Western cultural contexts might require additional genetics education and/or **cultural translation** to truly be considered informed consent. For example, Foster *et al*., when working with Apache, proposed "communal discourses" before the study begins to ensure both that the study aims are communicated in a culturally relevant way and that the community's method of consent is also honored with the potential for the community to be able to integrate their cultural preferences in how the study will be conducted and results reported [26, 27]. There is the potential, however, that this will involve large costs of time and money in attempting to culturally translate the study aims to the community, so discussion and thought should occur early on to ensure that this can be done in a feasible way.

In addition, there is a danger of **potential stigmatization and racism** that can often be higher in non-Western groups relative to most European-ancestry ones. For example, some groups, particularly very small ones, have a higher chance of being stigmatized based on genetic research (e.g. "that's the schizophrenic tribe"). Certain colonial and imperialistic notions of identity can also be perpetuated by genetic studies that contribute to beliefs about "disappearing cultures" doomed to vanish through genetic admixture. Moreover, for many indigenous groups, kinship can involve social, political, and even spiritual relationships rather than only biological/genetic connections [28, 29]. Sometimes new individuals could even be accepted as part of many indigenous groups without any biological relation to the group. However, blood quanta rules were developed by European-Americans originally to limit tribal enrollment often with the eventual goal of eliminating Native American tribes when they were no longer able to meet the threshold [30-32]. Thus, genetic studies could further perpetuate the ability of outsiders to define belonging in particular groups. Furthermore, the potential for stigma and racism can be particularly high given new advances in population genetics of social behavior, which could be applied with negative consequences, as is the risk with new

genome-wide association studies (GWAS) of traits like educational attainment (including even when done solely within Western cultural contexts [33]).

Another issue that is often more relevant in non-Western contexts than in Western contexts is the potential danger for **commodification of the body, culture and sacred traditions**. In many non-Western groups, particularly indigenous communities, blood and biological samples can be deeply meaningful due to different cultural understandings of the body and self. Frank Dukepoo, a Hopi and Laguna geneticist, remarked, "To us, any part of ourselves is sacred. Scientists say it's just DNA. For an Indian [Native American], it is not just DNA, it's part of a person, it is sacred, with deep religious significance. It is part of the essence of a person" [34]. Similarly, in Native Hawaiian culture, there is the belief that all things possess mana, a life force that makes all bodily samples sacred [35]. Due to these beliefs, proper care and respect for the biological samples from present-day or ancient humans is necessary to accord with the cultural values of the relevant groups. This will likely be different in different cases, so it is important to work on a case-by-case basis with each group (including even Western groups where this might be an issue). Sometimes this might bring up irreconcilable differences as Dukepoo points out with the HGDP's process of immortalizing cell lines, saying, "The idea that some of the tissue, part of that person may be immortalized in these cell lines upsets many people because many Indian tribes hold a strong belief that you can't be buried with a portion of you wandering around the earth; you must be buried whole" [36]. In addition, many indigenous groups might want the samples or human remains returned to them after the study (e.g. to be buried in a culturally appropriate way) and these requests should be honored as much as possible [37]. As an example of sensitivities in this area, the Chaco Canyon paper mentioned previously was criticized for the "culturally-insensitive descriptions" of the ancient individuals and "objectifying" terminology (e.g. "cranium 14" and "burial 14") [11].

One issue that is shared between many Western and non-Western groups is the **potential conflicts of the results of genetic studies with different worldviews**. In Western cultures this can take the form of revealing different family histories than expected or conflicts with origin stories (e.g. literalist biblical views of the book of Genesis). Similarly, in non-Western contexts genetic studies can yield conclusions that conflict with origin stories of different cultures, which could harm their self-identity or even potential claims to their land [38]. On the other hand, it is important for the science not to be biased or presented in inaccurate ways to fit a particular pre-conceived narrative – a repeated theme in aDNA studies is that they have revealed features of the past that conflicted with prior expectations both of present-day groups in the area and of scholars. Balancing scientific models with indigenous knowledge can be difficult, but it is vitally important, particularly for groups that have experienced significant past and present oppression.

One key issue underlying many of the other issues is the concept of **sovereignty** and who gets to decide what type of genetic studies are conducted (including what questions are being asked) and how they are conducted. Many non-Western groups (as well as Western ones) believe that they should have more of a say in the studies of not only present-day DNA from their group but also genetic studies of ancient remains in

their historical lands. On the other hand, depending on how old the ancient individuals are, some argue that the human remains belong, in a certain sense, to the world as a whole rather than to the particular community that happens to have lived in the area most recently. Related to this issue is the method by which such a question might be answered. Many indigenous groups believe they are spiritual care-takers of the land they are living on, so even if they might not have originated from a particular area that they live in now and only happened to have moved into that area more recently (e.g. 500 years ago) or if they are not directly genetically related to a particular ancient individual on their land (e.g. due to past population replacement), they still should have sovereignty over any material from that land. On the other hand, what should happen if there are competing claims to ownership of a particular ancient individual? If claims to ownership are based on factors that are not easily measurable (e.g. spiritual connections), how are such claims to be evaluated and adjudicated? Due to these difficulties, it might not be possible to develop standards that are universally applicable, but the different competing interests must be carefully balanced for each case.

**Principles for Ethically Responsible Genetics Research in Non-Western Groups**

Given the issues highlighted above, several principles can be put forth to undergird ethically-responsible genetics research in non-Western contexts. For the purposes of this essay, we can start with four principles widely agreed upon in Western biomedical ethics: non-maleficence, justice, beneficence, and autonomy [39] (where these principles themselves come from is outside the scope of this section). These principles have correlates in many other cultures and can be explained using culturally specific terms, but for the purposes of this section, they will be explained from a Western cultural viewpoint to show that even within this lens, conclusions can be reached that account for the issues described above. At the same time, even within Western culture some of the principles are interpreted in different ways by different people, and this essay will engage to some degree with this diversity of interpretation both within Western culture and in non-Western cultures (particularly for the principles of autonomy and justice).

The principle of **non-maleficence** is often framed in the Hippocratic phrase: “Do no harm.” The difficulty with this principle is that it can be difficult to define harm, particularly if the claims of harm are psychological or cultural in nature. Nevertheless, the wide body of literature on historical trauma and its relation to public health [40, 41] should indicate that one of the ethical responsibilities (balanced amongst the others) of teams carrying out genetic research is to attempt to avoid contributing to the harm of individuals or groups as the groups themselves perceive it. An important mechanism for preventing potential harm is considering the concerns of the communities most impacted by the research. One way to ensure that this occurs is to engage with the relevant communities throughout the process of the research to get their feedback on as much of the study as possible. This potentially could include the choice of which ancient individuals are to be studied, the process by which the DNA is extracted, which scientific questions are asked, and how the results are interpreted and communicated (e.g. the terms used to describe particular communities or ancient individuals) [42, 43].

This of course needs to be balanced with practical concerns about time and resources as well as scientific accuracy, but providing the communities some seat at the table to voice their views is a significant step towards non-maleficence. If a member of the relevant community were leading the study this likely would improve this situation even more. Some examples of this are research on the Pima Indians in Arizona via the Translational Genomics Research Institute [44], research in Alaska Native communities with the SouthCentral Foundation [45], and studies with the Tlingit peoples and the ancient Shuká Káa individual [46].

At the same time, this principle should have a long-term vision such that the integrity of the scientific enterprise is maintained. Results must not be biased such that incorrect information is disseminated for the sake of fitting a particular narrative (this is non-maleficence towards society as a whole and even the group long-term, since there is a reasonable likelihood that the authentic scientific result will be discovered eventually). On the flip side, there are a multitude of implicit and explicit biases that *all* scientists can have that might skew their work, and sometimes the community's perspective could correct or counterbalance these biases.

The particular way in which the scientific results and cultural interpretation are balanced will likely need to be determined on a case-by-case basis, but it could potentially involve writing the scientific paper and a cultural contextualization as a separate document. An extreme example of this might be a case in which a particular genetic discovery could help to foster a civil war between two groups (e.g. showing one of the groups to be more closely related to a particular revered ancient individual or group). On the one hand, hiding the finding would lead to a bias in the science and an obscuring of the truth, but on the other hand, it would seem that withholding the publication of a particular result is less of an evil than putting forth something that would knowingly contribute to great harm. In this particular case, it might be possible to have the true result present somewhere in the paper while attempting to find a way to present the results publicly in a way that would not contribute to the civil war. In a more common example, a group might have a different origin story than is suggested by a genetic study. In these cases, it might be possible, particularly with the help of the community, to write the paper in such a way that the scientific result is clear while still allowing the group to interpret it in a way that can be reconciled with their particular cultural narrative (though, admittedly, this might be difficult in many situations and in some cases not possible at all, but this principle should be adapted to each individual case).

The principle of **justice** can be interpreted in many different ways, but in a restorative justice lens, this principle should concern the restoration and repair of past wrongs. This is relevant in the cases of human remains that might currently be deemed "culturally unaffiliated" but is known to have been taken from graveyards (e.g. Native American cultural sites) without local community permission, frequently after violent expropriation. In these cases, it can be argued that justice should involve some commitment to repairing the damages done and contributing to the betterment of the groups harmed in this process. This could involve working with the groups during the study of the human remains so they benefit and potentially returning the remains if

there is a clear group or set of groups to which the ancient individuals can be returned. One might think it is unfair to expect scientists today to make up for the wrong doings of prior groups and individuals. A detailed treatment of this complex topic is beyond the scope of this essay, but one important consideration is that if a research group is benefitting from past wrongs (e.g. studying human remains previously stolen from particular groups) without any attempt of repairing the situation, it can be argued that they are perpetuating the past wrongs and thus performing a wrong themselves. Another possible claim is that keeping the remains in a university or government museum would help society as a whole (especially compared to burying the material). However, if there are clear groups that should have say over the remains, then justice might mean their views should take precedence as some restitution for past wrongs. Another type of justice, participatory justice, advocates for direct participation of the individuals/groups most impacted by particular decisions [47, 48], a concept which will be discussed in the section on autonomy. Other views of justice, such as retribution for wrong deeds, distribution of goods, or general fairness and equity, will not be discussed here but intersect with many of the other principles.

The principle of **beneficence** can be applied in many different ways. One clear application is the benefit to society of the scientific discoveries derived from this research. The increase in human knowledge is a tangible, public good that, in theory, is accessible for all after publication of the study. Understanding human history is relevant to all people groups, and sometimes the results are eventually applicable towards improving human health (e.g. gene mapping or precision medicine), which could benefit people in additional ways. However, the benefit should ideally be distributed in such a way that it is not biased towards only those with particular education levels or in particular spheres of society. Thus, if possible, it would be helpful to create opportunities for the distribution of the results of the study in a culturally appropriate way (e.g. language translation, education about basic genetics concepts, and communication of the results in an understandable, culturally relevant manner). It can be argued that groups with greater relation to the DNA should have a greater share of the benefits due to the greater amount of shared history, culture, and connection with the individuals whose DNA is being studied. This is often accounted for in Western groups as the knowledge gained from the study is more often accessible (including via media outlets) and potential medical benefits more readily benefit these groups, but for groups that are more isolated from academia and biomedical advances, additional effort should be made to ensure the benefit adequately reaches them.

The principle of **autonomy** is usually interpreted in Western culture on an individual basis to mean that the subjects involved in the study should have freedom and should not be compelled to act in particular ways. This can be applied more easily in cases of DNA from present-day individuals and communities in which they can have say over how their DNA is being used and can choose to no longer allow their DNA to be used in particular ways at any point (similar to a clinical study) [37]. However, even in Western culture there is a concept of autonomy on a group level, such as the rights of different nations to govern themselves, including the people and material within their borders (for a related discussion on group rights in contrast with individual rights see

[49]. Similarly, groups that are more communitarian might choose to make more of their decisions on a group rather than individual level, and in general the group should be given the autonomy to govern themselves in this manner. In addition, if the group is not adequately represented in the government or agencies that made the laws governing the process of DNA collection, then proper group consent should be obtained. This is admittedly difficult, because there are many cases where sub-groups of the larger group are not well represented in the government (e.g. certain groups that are traditionally lower caste in India or African-Americans in the US), and if these groups do not have clear structures for decision-making regarding issues specifically affecting them on a group level, then "group consent", in a strict sense, is not possible, and it will have to be modified, such as perhaps to approval by many community members after consultation and absence of wide-spread disagreement. In these cases, though, it seems reasonable that, at the very least, the researchers should consider not continuing the study if it goes against the wishes of a large fraction of the group. This might be analogous to an advertisement or movie where some individuals of the group might be willing to participate in the ad (e.g. to earn money), but it is perceived negatively by the majority of the group.

In the case of ancient individuals, autonomy could be conceptualized as a basic respect for the skeletal remains as human individuals (though deceased) [42]. Similar to the case of a deceased individual whose will must be determined by her or his closest relatives, the treatment of an ancient individual (and her/his DNA), it could be argued, should be determined at least in part by the closest relatives (if they can be identified). One might claim that after several hundreds or thousands of years, assuming a linear sense of time, communities should no longer be able to exercise claim to particular ancient individuals. However, the strong cultural connection many groups feel towards ancient individuals should be considered given these groups often are, in fact, genetically more closely related to the ancient individuals than other groups might be. Even in cases where this is not true (e.g. a ~12,800 year old individual from Montana was found to be more closely related to present-day indigenous South Americans rather than to all indigenous North Americans [50] studied up until the time of publication), it could be argued that the indigenous groups in the area still should have some proxy decision making power due to their cultural connection and physical proximity (similar to an individual who chooses another person to be their proxy decision maker even if she/he is not a biological relative). This is admittedly vague, since many groups might feel connections to different ancient individuals despite variable levels of evidence to substantiate those claims. However, there are a variety of ways one might assess such claims (detailed more in the Practical Applications section), and if many external sets of people accept the connection as valid, the research group should take the input of the relevant group seriously.

But what about the autonomy of scientists? Should not scientists, even if they are not from the regions of the world they are studying, have full autonomy to study anything and report whatever they discover? In ethical decision making, the various principles must be balanced properly. It is true that scientific autonomy is extremely important for the advancement of human knowledge, but on the other hand, this

autonomy is not an absolute principle that overrides the others. The other principles should provide guidance and sometimes boundaries for the sampling of remains and the framing and distribution of results, particularly in cases of highly uneven power dynamics. Sometimes these principles might slow the scientific process, but oftentimes they might enrich the study, for example by providing previously un-considered questions to a study or a different cultural lens for interpreting the results. The four fields of genetics, archaeology, linguistics, and anthropology must be synthesized, and involving the communities more closely in the work can help to bring the anthropological lens that is often missing from these syntheses (even if it is from the view of a present-day group that might have variable degrees of cultural connectivity to the ancient groups). It is true that due to practical concerns, papers often might be written with a particular focus (e.g. a genetics focused paper might have less synthesis with material from the other disciplines if the connections are less clear or for the sake of brevity, but attempts at syntheses should be made when possible, and adding a different cultural lens to the results could enrich many studies).

**Comparisons to Genetic Studies in Western Contexts**

The four principles of bioethics can help illuminate some of the potential differences between genetic research done in many Western contexts relative to non-Western ones. They can help to answer common questions like: Why is "special" consultation required when studying many Native American groups but not when studying Europeans or European-ancestry Americans? What is the difference between this situation and, for example, a Swedish person studying ancient Spanish individuals or a Chinese-American studying ancient Hungarians? The differences might become clearer in looking at the application of the principles to the particular historical and cultural contexts of these cases. First the case of present-day DNA will be examined, focusing primarily on de-identified DNA, and then the case of aDNA will be explored.

In the case of present-day DNA, most studies involve de-identified DNA, where the individual's identity is not available to the researchers. Even in these cases, individual informed consent is almost always required to obtain the DNA for both Western and non-Western individuals. The individual is informed for what use the DNA is being collected, including potentially more broad consent to allow the DNA to be used for alternative studies later in the future (e.g. even if it was collected originally for one disease, it could be used as a control for another study or for broader population genetics studies). In addition to individual consent, group consent is arguably still important, because a genetic study of one individual can allow geneticists to make claims about the entire group that the individual comes from. However, it can be argued that even in Western cultural contexts a form of group consent (as defined above) occurs, because usually the communities in particular Western countries are represented by their governments (with proper qualifications), so that, for example, a French individual has some say in French governmental policies. The governmental policies regulate the genetic studies, or even if there are no explicit policies, the government at least has the power to make such policies if the people desired them. Thus, the principle of autonomy, both on an individual and group level is more often

fulfilled. The principle of non-maleficence is more often fulfilled, because the government could produce new policies if damage were being done (and often less damage is done, because the cultural assumptions of DNA use are more concordant with the researchers' assumptions about the world). The principle of justice in terms of restorative and reparative practices are less often at play if Europeans are studying their own ancestors (rather than remains sometimes taken without consent from another group, for example under colonial conditions). Lastly, the principle of beneficence often is fulfilled in Europe and European-Americans, because the results of the study are often communicated to the groups most related to the remains (through scientific journals and news outlets), and the biomedical implications more often reach these communities.

Some of these principles may not be fulfilled in the same way for some non-Western groups. For example, North American Native American tribes and indigenous groups in some parts of South America are not adequately represented by the federal government due to the past and continual effects of colonial history [51]. Moreover, many of these groups also place a higher value on group consent. Thus, the principle of autonomy on a group level is not completely fulfilled in these cases (as described in the autonomy section above). The principle of non-maleficence should consider the fact that the cultural assumptions of many non-Western groups often differ to a greater extent from those of many Western scientists, so the studies could potentially harm them (in ways described above). The principle of justice should take into account the past negative experiences of many groups with Western scientists and seek some restoration and reconciliation. Lastly, the principle of beneficence is frequently less fulfilled for non-Western groups because the results are often not communicated to them, and the biomedical implications also often do not reach these communities to the same extent as they do for many Western groups.

In the case of aDNA, individual consent usually does not apply because the deceased individual cannot provide consent. However, should group consent apply here? If so, what should this look like? The case of post-mortem DNA studies is arguably on the spectrum between present-day and ancient DNA and thus might be instructive. Already there is some sense of group consent in many Western contexts, because the countries have the ability to provide policies on the issue. Most European countries have some data protection policies for deceased patients (reviewed here [52]), and in the countries with explicit policies on the issue, typically the relatives have post-mortem control of the individual unless expressly permitted by the person while alive. It can be argued that ethical principles for aDNA research should be similar to these policies, except perhaps with control spread among more people due to the greater age of the ancient individual (sometimes thousands of years old). For this to be parallel, the policies governing the study of the ancient individual would be determined by the group most related to the individual (even while acknowledging the difficulties in deciding what "related" means with the need to balance cultural and biological relationships, as well as whose land the individual was from).

In archaeology and aDNA research, this framework is not usually followed exactly in this way in Europe, because the governmental permissions required for

archaeologists to excavate and transfer the human remains out of each country vary between countries (or do not exist), and studies of ancient individuals generally do not require consultation with present-day groups most related to the ancient individuals. One could argue that this means the current status quo in European archaeology should change to provide more guidelines or regulations. Alternatively, one could argue that in most Western contexts it is clear to many in the government and the general public that such activities are taking place, and it would be possible for such activities to be more regulated if they desired. There is currently no widespread negative response to such activities (e.g. negative responses are not widespread after media coverage of the studies), and present-day European communities generally have their voices heard through their government (though not in all cases), so their lack of regulations on the studies could be argued to be one form of group consent for the process. In Western contexts, the fact that the communities have governmental representation and the potential ability to regulate such studies (as well as the involvement, in many cases, of local archaeologists and sometimes geneticists of the communities in the area) make it such that the principle of autonomy (including weak forms of group consent) are not strongly violated. In addition, the principle of non-maleficence is more often fulfilled, because the cultural assumptions are more concordant with the study of the past through the sampling of the human remains. Furthermore, the principle of justice in terms of restorative and reparative practices are often less at play in many parts of Europe, because the human remains often were not taken from one area to another place without permission. Lastly, the principle of beneficence is frequently fulfilled in Europe and European-Americans, because the results of the study are often communicated to the groups most related to the remains (through scientific journals and news outlets).

Similar to the case with present-day DNA, the four principles are often not fulfilled to the same extent for non-Western groups in aDNA research. The principle of autonomy on a group level might not be completely fulfilled due to the lack of governmental representation to make policies specifically designed for the at-risk group, as noted previously. Even in cases where the group has governmental representation, it can be difficult in some areas of the world where the absence of governmental policies or regulations on the issue might be due to lack of knowledge about the research or lack of infrastructure to regulate such activities. In these cases, one could potentially work with local museums or other organizations as a representative for local views and participation in the project. Connected to this, the principle of non-maleficence might not be fulfilled if the study does harm to the group due to their cultural beliefs (as outlined previously). The principle of justice might be unfulfilled, especially in cases where the bones were taken without permission from certain areas and put in museums or other locations (e.g. amateur excavation of Native American sites to provide materials to sell to museums and private collections). In such cases, restorative justice would suggest the groups related to the ancient individual should have some control over the human remains, including returning it if desired. This has difficulties due to complex histories in many parts of the world where human remains might have been obtained through means considered to be legal at the time though with laws that were

unjust to certain people groups. Each scenario should be taken on a case-by-case basis to pursue justice in the proper way for each context. Lastly, the principle of beneficence is often not fulfilled fully unless extra effort is taken to ensure the results reach the communities most related to the ancient individual.

It must be noted here that it is not always true that all of these four principles are fulfilled even in Europe or European-Americans. For example, if injustices were done in modern history in the collection of the human remains (e.g. remains taken from one country under military occupation and placed in museums of other countries), it might be incumbent upon researchers to help correct this in the process of the study. In addition, if the results of the study are not easily accessible to the communities where the remains are from, then it might be helpful to find ways to disperse the results as a way of benefitting the communities in some way. Lastly, the researchers should be cognizant of potential harms of the results (e.g. political discord) and should find ways to mitigate such damage if possible, while at the same time fulfilling their obligation to faithfully report what the data show and to highlight aspects of the findings that are surprising relative to previous understandings.

**Practical Applications for the Four Principles to Genetic Research**

The principles explained above must be balanced and applied on a case-by-case basis for each genetic study. Below I will attempt to provide some practical suggestions for how these principles might be applied to studies of ancient and present-day DNA. In each study of this thesis, I will show in its respective section how we attempted (sometimes imperfectly) to apply these principles to each situation. The suggestions will be framed chronologically:

Before the study is conducted, the research team should consider: Where are the human remains coming from? Which are the groups most closely related to the individuals whose DNA is being studied? If it is present-day DNA, then it is the groups of the individuals whose DNA is being collected. If it is aDNA, what is the origin of the DNA? Are there groups that exist now that are closely related to the DNA and/or who live on the land where the ancient individuals were found? Relationship to the DNA can be defined culturally and/or biologically, which could be difficult if the present-day individuals in the region are not sequenced or if the ancient individual is equally genetically related to many groups (and, unfortunately, it will intrinsically be difficult because the genetic ancestry of the ancient individual will not be known before the study), so discussions should occur to properly balance the geographic, cultural, and biological relationships. The Northern Native American context should not be forced upon different contexts in other parts of the world, but nevertheless, the NAGPRA procedure for determining cultural affiliation could be instructive and helpful in some cases. The NAGPRA affiliation criteria weigh factors such as oral traditions, geographic proximity, linguistics, ethnographic documentation, historical evidence, material cultural similarities, or other information (including expert opinion) [53, 54]. The research team could potentially use similar criteria to determine the relevant communities, seeking advice from the relevant sources (e.g. archaeologists, anthropologists, linguists, local

museums, or cultural organizations) as needed. Some of the same sources could sometimes also provide contacts for the relevant communities to engage.

Once the relevant group(s) are found, are the views of these groups with regards to DNA sampling adequately expressed in the current policies governing the collection process? If not, these groups should be consulted if possible, and ideally the mechanism of group consent specific to that culture should be followed. If this is not feasible, it should at least be ensured that the group has general agreement with the study (autonomy). As noted previously, there might be difficult cases such as countries where the people related to the ancient individuals are generally well-represented by their government and other institutions but the potential policy makers have little knowledge that such research is taking place so no policies have been made. In these cases, the research team could potentially try to work with representative organizations, such as local museums or cultural organizations, and follow their cultural protocol after explaining the study to them.

As the study is being planned, the research team should think through potential ways the study could produce harm and try to prevent this if possible (non-maleficence). They should also think of potential ways to benefit groups most closely related to the DNA, including distribution of results to the groups (potentially entailing language and cultural translation) or possibly even restoring remains to where they originally belonged (beneficence and justice).

On a practical level, this should entail that at least one individual from the research team act as the connection between the cultural group(s) and the scientific team (ideally, individuals from the groups themselves would be part of the research team). The connecting individual(s) can communicate the aims of the study before it is conducted and then share the group's thoughts with the research team. The groups of interest should be considered as partial stake-holders in the work, somewhat similar to the collaborators on the research team. Just as archaeologists (and sometimes linguists) help frame the research questions for geneticist analysts and then assist in the interpretation of results throughout the study, the relevant present-day groups should have some input into the study if possible, or at least have a chance to give feedback about the study in a way that has a chance to ultimately be taken on board before publication (the amount of input they provide might vary based on their interest, similar to the variable input of different collaborators, some of whom decide to approve publication with no suggested changes to the manuscript).

As the study is underway, the main individuals performing genetic analyses could work directly with the connecting individual(s) to communicate with the groups (sharing results and getting feedback on the study from the groups). If feasible, the analyst(s) could also meet with the groups to communicate with them directly. This is similar to the communication between genetic analysts and the archaeologists that occurs during the studies where genetic results are shared and the archaeologists share their perspectives and help devise potential new analyses based on the archaeological context. As the manuscript is being written and particularly before publication, the perspectives of groups should be incorporated in a similar way to those of other collaborators. This can be done with the recognition that the focus of the paper might

be predominantly oriented on the storyline of the genetics, and separate material can be created that more fully emphasize the cultural perspective (similar to how separate papers are written from a primarily archaeology-driven perspective that use the genetic results to argue for particular historical models). After the study is finished, the research team should ensure that the results are available to the relevant groups. Ideally this could be done through exhibits or public presentations in museums, local schools, cultural organizations, or other culturally appropriate venues. At a minimum researchers should ensure that the scientific findings are accessible to the relevant communities through a publication that is freely and publicly available (not inaccessible behind a journal paywall), and if possible through a translation of the key findings into the languages commonly spoken by the community so that they are benefitting from the study in an equal way.

**Addressing Potential Difficulties**

Below I will attempt to address potential difficulties with this approach by responding to realistic questions.

*1) Will this add substantially more time and money required for each study?*

This is an understandable concern given the need to publish papers in reasonable time frames and meet grant deadlines. At the same time, the effort invested in outreach could yield new insights that would otherwise not be possible without the engagement of groups related to the DNA. The amount of time, effort, and money for the consultation with the relevant groups will depend on each study and the research group. In the case of DNA from present-day individuals, the group consultation would in some cases be expected to be obtained in the first place to get permission from a large number of individuals for the study. If the group is well-represented in their government, then some form of group consent already occurs if there are policies designated by the relevant governing authorities. If there are no policies in place or if the group is not well-represented in the government and do not have a governing body designated by the people for making decisions about these issues, obtaining consent from large numbers of individuals in the population could be considered a form of group consent (assuming a large proportion of the group knows what is happening and is not against the process). The follow-up reporting of results to the relevant groups could be a natural corollary if the initial relationship is strong.

In the case of DNA from ancient individuals, it seems possible that this will not add a substantial amount of extra resources required for the study if there are local archaeologists and anthropologists who can work with the relevant groups, particularly if the ancient individuals were excavated from sites near those groups (i.e. the archaeologists would have already needed to be in that area to excavate the ancient individuals, so working with the present-day group(s) in the area would seem to be a natural fit). If the ancient individuals were obtained from areas far away from the areas they originally came from (e.g. museum collections), then the extra effort should improve the study as it will embed the research in the original location where the ancient individuals lived. For example, several of the ancient Cusco individuals in this

thesis were obtained from museum collections in the US, and they were repatriated to the Cusco region during the study. Lastly, it should be noted that if it is believed that a particular study would cost too much time or money to be done in an ethically proper way, perhaps the team should consider modifications to make the study both feasible and ethical. For example, it seems reasonable to suggest that a similar amount of effort taken to obtain human remains (i.e. finding and "engaging" ancient individuals) should also be used to find and engage with living communities related to those ancient individuals.

*2) What if there are too many groups to consult (in aDNA studies)?*

If a large number of groups exist in a region and all claim connection to particular ancient individuals, it might be difficult on a practical level to consult with all of the groups. The approach will vary depending on the particular case. It is possible that there could be meetings with multiple groups at once (though this might be logistically very challenging), or the priority could potentially be given based on a decision weighing factors such as oral traditions, geographic proximity, linguistics, ethnographic documentation, material cultural similarities, or other factors as is done in NAGPRA [53, 54]. If the groups do not agree with each other, the research team will have to determine how to proceed based on the particular situation. Nevertheless, the difficulty of some of these situations should not be a reason to abandon the ethical framework here any more than conflicts between different archaeologist and geneticist teams should be a reason to abandon aDNA research in any area.

*3) What if the group(s) consulted do not want the study to be conducted?*

In the case of present-day DNA, the answer to this is fairly clear, because it would not be possible to obtain DNA from the living individuals if they did not consent to the study. However, the case for aDNA is less clear given the ancient individuals are sometimes thousands of years removed from the groups most closely related to them, so many argue that these groups should not have veto power over the study of the ancient individuals. However, this section so far has argued that on the basis of the principles of non-maleficence, justice, beneficence, and autonomy, the groups most closely related to the DNA should be contributing members with input in the study in a similar way that archaeologists and/or museum curators contribute to genetic studies, with these individuals representative of the sometimes wide variety of stakeholders in aDNA research that have different goals and occasionally different perspectives on the work.

In the current structure for many archaeologist-geneticist collaborations, it is usually understood that the archaeologists or the curators, in a broad sense, have overall custodianship of the ancient individuals they provide for the study in the sense that the geneticists would not be allowed to pursue the study if the archaeologists did not agree to it, and the archaeologists could in theory decide to withdraw the individuals they provide at any moment (though once the archaeologists agree to pursue a study with the geneticists and the geneticists have invested effort, it is generally considered unprofessional for either side to back out after the study has

begun and more appropriate to work together to find an acceptable common ground on which to publish).

The communities related to the aDNA should be added to this structure in a role analogous to the archaeologist. In the ethical framework presented here, it would seem then that the groups most closely related to the DNA should have authority over the ancient individuals and thus be able to choose which research teams to work with or whether to conduct the study at all (similar to archaeologists choosing which geneticist teams to work with). This produces a shared custodianship of the ancient individuals rather than having it only with the archaeologists [55, 56]. It must be noted that this structure is not always possible, for example if the community would not like to engage due to lack of capacity or interest, but at very least attempts should be made to involve them in this manner.

Both sides should have a balanced perspective, however. If the ancient individuals are many thousands of years old, it is not necessarily clear that a single present-day group should have authority over them (e.g. to prevent a study), because the individuals and the cultural group they were part of could potentially have had different views than the existing present-day group(s) living in the region today (what if they would have wanted the study to have been conducted?). Moreover, due to the age of the individuals, it might be true that they belong in some sense, to the entire world rather than to a particular group. On the other hand, the study has greatest impact on these present-day groups, and doing the study against their wishes might have significant psychological and cultural harm on the group. In addition, why should the archaeologists have sole management control over the ancient individuals that they excavated, especially if it is not their land? The philosophy of "finders, keepers" was often used in the Western colonization of many lands with devastating consequences; the aDNA field should be cognizant of past mistakes and avoid the perpetuation of a colonial mindset. Thus, at very least it would seem reasonable to give the groups most closely related to the DNA (potentially via their representatives) some management control over the ancient individuals if they choose to become engaged as partners in the work (similar to the oversight archaeologists often currently have over many of the ancient individuals they excavate). In both cases, though, the groups and the archaeologist do not "own" the individuals; they are simply temporary stewards caring for them.

*4) What if the group(s) consulted have differing opinions either within their own group or between different groups?*

Most groups are not homogeneous, so even within a single group there might be differing opinions about a particular study. If multiple groups need to be consulted there might also be differing opinions between the different groups. In these cases, the balance of individual vs. group autonomy should be done by the group(s) themselves, so they can come to a decision in their culturally specific way. This is analogous to governmental permissions for studies where the policies are determined based on discussions between many individuals and the final decision might not match everyone's preferences but everyone still needs to follow the final guidelines. This will admittedly

be difficult for the cases of multiple groups that might not have a previously established mechanism to come to conclusions among the group members, but in these cases the research team should still do their best to allow the groups themselves to come to a decision through their own culturally appropriate mechanisms with enough time for the discussions.

But what if you are an individual inside or outside of the group that disagrees with the group's final decision? Do you have to follow it? One could argue this is analogous to situations where a citizen or external individual might disagree with particular policies of a government (e.g. tax law, social policies, foreign affairs). There are particular times when it might be considered ethical to violate a policy in the name of a higher principle(s). However, the principles of non-maleficence and group autonomy are such that it is usually better to follow the decision made by the group's designated decision-making process, since side-stepping this could de-legitimize that group's governing process. At the same time, individuals, especially those within the group but also those external to the group, could seek to persuade the group to change a particular policy.

It must be noted, however, that there are many ethical and philosophical complications of this framework for both present-day and ancient DNA studies. For example, in present-day DNA studies, what if one individual from the group wants her/his genome studied, but the group as a whole does not want the study? In the analogous example where an individual disagrees with his/her group's policies, the individual is often free to leave the group and revoke his/her group identity (e.g. citizenship). However, in the case of genetics, it is often not possible for individuals to escape the fact that their genomes still have information relevant for their group. These cases are complex, but at the moment the framework presented here would seem to imply that the research team has an ethical responsibility at least to withhold the group identity of the individual in the study.

In the case of aDNA, the possibility of differences in opinion between the individual and group is not present, but in this case, the age of the individuals (sometimes thousands of years old) presents another philosophical complication. The temporal gap is such that some might argue the decision-making power should be shared more broadly (including potentially to people less related to the ancient individuals). On the other hand, I have argued above that some priority in the decision-making should still be given to the groups most related to the ancient individual due to the principles of non-maleficence, autonomy, and participatory justice.

*5) What if the results of the study contradict the cultural narrative of the present-day groups?*

One of the major fears of many present-day indigenous groups is that the genetic studies will contradict their cultural narrative, such as their origin stories or other traditional knowledge. There is fear that this will diminish their traditional ways of knowing and thus harm the cultural identity of the group. Related to this are often fears that such studies will harm their claims to land ownership, because the studies might show that their ancestors did not originate from the place the group is residing now

(e.g. Out of Africa narrative for non-Africans today, Bering Land Bridge narrative for people living in the Americas today, or archaeology of Europe for Brexit arguments [57]). It is important for these issues to be communicated up front with the group early on as potential implications of the study.

At the same time, these discussions should promote ways for the group to consider how they might be able to integrate Western science with their cultural practices and traditional ways of knowing. If the discussions are framed properly without denigrating the group's traditional beliefs, they could lead to productive ways to promote cultural pride, traditional practices, and science education amongst their group as well as potentially sharing how their cultural perspectives should influence others, including the beliefs and practices of those in Western cultures. It is very important for the scientific findings not to be biased to produce a particular narrative, because this will produce long-term harm to the scientific enterprise, which is based on reproducibility and consistency of results. The most plausible scientific model will usually be discovered eventually, so trying to force the scientific results to say something the data do not support can be counter-productive in the long-term. On the other hand, the present-day groups do not need to view the world primarily through the lens of the scientific narrative. There might be ways for them to integrate the scientific results into their cultural perspective while still promoting their traditional ways of knowing and their cultural outlook. The discussions with the group could promote this as a possibility and allow the group to determine how they might want to integrate the genetic study with their particular culture.

During the discussions, it should also be communicated that a global perspective on population genetics should not provide grounds against claims to land ownership, because such a global perspective would reveal that ultimately all groups are immigrants to all lands. Individuals of French ancestry might be genetically related in part to early hunter-gatherers from the region, but even these individuals migrated from elsewhere (and the same can even be said of all groups in Africa based on past migrations within Africa). This "immigration" status of all people is clearly not grounds for anyone to then take away land. It is only a deceptive, selective use of population genetics that could lead to such conclusions.

*6) What about the potential for increased bias if members study DNA related to their own communities?*

Some argue that to prevent biases genetic studies should be done by "third-parties" who have more cultural and genetic distance from the group they are studying (e.g. having a Gambian scientist study ancient Swedish genomes or a Japanese scientist study West African genetic variation rather than having French individuals study French or Spanish genetics). Proponents of this would argue that although scientists with greater cultural proximity to the group they are studying have more background knowledge of that group, they also have more biases (e.g. from potentially unsubstantiated history they have learned when growing up). In response to this it should be noted that all people have some biases, but the peer review process and the

scientific community as a whole should hopefully be able to correct possible errors or claims in publications that are not substantiated with the data.

*7) What if the community is unable to understand the goals, methods, or results of the study?*

In cases where there are difficulties with language or cultural translation, particularly with groups with very limited formal Western education (e.g. Onge) where explaining concepts of genetics and DNA might be difficult, there might be other ways to communicate in terms the group can understand (for example, art, songs, or genealogies connecting people). The approach will need to be specific for each case, but adequate cultural and language translation is necessary for proper informed consent to ensure that the group has true autonomy in their decision of whether or not to participate in the study, and if truly informed consent is not possible, the study should not proceed.

*8) Should studies be conducted with previously published data derived from samples that were already collected in what some might consider to be unethical ways?*

This is a complex issue, which should be taken on a case by case basis. Researchers should recognize that the use of such data could sometimes be further damaging to those who were negatively impacted. In an extreme case, data and results from old Nazi studies are rarely ever used due their potential to elicit trauma from memories of past evil [58-60]. In a case more similar to the ones described in this essay, the continual use of HeLa cells without the permission of the Lacks family has been criticized even with the acknowledgment that the cell lines have produced enormous benefit for the world (see context here [61]). Nevertheless, there are many complex cases, because in some past studies the investigators might have been following the expectations and guidelines of their time, yet at the present their approach might be viewed as having not fulfilled all ethical principles in the most ideal way possible. In addition, many investigators have different views on the ethics of past studies and whether (and to what extent) some studies were ethically problematic.

There are a variety of different approaches to these cases. If an individual or group considers a past study fully consistent with their ethical principles (e.g. the view of some with regards to the Chaco Canyon study [62]), they often will make full use of the data even if others believe the study was unethical. In some cases, however, they might choose to not use the data purely for the sake of not offending those who view the study as unethical. Along the spectrum, some might have different degrees of reservations with various studies conducted in the past. In these cases, some decide not to cite or mention any studies they believe were done unethically, while others decide to cite the studies if their existence is relevant but not use the data for primary analyses (the studies in this thesis usually followed this approach).

Even with the principles and ethical framework presented in this section, there likely will be many gray areas due to the need to balance the different principles and the different ethical standards at different times in the past. For example, in the case of the Karitiana and Surui DNA sample collection, some of the contacts were more negative

with promises unfulfilled and no further communication, while others contacts (including the ones that led to the actual samples that turned into immortalized cell lines) were done with full informed consent with no evidence for a deviation from ethical standards. In this case, the first collection occurred in 1987 by geneticist Francis Black with full informed consent and were stocked and sold by the Coriell Cell Repositories [63]. In contrast, a second set in 1996 was collected from the Karitiana with the accompaniment of a British television crew along with promises of significant health benefits. The Karitiana denounced the second team and requested compensation due to their failed promises. Much of the data from the set with full informed consent were unified in a sample panel that is widely used by researchers across the world today [64]. The data currently being used in this thesis are only from this panel, meaning they are being used in a manner fully consistent with the original intent of the agreement of the researchers and the different groups. In all cases, though, the investigators should be careful about how the data was originally generated and whether their proposed study is using the data in a manner consistent with the original intent of the data collection. They should also consider whether, as in the case of the Karitiana, perspectives of the group involved may have changed over time and differed from original views.

*9) How should this ethical framework be regulated (for aDNA studies)?*

Governmental regulations like IRBs have been useful in guiding biomedical research, particularly for human studies. They provide constraints on studies to prevent research deemed to be unethical, and they often help to guide scientists in how they should think about the ethical underpinnings of their work. On the other hand, they add a significant amount of time and effort to studies. More importantly, it can be argued that they are not flexible enough to guide population genetics research through the widely varied contexts of the many cultures in this world. This is a crucial point, because the enormous differences in cultural contexts are such that it would be very difficult to create (and enforce) governmental policies that fit all cases. In addition, it could even potentially cause harm if the IRB committee does not understand the particular culture and they attempt to impose their own standards to the group in a way that does not fit its unique historical-cultural situation.

There are other potential ways to create accountability for aDNA studies. For example, journals could require investigators to fill out a form detailing their ethical framework in the study, including to what extent local communities were involved. This could also include a statement about permissions and consent and a rationale for how individual and group consent and consultation was conceived and whether (and how) either of them should apply to the study. This statement could potentially be published with the manuscript so that the broader community, including geneticists and archaeologists, could comment on the approach of the investigators. Many recent publications on ancient Native Americans (including the 3 in this thesis) opted to include an ethics statement in the manuscript [46, 50, 65-68].

**Concluding Thoughts**

The framework presented here is only one voice amongst many who have studied and thought about these issues in depth. Nevertheless, it is hopefully still a useful contribution in producing research that is ethically robust and beneficial towards all. Even if there are disagreements with my particular application of the ethical principles, this section was written to advance the dialogue so that all sides can work together towards common goals. As the field changes, the dialogue will likely also advance and change, but this section reflects my current thinking and that of potentially many others, and it forms the basis for the studies that will be presented in the rest of this thesis. At the beginning of each section of this thesis, I will highlight how our studies hopefully are consistent with the framework outlined here (albeit imperfectly in some cases).

**Acknowledgments:**

I thank Jeantine Lunshof, Mary Prendergast, Natalie Kofler, Jakob Sedig, Iosif Lazaridis, Kendra Sirak, and David Reich for their comments on this essay. I would also like to thank members of the SING (Summer Internship for Indigenous Peoples) consortium and the Reich lab for helpful discussions on these issues. Any errors or mis-statements, however, are my responsibility.

**References:**

1. Pacheco CM, Daley SM, Brown T, Filippi M, Greiner KA, Daley CM: **Moving forward: breaking the cycle of mistrust between American Indians and researchers.** *American journal of public health* 2013, **103:**2152-2159.
2. Garrison NA: **Genomic justice for Native Americans: impact of the Havasupai case on genetic research.** *Science, Technology, & Human Values* 2013, **38:**201-223.
3. Dalton R: **Tribe blasts' exploitation'of blood samples.** Nature Publishing Group; 2002.
4. Awang SS: **Indigenous nations and the human genome diversity project.** *Indigenous knowledges in global contexts: Multiple readings of our world* 2000**:**120-136.
5. Cunningham H: **Colonial encounters in postcolonial contexts: Patenting indigenous DNA and the Human Genome Diversity Project.** *Critique of Anthropology* 1998, **18:**205-233.
6. Hasian J, Marouf, Plec E: **The cultural, legal, and scientific arguments in the human genome diversity debate.** *Howard Journal of Communication* 2002, **13:**301-319.
7. Singeo L: **The patentability of the Native Hawaiian genome.** *American journal of law & medicine* 2007, **33:**119-139.

8. Rose JC, Green TJ, Green VD: **NAGPRA is Forever: Osteology and the Repatriation of Skeletons.** *Annual Review of Anthropology* 1996, **25:**81-103.
9. Fine-Dare KS: *Grave injustice: the American Indian repatriation movement and NAGPRA.* U of Nebraska Press; 2002.
10. Liebmann M: **Postcolonial cultural affiliation: essentialism, hybridity, and NAGPRA.** *Archaeology and the postcolonial critique* 2008**:**73-90.
11. Claw KG, Lippert D, Bardill J, Cordova A, Fox K, Yracheta JM, Bader AC, Bolnick DA, Malhi RS, TallBear K: **Chaco Canyon Dig Unearths Ethical Concerns.** *Human biology* 2017, **89:**177.
12. Mathaiyan J, Chandrasekaran A, Davis S: **Ethics of genomic research.** *Perspectives in clinical research* 2013, **4:**100.
13. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
14. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S: **The promise of discovering population-specific disease-associated genes in South Asia.** *Nature genetics* 2017, **49:**1403.
15. Mahajan S: **The challenges of genetic research in India.** *Issues in Medical Ethics* 2002, **10:**143-145.
16. Murthi M, Guio A-C, Dreze J: **Mortality, fertility, and gender bias in India: A district-level analysis.** *Population and development review* 1995**:**745-782.
17. Aggarwal S, Phadke SR: **Medical genetics and genomic medicine in India: current status and opportunities ahead.** *Molecular genetics & genomic medicine* 2015, **3:**160.
18. Sanneving L, Trygg N, Saxena D, Mavalankar D, Thomsen S: **Inequity in India: the case of maternal and reproductive health.** *Global health action* 2013, **6:**19145.
19. Dickert N, Sugarman J: **Ethical goals of community consultation in research.** *American journal of public health* 2005, **95:**1123-1127.
20. Darwish A-FE, Huber GL: **Individualism vs collectivism in different cultures: a cross-cultural study.** *Intercultural Education* 2003, **14:**47-56.
21. Fiske AP, Kitayama S, Markus HR, Nisbett RE: **The cultural matrix of social psychology.** 1998.
22. Kim U: *Individualism and collectivism: A psychological, cultural and ecological analysis.* NIAS Press; 1995.
23. McInerney DM, Ali J: **Indigenous motivational profiles: do they reflect collectivism?: a cross-cultural analysis of similarities and differences between groups classified as individualist and collectivist cultures.** *Indigenous peoples* 2013**:**211-232.
24. Hossain Z, Skurky T, Joe J, Hunt T: **The sense of collectivism and individualism among husbands and wives in traditional and bi-cultural Navajo families on the Navajo Reservation.** *Journal of Comparative Family Studies* 2011, **42:**543-562.
25. Podsiadlowski A, Fox S: **Collectivist value orientations among four ethnic groups: Collectivism in the New Zealand context.** *New Zealand Journal of Psychology* 2011, **40:**5-18.

26. Foster MW, Bernsten D, Carter TH: **A model agreement for genetic research in socially identifiable populations.** *The American Journal of Human Genetics* 1998, **63:**696-702.
27. Foster MW, Sharp RR, Freeman WL, Chino M, Bernsten D, Carter TH: **The role of community review in evaluating the risks of human genetic variation research.** *The American Journal of Human Genetics* 1999, **64:**1719-1727.
28. Alfred T, Corntassel J: **Being Indigenous: Resurgences against contemporary colonialism.** *Government and opposition* 2005, **40:**597-614.
29. Dudgeon P, Bray A: **Indigenous Relationality: Women, Kinship and the Law.** *Genealogy* 2019, **3:**23.
30. TallBear K: **Native-American-DNA. com.** *Revisiting race in a genomic age* 2008, **235**.
31. Schmidt RW: **American Indian identity and blood quantum in the 21st century: a critical review.** *Journal of Anthropology* 2012, **2011**.
32. Thornton R: **Tribal membership requirements and the demography of 'old'and 'new'Native Americans.** *Population Research and Policy Review* 1997, **16:**33-42.
33. Adam D: **The promise and peril of the new science of social genomics.** *Natur* 2019, **574:**618-620.
34. John JS: *Native American Scientists.* Children's Press; 2000.
35. Blaisdell K: **Historical and cultural aspects of Native Hawaiian health.** *Social Process in Hawaii* 1989, **32:**1-21.
36. Dukepoo F: **Sensitivities and concerns of research in native American communities.** https://bioethicsarchive.georgetown.edu/nbac/transcripts/jul98/native.html; 1998.
37. Arbour L, Cook D: **DNA on loan: issues to consider when carrying out genetic research with aboriginal families and communities.** *Public Health Genomics* 2006, **9:**153-160.
38. Ilkilic I, Paul NW: **Ethical aspects of genome diversity research: genome research into cultural diversity or cultural diversity in genome research?** *Medicine, Health Care and Philosophy* 2009, **12:**25-34.
39. Childress JF, Beauchamp TL: *Principles of biomedical ethics.* Oxford University Press New York; 2001.
40. Mohatt NV, Thompson AB, Thai ND, Tebes JK: **Historical trauma as public narrative: A conceptual review of how history impacts present-day health.** *Social Science & Medicine* 2014, **106:**128-136.
41. Gone JP, Hartmann WE, Pomerville A, Wendt DC, Klem SH, Burrage RL: **The impact of historical trauma on health outcomes for indigenous populations in the USA and Canada: A systematic review.** *American Psychologist* 2019, **74:**20.
42. Bardill J, Bader AC, Garrison NA, Bolnick DA, Raff JA, Walker A, Malhi RS, Summer internship for IpiGC: **Advancing the ethics of paleogenomics.** *Science* 2018, **360:**384-385.

43. Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Nanibaa'A G: **A framework for enhancing ethical genomic research with Indigenous communities.** *Nature communications* 2018, **9:**2957.
44. Smith-Morris CM: **Reducing diabetes in Indian country: lessons from the three domains influencing Pima diabetes.** *Human Organization* 2004**:**34-46.
45. Dirks LG, Shaw JL, Hiratsuka VY, Beans JA, Kelly JJ, Dillard DA: **Perspectives on communication and engagement with regard to collecting biospecimens and family health histories for cancer research in a rural Alaska Native community.** *Journal of community genetics* 2019**:**1-12.
46. Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, Mitchell J, Worl R, Dixon EJ, Fifield TE, et al: **Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity.** *Proc Natl Acad Sci U S A* 2017, **114:**4093-4098.
47. Stein MA, Lord JE: **Jacobus tenBroek, participatory justice, and the UN Convention on the Rights of Persons with Disabilities.** *Tex J on CL & CR* 2007, **13:**167.
48. Honeyman C, Hudani S, Tiruneh A, Hierta J, Chirayath L, Iliff A, Meierhenrich J: **Establishing collective norms: Potentials for participatory justice in Rwanda.** *Peace and Conflict* 2004, **10:**1-24.
49. Jones P: **Group rights.** *Stanford Encyclopedia of Philosophy* 2016.
50. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al: **The genome of a Late Pleistocene human from a Clovis burial site in western Montana.** *Nature* 2014, **506:**225-229.
51. Starks GL: **Minority representation in senior positions in US federal agencies: A paradox of underrepresentation.** *Public Personnel Management* 2009, **38:**79-90.
52. Bak MA, Ploem MC, Ateşyürek H, Blom MT, Tan HL, Willems DL: **Stakeholders' perspectives on the post-mortem use of genetic and health-related data for research: a systematic review.** *European Journal of Human Genetics* 2019**:**1-14.
53. Schillaci MA, Bustard WJ: **Controversy and conflict: NAGPRA and the role of biological anthropology in determining cultural affiliation.** *PoLAR: Political and Legal Anthropology Review* 2010, **33:**352-373.
54. Kuprecht K: **The concept of "cultural affiliation" in NAGPRA: its potential and limits in the global protection of indigenous cultural property rights.** *International Journal of Cultural Property* 2012, **19:**33-63.
55. Nicholas G, Bannister K, Brown M, Hamilakis Y, Ouzman S, Vitelli K, Nicholas G, Bannister K: **Copyrighting the past? Emerging intellectual property rights issues in archaeology.** *Current Anthropology* 2004, **45:**327-350.
56. Watkins J: *Indigenous archaeology: American Indian values and scientific practice.* Rowman & Littlefield; 2000.
57. Brophy K: **The Brexit hypothesis and prehistory.** *antiquity* 2018, **92:**1650-1658.
58. Cohen Jr MM: **Overview of German, Nazi, and holocaust medicine.** *American Journal of Medical Genetics Part A* 2010, **152:**687-707.

59. Oehler-Klein S, Preuss D, Roelcke V: **The use of executed Nazi victims in anatomy: Findings from the Institute of Anatomy at Giessen University, pre- and post-1945.** *Annals of Anatomy-Anatomischer Anzeiger* 2012, **194:**293-297.
60. Mills J: **Pandora's box closed: The Royal Air Force Institute of Aviation Medicine and Nazi medical experiments on human beings during World War II.** *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 2020, **79:**101190.
61. Beskow LM: **Lessons from HeLa cells: the ethics and policy of biospecimens.** *Annual review of genomics and human genetics* 2016, **17:**395-417.
62. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, Rohland N, Mallick S, Stewardson K, Kistler L, et al: **Archaeogenomic evidence reveals prehistoric matrilineal dynasty.** *Nat Commun* 2017, **8:**14115.
63. **Karitiana: Biopiracy and the Unauthorized Collection of Biomedical Samples** [https://pib.socioambiental.org/en/povo/karitiana/389]
64. Fullerton SM, Lee SS: **Secondary uses and the governance of de-identified data: lessons from the human genome diversity panel.** *BMC medical ethics* 2011, **12:**16.
65. Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Morseburg A, Johnson JR, Potter A, et al: **Ancient human parallel lineages within North America contributed to a coastal expansion.** *Science* 2018, **360:**1024-1027.
66. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T: **Early human dispersals within the Americas.** *Science* 2018**:**eaav2621.
67. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E: **Reconstructing the deep population history of Central and South America.** *Cell* 2018, **175:**1185-1197. e1122.
68. Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS: **A time transect of exomes from a Native American population before and after European contact.** *Nature communications* 2016, **7:**1-11.

# Chapter 2: Using modern DNA to explore the history of South Asia with relevance for gene mapping

This section is based on the following papers I contributed to during my PhD:

Narasimhan, V.*; Patterson, N.*; Moorjani, P.; Rohland, N.; Bernardos, R.; Mallick, S.; Lazaridis, I.; **Nakatsuka, N.;** Olalde, I.; Lipson, M.; Kim, A.; Olivieri, L.; Coppa, A.; Vidale, M.; Mallory, J.; Moiseyev, V.; Kitov, E.; Monge, J.; Adamski, N.; Neel, A.; Broomandkhoshbacht, N.; Candilio, F.; Callan, K.; Cheronet, O.; Culleton, B.; Ferry, M.; Fernandes, D.; Gamarra, B.; Gaudio, D.; Hajdinjak, M.; Harney, E.; Harper, T.; Keating, D.; Lawson, A.; Mah, M.; Mandl, K.; Michel, M.; Novak, M.; Oppenheimer, J.; Rai, N.; Sirak, K.; Slon, V.; Stewardson, K.; Zalzala, F.; Zhang, Z.; Akhatov, G.; Bagashev, A.; Bagnera, A.; Baitanayev, B.; Bendezu-Sarmiento, J.; Bissembaev, A.; Bonora, G.; Chargynov, T.; Chikisheva, T.; Dashkovskiy, P.; Derevianko, A.; Dobes, M.; Douka, K.; Dubova, N.; Duisengali, M.; Enshin, D.; Epimakhov, A.; Freilich, S.; Fribus, A.; Fuller, D.; Goryachev, A.; Gromov, A.; Grushin, S.; Hanks, B.; Judd, M.; Kazizov, E.; Khokhlov, A.; Krygin, A.; Kupriyanova, E.; Kuznetsov, P.; Luiselli, D.; Maksudov, F.; Mamedov. A.; Mamirov, T.; Meiklejohn, C.; Merrett, D.; Micheli, R.; Mochalov, O.; Mustafaokulov, S.; Nayak, A.; Pettener, D.; Potts, R.; Razhev, D.; Rykun, M.; Sarno, S.; Savenkova, T.; Sikhymbaeva, K.; Slepchenko, S.; Soltobaev, O.; Stepanova, N.; Svyatoko, S.; Tabaldiev, K.; Teschler-Nicola, M.; Tishkin, A.; Tkachev, V.; Vasilyev, S.; Velemínsky, P.; Voyakin, D.; Yermolayeva, A.; Zahir, M.; Zubkov, V.; Zubova, A.; Shinde, V.; Lalueza-Fox, C.; Meyer, M.; Anthony, D.; Boivin, N.; Thangaraj, K.; Kennett, D.; Frachetti, M.; Pinhasi, R.; Reich, D.. "The Formation of Human Populations in South and Central Asia." *Science*. 6 Sept. 2019. 365(6457).

Shinde, V.*; Narasimhan, V.*; Rohland, N.; Mallick, S.; Mah, M.; Lipson, M.; **Nakatsuka, N.**; Adamski, N.; Broomandkhoshbacht, N.; Ferry, M.; Lawson, A.; Michel, M.; Oppenheimer, J.; Stewardson, K.; Jadhav, N.; Kim. Y.; Chaterjee, M.; Munshi, A.; Panyam, A.; Waghmare, P.; Yadav, Y.; Patel, H.; Kaushik, A.; Thangaraj, K.; Meyer, M.; Patterson, N.; Rai, N.; Reich, D.. "An Ancient Genome from the Indus Valley Civilization Lacks Ancestry from Steppe Pastoralists or Western Iranian Farmers." *Cell*. 4 Sept. 2019. S0092-8674(19)30967-5.

**Nakatsuka, N.**; Moorjani, P.; Rai, N.; Sarkar, B.; Tandon, A.; Patterson, N.; Bhavani, G.; Girisha, K.; Mustak, M.; Srinivasan, S.; Kaushik, A.; Vahab, S.; Jagadeesh, S.; Satyamoorthy, K.; Singh, L.; Reich, D.*; Thangaraj, K.*. "The Promise of Discovering Population-Specific Disease-Associated Genes in South Asia." *Nature Genetics*. 1 Sept. 2017. 49(9): 1403-1407.

**Overview:**

The primary study in this chapter concerns the history of South Asia within the past ~2,000 years (~70 generations). The history of South Asia prior to this period was elucidated in Narasimhan *et al*. [1] and Shinde *et al*. [2], who demonstrated that an individual from the Indus Valley Civilization, as well as individuals with similar ancestry found in Central Asia, had Iranian-related ancestry that split from other Iranian plateau lineages over 12,000 years ago (thus, prior to the advent of farming in Iran). This suggests that farming in South Asia was developed by local South Asian hunter-

gatherers rather than coming in accompanied by major gene flow from Iranian farmers (in contrast to Europe, where farming was introduced via a large-scale immigration of Anatolian farmers [3]). The first mixture of Iranian-related ancestry with ancestral ancient South Indian ancestry was estimated to have occurred ~7400-5700 BP [1]. Around 4,000 BP, ancestry from Steppe pastoralists arrived and mixed into South Asia, though there are still some South Asian groups today that lack this ancestry in Southern India. Approximately 2,000 BP, groups in South Asia began to undergo strong founder events [4] that are the primary topic of this thesis chapter.

**Ethics:**

The ethics section in Chapter 1.2 primarily focused on aDNA, but it is equally important to follow proper ethical protocol in studies of present-day people. The large majority of the newly collected samples in this study were from India, so group consent was primarily conceived as following the governing structures present there. In India, the guidelines for good clinical practice in 2001 from the Indian Council of Medical Research emphasized the following principles: non-maleficence, beneficence, institutional arrangement, risk minimization, ethical review, voluntariness, and compliance [5]. We followed these principles in our study. For example, for institutional arrangement and ethical review, all samples were collected under the supervision of ethical review boards in India, namely, the approval of the Institutional Ethical Committees (IEC) of the CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India; Kasturba Hospital, Manipal, India; the Centre for Human Genetics, Banagalore, India; and the Fetal Care Research Foundation, Chennai, India. For voluntariness and compliance, on an individual consent level, informed consent was obtained from all subjects participating in the study for broad use of the material in medical and population genetics studies.

As with many countries, however, there is a high variance in the amount of representation that different groups have in their local and national governments (often for a variety of historical reasons), so it is possible that the regulations governing the genetic studies did not fully represent the wishes of all of the groups. Nevertheless, the sample collection and study design were led by Indians following the principles above such that if the groups had widespread objection to the study, their group would not have been included. The study was conducted to improve clinical genetics particularly in founder groups, but it will also have implications for improved health throughout India, in line with the beneficence principle. The potential follow-up to this study in terms of prenatal screening has the biggest potential risk for harm, but with proper care and precautions, these future studies could be conducted in an ethically responsible way that is sensitive to the complex cultural and socioeconomic environment of the local groups throughout South Asia.

# The promise of discovering population-specific disease-associated genes in South Asia

Nathan Nakatsuka[1,2], Priya Moorjani[3,4], Niraj Rai[5], Biswanath Sarkar[6], Arti Tandon[1,4], Nick Patterson[4], Gandham SriLakshmi Bhavani[7], Katta Mohan Girisha[7], Mohammed S Mustak[8], Sudha Srinivasan[9], Amit Kaushik[10], Saadi Abdul Vahab[11], Sujatha M. Jagadeesh[12], Kapaettu Satyamoorthy[11], Lalji Singh[13], David Reich[1,4,14,*], Kumarasamy Thangaraj[15,*]

[1]Department of Genetics, Harvard Medical School, New Research Building, 77 Ave. Louis Pasteur, Boston, MA 02115, USA
[2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Biological Sciences, Columbia University, 600 Fairchild Center, New York, NY 10027, USA
[4]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA
[5]Present address: Birbal Sahni Institute of Palaeosciences, Lucknow, Uttar Pradesh 226007, India
[6]Superintending Anthropologist (Physical) (Rtd.), Anthropological Survey of India, 27 Jawaharlal Nehru Road, Kolkata 700016, India
[7]Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal, India
[8]Department of Applied Zoology, Mangalore University, Mangalagangothri 574199, Mangalore, Karnataka, India
[9]Centre for Human Genetics, Biotech Park, Electronics City (Phase I), Bangalore 560100, India
[10]Amity Institute of Biotechnology, Amity University, Sector125, Noida 201303, India
[11]School of Life Sciences, Manipal University, Manipal 576104, India
[12]Fetal Care Research Foundation, 197 Dr. Natesan Road, Chennai 600004, India
[13]Present address: Genome Foundation, Hyderabad 500076, India
[14]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
[15]CSIR-Centre for Cellular and Molecular Biology, Habsiguda, Hyderabad, Telangana 500007, India
*co-senior authors

**Supplementary Material:** All supplementary material can be found in the supplement of Nakatsuka *et al.*, 2016 *Nature Genetics*.

# Abstract

The more than 1.5 billion people who live in South Asia are correctly viewed not as a single large population, but as many small endogamous groups. We assembled genome-wide data from over 2,800 individuals from over 260 distinct South Asian groups. We identified 81 unique groups, 14 of which had estimated census sizes of more than 1 million, that descend from founder events more extreme than those in Ashkenazi Jews and Finns, both of which have high rates of recessive disease due to founder events. We identified multiple examples of recessive diseases in South Asia that are the result of such founder events. This study highlights an underappreciated opportunity for decreasing disease burden among South Asians through discovery of and testing for recessive disease-associated genes.


# Body

South Asia is a region of extraordinary diversity, containing over 5,000 anthropologically well-defined groups, many of which are endogamous communities with substantial barriers to gene flow, owing to cultural practices that restrict marriage between groups [6]. Of the tiny fraction of South Asian groups that have been characterized with genome-wide data, many exhibit large allele-frequency differences from their close neighbors [4, 7, 8], reflecting strong founder events whereby a small number of ancestors gave rise to many descendants [4]. The pervasive founder events in South Asia present a potential opportunity for decreasing disease burden in South Asia, as highlighted by studies of founder groups of European ancestry – including Ashkenazi Jews, Finns, Amish, Hutterites, Sardinians, and French Canadians – which have resulted in the discovery of dozens of recessive disease-causing mutations in each group. Prenatal testing for these mutations has substantially reduced recessive disease burden in all of these communities [9, 10].

Here, we carried out new genotyping of 1,663 samples from 230 endogamous groups in South Asia by using the Affymetrix Human Origins single nucleotide polymorphism (SNP) array [11]. We combined the newly collected data with previously reported data, thus yielding four data sets (Figure 1A). The Affymetrix Human Origins SNP array data comprised 1,955 individuals from 249 groups in South Asia, to which we added data for 7 Ashkenazi Jews. The Affymetrix 6.0 SNP array data comprised 383 individuals from 52 groups in South Asia [4, 12]. The Illumina SNP array data comprised 188 individuals from 21 groups in South Asia [13] and 21 Ashkenazi Jews [13, 14]. The Illumina Omni SNP array data comprised 367 individuals from 20 groups in South Asia [15]. We merged 1000 Genomes Phase 3 data [16] (2,504 individuals, including 99 Finns, from 26 different groups) with each of these data sets. We removed SNPs and individuals that had a high proportion of missing genotypes or were outliers in Principal Components Analysis (PCA) (Figure 1B; Supplementary Note of Nakatsuka *et al*. [17]). The total number of unique groups analyzed in this study was 263, after accounting for groups represented in multiple data sets. To our knowledge, this work provides the richest set of genome-wide data from anthropologically well-documented groups from any region in the world.

We devised an algorithm to quantify the strength of the founder events in each group on the basis of identity by descent (IBD) segments, large stretches of DNA shared from a common founder in the last approximately 100 generations (Figure 2). We computed an IBD score (the average length of IBD segments between 3 and 20 centimorgans (cM) detected between two genomes, normalized to sample size) as a measure of the strength of the founder event in each group's history. Because we were interested in characterizing the effects of recessive diseases that did not originate from consanguineous marriages of close relatives, we ignored self-matches (internal homozygosity) in IBD calculations. We removed all individuals with evidence of recent relatedness (within several generations) to others in the data set by computing the IBD between all pairs of individuals in each group and removing one individual from each pair with an outlying number of IBD segments. (Our focus on founder events rather than recent relatedness also explains our choice to exclude IBD segments >20 cM in size.) We validated the effectiveness of this procedure by simulation (Supplementary Table 1 of Nakatsuka *et al.* [17] and Methods).

We expressed IBD scores for each group as a fraction of the IBD scores of the 1000 Genomes Project Finns merged into each respective data set. Because all the SNP arrays analyzed included more SNPs ascertained in Europeans than in South Asians, the sensitivity of our methods to founder events was greater in Europeans than in South Asians, and thus our estimates of founder event strengths in South Asian groups are conservative underestimates relative to those in Europeans. (Supplementary Figure 1 of Nakatsuka *et al.* [17] demonstrates this effect empirically and shows that it results in less bias for the strong founder events that were the focus of this study.) We computed standard errors for these ratios by using a weighted block jackknife across chromosomes and concluded significance when the 95% confidence intervals did not overlap 1. We further carried out computer simulations to validate our procedure. The simulations suggested that we did not substantially overestimate the magnitudes of modest founder events, because for a simulated founder event with half the magnitude of that in Finns, we never inferred the score to be significantly greater than that in Finns. The simulations also suggested that our procedure was highly sensitive to detecting strong founder events, because for sample sizes of at least 5, the algorithm's sensitivity was >95% for determining that a group with twice the bottleneck strength as that of Finns has an IBD score significantly greater than that of Finns (Supplementary Figure 2 and Supplementary Table 2 of Nakatsuka *et al.* [17]). We also used two additional non-IBD based methods to measure the strength of founder events and, in cases in which a comparison was possible, found that these results were highly correlated with our IBD scores (Supplementary Text and Supplementary Table 3 of Nakatsuka *et al.* [17]).

We inferred that 81 out of 263 unique groups (96 out of 327 groups if not considering the overlap of groups among data sets) had an IBD score greater than those of both Finns and Ashkenazi Jews (Figure 3). These results did not change when we added back the outlier samples that we removed in quality control. A total of 14 of these groups have estimated census sizes of over 1 million (Figure 3; Table 1). However, the groups with smaller census sizes are also important: outside of South Asia, groups with small census sizes and extremely strong founder events, such as Amish, Hutterites,

and people of the Saguenay-Lac Saint-Jean region have led to the discovery of dozens of novel disease-causing variants. We also searched for IBD across groups – screening for cases in which the across-group IBD score was at least one-third of the within-group IBD score of Ashkenazi Jews – and found 77 cases of clear IBD sharing, which typically followed geography, religious affiliation, or linguistic grouping (particularly for Austroasiatic speakers) (Supplementary Table 4 of Nakatsuka *et al.* [17]). Pairs of groups with high shared IBD and descent from a common founder event probably share risk for the same recessive diseases. However, these cross-group IBD sharing patterns did not drive our observations, because we identified 68 unique sets of groups without high IBD to other groups with significantly higher estimated IBD scores than both Finns and Ashkenazi Jews.

Our evidence that very strong founder events affect a large fraction of South Asian groups presents an opportunity to decrease disease burden in South Asia. This source of risk for recessive diseases is very different from risk due to marriages among close relatives, which is also a major cause of recessive disease in South Asia. To determine the relative impact of these factors, we computed $F_{ST}$, a measurement of allele-frequency differentiation, between each group in the data set and a pool of other South Asian groups chosen to be closest in terms of ancestry proportions. We found that inbreeding was not driving many of these signals, because 89 unique groups had higher $F_{ST}$ scores than those of Ashkenazi Jews and Finns, even after the $F_{ST}$ score was decreased by the proportion of allele-frequency differentiation due to inbreeding. These results show that although most studies mapping recessive disease-associated genes in South Asia have focused on families that are the products of marriages between close relatives, recessive diseases are also likely to occur at an elevated rate, even in nonconsanguineous cases, because of recent shared ancestors.

As an example of the promise of founder-event disease mapping of disease-associated genes in South Asia, we highlight the case of the Vysya, who have a census size of more than 3 million and an estimated IBD score approximately 1.2-fold higher than that of Finns (Figure 3). The Vysya have an approximately 100-fold higher rate of butyrylcholinesterase deficiency than other groups, and Vysya ancestry is a known counterindication for the use of muscle relaxants, such as succinylcholine or mivacurium, that are given before surgery [18]. Butyrylcholinesterase deficiency is likely to occur at an elevated rate, owing to the founder event in the Vysya's history, and we expect that, like Finns, the Vysya probably exhibit a higher rate of many other diseases than do other groups. Other examples of recessive disease-associated genes with a likely origin in founder events are known anecdotally in South Asia, thus highlighting the importance of systematic studies to identify these genes [19].

To demonstrate how a new recessive disease in a founder-event group can be mapped, we carried out genome-wide SNP genotyping in 12 patients from southern India who had progressive pseudorheumatoid dysplasia (PPD), a disease known to be caused by mutations in *WISP3* [20, 21]. Of the six individuals with the WISP3 p.Cys78Tyr substitution [20, 21], five were from nonconsanguineous marriages, and we found a much higher fraction of IBD at the disease-mutation site than in the rest of the genome in these individuals (Supplementary Figure 3A; Supplementary Figure 4A of Nakatsuka *et*

*al.* [17]), in agreement with the WISP3 p.Cys78Tyr substitution originating from a founder event and causing PPD in these patients. This pattern contrasted with those in the six patients with different disease variants as well as those in six patients who carried a mutation in *GALNS* causing a different disease (mucopolysaccharidosis IVA (MPS IVA)), who were from primarily consanguineous marriages and who lacked substantial IBD across their disease mutation sites. Thus, these results suggested that in these groups, the driver of the recessive diseases was marriage between close relatives (Supplementary Note of Nakatsuka *et al.* [17]). This example highlights how not only marriages of close relatives but also founder events are substantial causes of rare recessive disease in South Asia.

The evidence of widespread strong founder events presents a major opportunity for discovering disease-associated genes and implementing public-health interventions in South Asia that is not widely appreciated (Supplementary Table 5 of Nakatsuka *et al.* [17]). The current paradigm for mapping recessive disease-associated genes in South Asia is to collect cases in tertiary medical centers and map diseases in individuals with the same phenotype, a procedure often carried out by experimenters blinded to information about group affiliation, as was the case in our PPD study, in which we did not have access to the identity of the ancestral groups. However, our results suggested that collecting information on group affiliation may greatly strengthen the power of these studies. A fruitful way to approach gene mapping would be to proactively survey communities known to have strong founder events, searching for congenital diseases that occur at high rates in these communities. This approach was pioneered in the 1950s in studies of the Old Order Amish in the U.S., a founder population of approximately 100,000 individuals in whom many dozens of recessive diseases were mapped. That research program was crucial to founding modern medical genetics and provided extraordinary health benefits. Our results suggest that the potential for disease gene mapping in South Asia would be orders of magnitude greater.

Mapping of recessive diseases may be particularly important in communities practicing arranged marriages, which are common in South Asia. An example of the power of this approach can be found in *Dor Yeshorim*, a community genetic testing program among religious Ashkenazi Jews [22], which visits schools, screens students for common recessive disease-causing mutations previously identified to segregate at a higher frequency in the target group, and enters the results into a confidential database. Matchmakers query the database before making suggestions to the families and receive feedback about whether the potential couple is “incompatible” in the sense of both being carriers for a recessive mutation at the same gene. Because approximately 95% of community members whose marriages are arranged participate in this program, recessive diseases like Tay-Sachs have virtually disappeared in these communities. A similar approach should work as well in South Asian communities. Given the potential for saving lives, this or similar types of research could be a valuable investment for future generations [23].

## Supplementary Data:

The supplement of this study includes an Excel spreadsheet detailing all groups and their scores on the IBD, $F_{ST}$, and group-specific drift analyses. Also included are 7 supplementary figures and 5 supplementary tables.

## Acknowledgements:

We are thankful to the many Indian, Pakistani, Bangladeshi, Sri Lankan, and Nepalese individuals who contributed the DNA samples analyzed here including the PPD and MPS patients. We would like to thank Pier Palamara for helpful discussions about IBD and his help with ARGON. We are grateful to Analabha Basu and Partha P. Majumder for early sharing of data. Funding was provided by an NIGMS (GM007753) fellowship to NN, a Translational Seed Fund grant from the Dean's Office of Harvard Medical School and an HG006399 to DR, Council of Scientific and Industrial Research, Government of India grant to KT, and an NIGMS grant 115006 to PM. SS and SMJ acknowledge the funding from the Department of Biotechnology (BT/PR4224/MED/97/60/2011) and Department of Science and Technology (SR/WOS-A/LS-83/2011), Government of India. Funding for the mutation analysis of Indian patients with PPD was provided by Indian Council of Medical Research (BMS 54/2/2013) to KMG. DR is an Investigator of the Howard Hughes Medical Institute. The informed consents and permits associated with the newly reported data are not consistent with fully public release. Therefore, researchers who wish to analyze the data should send the corresponding authors a PDF of a signed letter containing the following language: "(a) We will not distribute the data outside my collaboration, (b) We will not post data publicly, (c) We will make no attempt to connect the genetic data to personal identifiers, (d) We will not use the data for commercial purposes."

## Author Contributions:

N.N., P.M., D.R., and K.T. conceived the study. N.N., P.M., N.R., B.S., A.T., N.P. and D.R. performed analysis. G.B., K.M.G., M.S.M., S.S. A.K., S.A.V., S.M.J., K.S., L.S. and K.T. collected data. N.N., D.R., and K.T. wrote the manuscript with the help of all co-authors.

## Competing Financial Interests:

The authors declare no competing financial interests.

# References:


1. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M: **The formation of human populations in South and Central Asia.** *Science* 2019, **365:**eaat7487.
2. Shinde V, Narasimhan VM, Rohland N, Mallick S, Mah M, Lipson M, Nakatsuka N, Adamski N, Broomandkhoshbacht N, Ferry M: **An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers.** *Cell* 2019, **179:**729-735. e710.
3. Feldman M, Fernández-Domínguez E, Reynolds L, Baird D, Pearson J, Hershkovitz I, May H, Goring-Morris N, Benz M, Gresky J: **Late Pleistocene human genome suggests a local origin for the first farmers of central Anatolia.** *Nature communications* 2019, **10:**1-10.
4. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
5. Mathaiyan J, Chandrasekaran A, Davis S: **Ethics of genomic research.** *Perspectives in clinical research* 2013, **4:**100.
6. Mastana SS: **Unity in diversity: an overview of the genomic anthropology of India.** *Ann Hum Biol* 2014, **41:**287-299.
7. Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BV, Rasanayagam A, Hammer MF: **Female gene flow stratifies Hindu castes.** *Nature* 1998, **395:**651-652.
8. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, et al: **Ethnic India: a genomic view, with special reference to peopling and structure.** *Genome Res* 2003, **13:**2277-2290.
9. Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, Rehnstrom K, Esko T, Magi R, Inouye M, Lappalainen T, et al: **Distribution and medical impact of loss-of-function variants in the Finnish founder population.** *PLoS Genet* 2014, **10:**e1004494.
10. Arcos-Burgos M, Muenke M: **Genetics of population isolates.** *Clin Genet* 2002, **61:**233-247.
11. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
12. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L: **Genetic evidence for recent population mixture in India.** *Am J Hum Genet* 2013, **93:**422-438.
13. Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Magi R, Metspalu E, Remm M, et al: **Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia.** *Am J Hum Genet* 2011, **89:**731-744.
14. Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, et al: **The genome-wide structure of the Jewish people.** *Nature* 2010, **466:**238-242.

15. Basu A, Sarkar-Roy N, Majumder PP: **Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure.** *Proc Natl Acad Sci U S A* 2016.
16. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.
17. Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S: **The promise of discovering population-specific disease-associated genes in South Asia.** *Nature genetics* 2017, **49:**1403.
18. Manoharan I, Wieseler S, Layer PG, Lockridge O, Boopathy R: **Naturally occurring mutation Leu307Pro of human butyrylcholinesterase in the Vysya community of India.** *Pharmacogenet Genomics* 2006, **16:**461-468.
19. Anju Shukla MH, Anshika Srivastava, Rajagopal Kadavigere, Priyanka Upadhyai, Anil Kanthi, Oliver Brandau, Stephanie Bielas, Katta Girisha: **Homozygous c.259G>A variant in ISCA1 is associated with a new multiple mitochondrial dysfunctions syndrome.** *bioRxiv* 2016.
20. Dalal A, Bhavani GS, Togarrati PP, Bierhals T, Nandineni MR, Danda S, Danda D, Shah H, Vijayan S, Gowrishankar K, et al: **Analysis of the WISP3 gene in Indian families with progressive pseudorheumatoid dysplasia.** *Am J Med Genet A* 2012, **158A:**2820-2828.
21. Bhavani GS, Shah H, Dalal AB, Shukla A, Danda S, Aggarwal S, Phadke SR, Gupta N, Kabra M, Gowrishankar K, et al: **Novel and recurrent mutations in WISP3 and an atypical phenotype.** *Am J Med Genet A* 2015, **167A:**2481-2484.
22. Raz AE: **Can population-based carrier screening be left to the community?** *J Genet Couns* 2009, **18:**114-118.
23. Rajasimha HK, Shirol PB, Ramamoorthy P, Hegde M, Barde S, Chandru V, Ravinandan ME, Ramchandran R, Haldar K, Lin JC, et al: **Organization for rare diseases India (ORDI) - addressing the challenges and opportunities for the Indian rare diseases' community.** *Genet Res (Camb)* 2014, **96:**e009.
24. Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, et al: **Global diversity, population stratification, and selection of human copy-number variation.** *Science* 2015, **349:**aab3761.
25. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al: **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.** *Nature* 2016, **538:**201-206.
26. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526:**75-81.
27. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
28. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.

29. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: **Whole population, genome-wide mapping of hidden relatedness.** *Genome Res* 2009, **19:**318-326.
30. Hoaglin BIaD: *How to Detect and Handle Outliers.* 1993.
31. Palamara PF: **ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process.** *Bioinformatics* 2016.
32. Browning SR, Browning BL: **Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.** *Am J Hum Genet* 2007, **81:**1084-1097.
33. Durand EY, Eriksson N, McLean CY: **Reducing pervasive false-positive identical-by-descent segments detected by large-scale pedigree analysis.** *Mol Biol Evol* 2014, **31:**2212-2222.
34. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data.** *Genetics* 2013, **194:**459-471.
35. Bidchol AM, Dalal A, Shah H, S S, Nampoothiri S, Kabra M, Gupta N, Danda S, Gowrishankar K, Phadke SR, et al: **GALNS mutations in Indian patients with mucopolysaccharidosis IVA.** *Am J Med Genet A* 2014, **164A:**2793-2801.

| Group | Sample Size | IBD Score | IBD Rank | $F_{ST}$ Rank | Drift Rank | Census Size | Location |
|---|---|---|---|---|---|---|---|
| Gujjar | 5 | 11.6 | 19 | 33 | 46 | 1,078,719 | Jammu and Kashmir |
| Baniyas | 7 | 9.6 | 24 | 22 | 18 | 4,200,000 | Uttar Pradesh |
| Pattapu_Kapu | 4 | 9.5 | 25 | 24 | 21 | 13,697,000 | Andhra Pradesh |
| Vadde | 3 | 9.2 | 26 | 30 | 26 | 3,695,000 | Andhra Pradesh |
| Yadav | 12 | 4.4 | 48 | 87 | 67 | 1,124,864 | Puducherry |
| Kshatriya_Aqnikula | 4 | 2.4 | 75 | 109 | NA | 12,809,000 | Andhra Pradesh |
| Naga | 4 | 2.3 | 76 | NA | NA | 1,834,483 | Nagaland |
| Kumhar | 27 | 2.3 | 77 | 35 | 197 | 3,144,000 | Uttar Pradesh |
| Reddy | 7 | 2.0 | 84 | 129 | 106 | 22,500,000 | Telangana |
| Brahmin_Nepal | 4 | 1.9 | 86 | 63 | 141 | 4,206,235 | Nepal |
| Kallar | 27 | 1.7 | 94 | 87 | 73 | 2,426,929 | Tamil Nadu |
| Brahmin_Manipuri | 17 | 1.6 | 99 | NA | NA | 1,544,296 | Manipur |
| Arunthathiyar | 18 | 1.3 | 108 | 109 | 81 | 1,192,578 | Tamil Nadu |
| Vysya | 39 | 1.2 | 110 | 46 | 35 | 3,200,000 | Telangana |

**Table 1. South Asian groups with estimated census sizes over 1 million and IBD scores significantly greater than those of Ashkenazi Jews and Finns.** 14 South Asian groups with IBD scores significantly higher than that of Finns, census sizes over 1 million, and sample sizes of at least 3 that are of particularly high interest for founder event disease gene mapping studies. For reference, Finns and Ashkenazi Jews (on the Human Origins array) would have IBD scores of 1.0 and 0.9, IBD ranks of 121 and 135, and $F_{ST}$ ranks of 109 and 129, respectively (the group-specific drift is difficult to compare for groups with significantly different histories, so they were not calculated for Finns or Ashkenazi Jews). NA, not available.

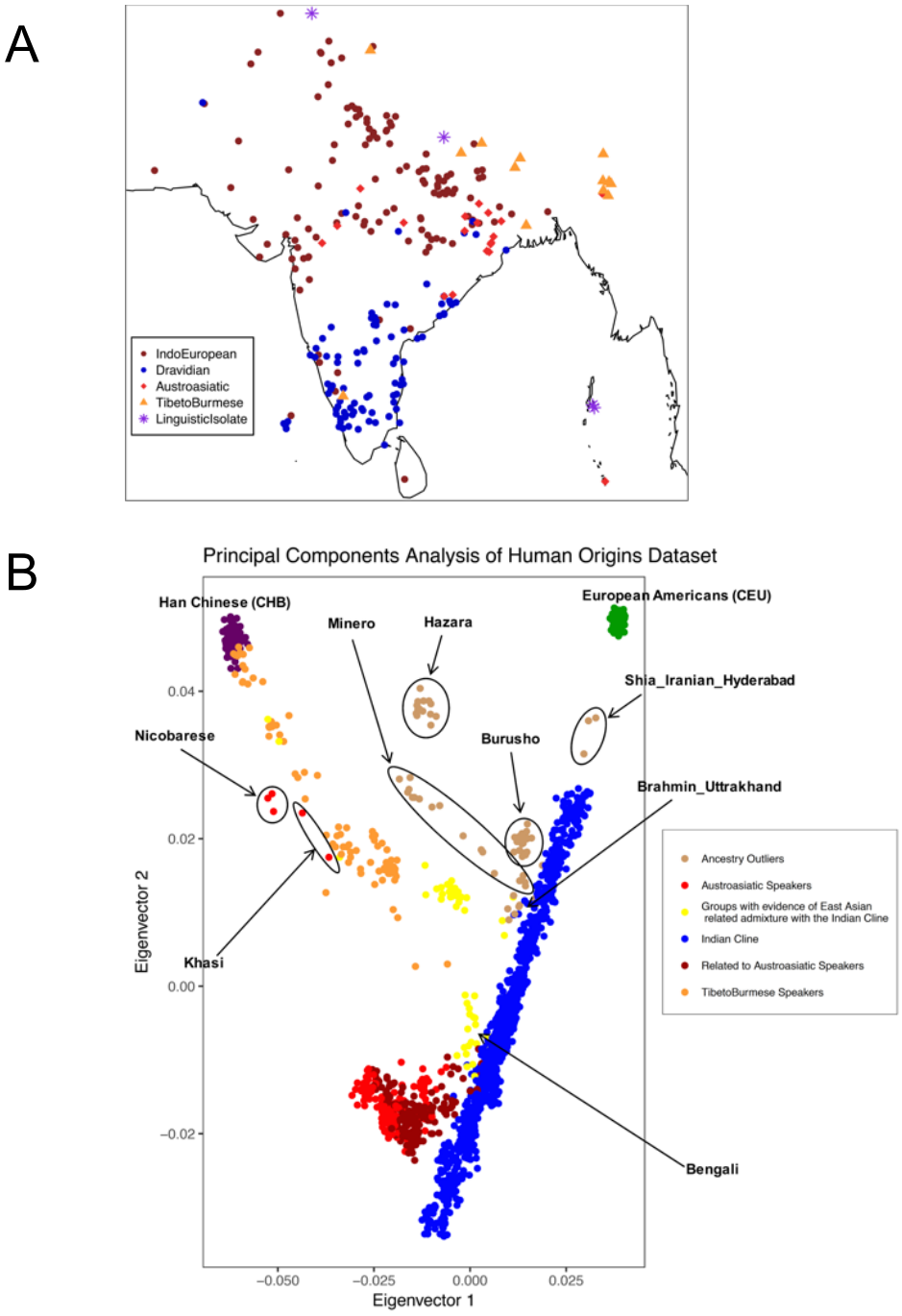


**Figure 1. Dataset overview. (A)** Sampling locations for all analyzed groups. Each point indicates a distinct group (random jitter was added to help in visualization at locations where there are many groups). **(B)** PCA of Human Origins dataset along with European Americans (CEU) and Han Chinese (CHB). There is a large cluster (blue) of IndoEuropean and Dravidian speaking groups that stretch out along a line in the plot and that are well-modeled as a mixture of two highly divergent ancestral populations (the "Indian Cline"). There is another larger cluster of Austroasiatic speakers (light red) and groups that cluster with them genetically (dark red). Finally, there are groups with genetic affinity to

East Asians that include Tibeto-Burman speakers (orange) and those that speak other languages (yellow).

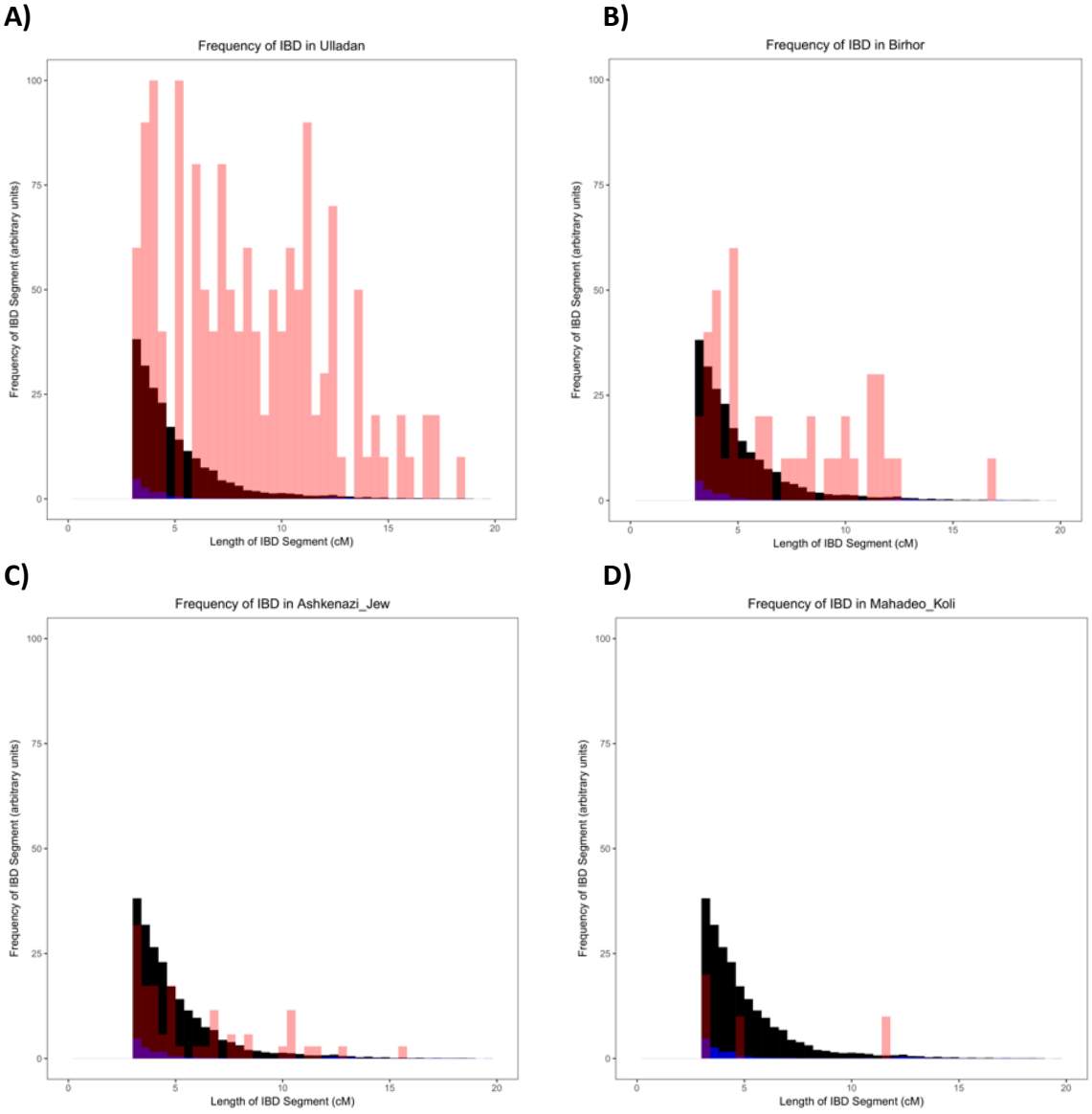


**Figure 2. Example histograms of IBD segments to illustrate the differences between groups with founder events of different magnitudes:** These histograms provide visual illustrations of differences between groups with different IBD scores. As a ratio relative to Finns (FIN; black), these groups (red) have IBD scores of: **(A)** ~26 in Ulladan, **(B)** ~3 in Birhor, **(C)** ~0.9 in Ashkenazi Jews, and **(D)** ~0.1 in Mahadeo_Koli. In each plot, we also show European Americans (CEU) with a negligible founder event in blue. Quantification of these founder events is shown in Figure 3 and Supplementary Table 5. The IBD histograms were normalized for sample size by dividing their frequency by $\left\{\binom{2n}{2} - n\right\}$, where $n$ is the number of individuals in the sample. All data for the figure are based on the Human Origins dataset.

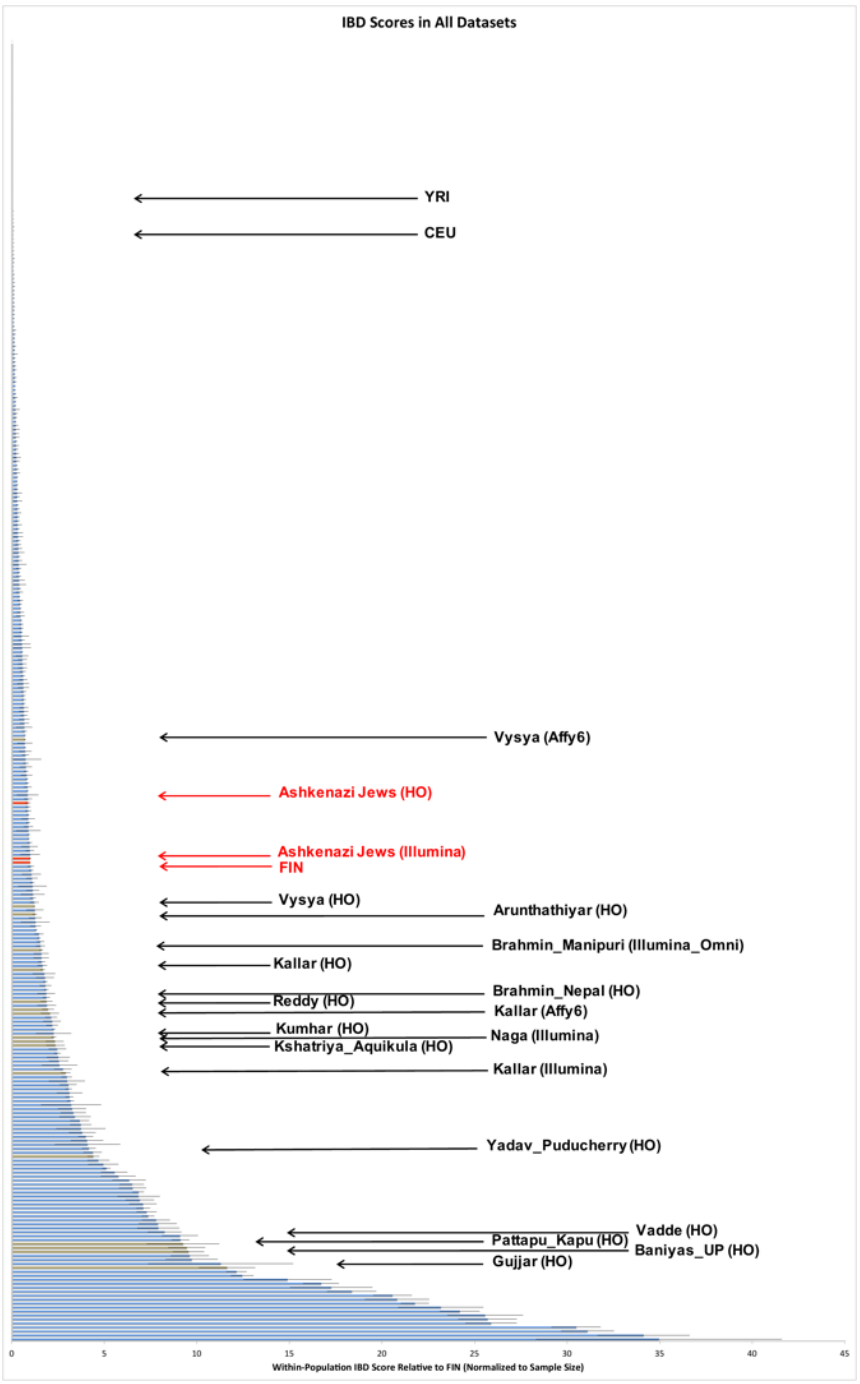


**Figure 3. IBD scores relative to Finns (FIN).** Histogram ordered by IBD score, roughly proportional to the per-individual risk for recessive disease due to the founder event. (These results are also given quantitatively for each group in Online Table 1 of Nakatsuka *et al.* [17].) We restrict to groups with at least two samples, combining data from all four genotyping platforms onto one plot. Data from Ashkenazi Jews and Finns are highlighted in red, and from South Asian groups with significantly higher IBD scores than that of Finns and census sizes of more than a million in brown. Error bars for each IBD score are standard errors calculated by weighted block jackknife over each chromosome. YRI=Yoruba (West African); CEU=European American.

# Materials and Methods:

## Data Sets:

We assembled a dataset of 1,955 individuals from 249 groups genotyped on the Affymetrix Human Origins array, of which data from 1,663 individuals from 230 groups are newly reported here (Figure 1A). We merged these data with the dataset published in Moorjani *et al*. [12], which consisted of 332 individuals from 52 groups genotyped on the Affymetrix 6.0 array. We also merged it with two additional datasets published in Metspalu *et al*. [13], consisting of 151 individuals from 21 groups genotyped on Illumina 650K arrays as well as a dataset published in Basu *et al*. [15], consisting of 367 individuals from 20 groups generated on Illumina Omni 1-Quad arrays. All the samples were collected with the approval of the Institutional Ethical Committees (IEC) of the CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India; Kasturba Hospital, Manipal, India; the Centre for Human Genetics, Banagalore, India; and the Fetal Care Research Foundation, Chennai, India.

These groups come from India, Pakistan, Nepal, Sri Lanka, and Bangladesh. All samples were collected under the supervision of ethical review boards in India with informed consent obtained from all subjects.

We analyzed two different Ashkenazi Jewish datasets, one consisting of 21 individuals genotyped on Illumina 610K and 660K bead arrays [14] and one consisting of 7 individuals genotyped on Affymetrix Human Origins arrays.

Our "Affymetrix 6.0" dataset consists of 332 individuals genotyped on 329,261 SNPs, and our "Illumina_Omni" dataset consists of 367 individuals genotyped on 750,919 SNPs. We merged the South Asian and Ashkenazi Jewish data generated by the other Illumina arrays to create an "Illumina" dataset consisting of 172 individuals genotyped on 500,640 SNPs. We merged the data from the Affymetrix Human Origins arrays with the Ashkenazi Jewish data and data from the Simons Genome Diversity Project [24, 25] to create a dataset with 4,402 individuals genotyped on 512,615 SNPs. We analyzed the four datasets separately due to the small intersection of SNPs between them. We merged in the 1000 Genomes Phase 3 data [26] (2,504 individuals from 26 different groups; notably, including 99 Finnish individuals) into all of the datasets. We used genome reference sequence coordinates (hg19) for analyses.

## Quality Control:

We filtered the data at both the SNP and individual level. On the SNP level, we required at least 95% genotyping completeness for each SNP (across all individuals). On the individual level, we required at least 95% genotyping completeness for each individual (across all SNPs).

To test for batch effects due to samples from the same group being genotyped on different array plates, we studied instances where samples from the same group *A* were genotyped on both plates 1 and 2 and computed an allele frequency difference at each SNP, $Diff_A^i = \left(Freq_{PopA,Plate1}^i - Freq_{PopA,Plate2}^i\right)$. We then computed the product of these allele frequencies averaged over all SNPs for two groups A and B genotyped on the same plates, $\frac{1}{n}\sum_{i=1}^{n}\left(Diff_A^i\right)\left(Diff_B^i\right)$, as well as a standard error from a weighted Block Jackknife across chromosomes. This quantity should be consistent with zero within a few standard errors if there are no batch effects that cause systematic differences across the plates, as allele frequency differences between two samples of the same group should be random fluctuations that have nothing to do with the array plates on which they are genotyped. This analysis found strong batch effects associated with one array plate, and we removed these samples from further analysis.

We used EIGENSOFT 5.0.1 *smartpca* [27] on each group to detect PCA outliers and removed 51 samples. We also developed a procedure to distinguish recent relatedness from founder events so that we could remove recently related individuals. We first identified all duplicates or obvious close relatives by using *PLINK* [28] "genome" and *GERMLINE* [29] to compute IBD (described in more detail below) and removed one individual from all pairs with a PI_HAT score greater than 0.45 and the presence of at least 1 IBD fragment greater than 30cM. We then used an iterative procedure to identify additional recently related individuals. For sample sizes above 5, we identified any pairs within each group that had both total IBD and total long IBD (>20cM) that were greater than 2.5 SDs and 1 SD, respectively, from the group mean. For sample sizes 5 or below, we used modified Z scores of 0.6745*(IBD_score - median(score))/MAD, where MAD is the median absolute deviation, and identified all pairs with modified Z scores greater than 3.5 for both total IBD and total long IBD as suggested by Iglewicz and Hoaglin [30]. After each round, we repeated the process if the new IBD score was at least 30% lower than the prior IBD score. Simulations showed that we were always able to remove a first or second cousin in the dataset using this method (Supplementary Table 1 of Nakatsuka *et al.* [17]). Together these analyses removed 53 individuals from the Affymetrix 6.0 dataset, 21 individuals from the Illumina dataset, 43 individuals from the Illumina Omni dataset, and 225 individuals from the Human Origins dataset.

After data quality control and merging with the 1000 Genomes Project data, the Affymetrix 6.0 dataset included 2,842 individuals genotyped on 326,181 SNPs, the Illumina dataset included 2,662 individuals genotyped on 484,293 SNPs, the Illumina Omni dataset included 2,828 individuals genotyped on 750,919 SNPs, and the Human Origins dataset included 4,177 individuals genotyped at 499,158 SNPs.

**Simulations to Test Relatedness Filtering and IBD Analyses**

We used *ARGON* [31] to simulate groups with different bottleneck strengths to test the IBD analyses and relatedness filtering. We used *ARGON*'s default settings, including a mutation rate of $1.65*10^{-8}$ per base pair (bp) per generation and a recombination rate

of $1*10^{-8}$ per bp per generation and simulated 22 chromosomes of size 130 Mb each. We pruned the output by randomly removing SNPs until there were 22,730 SNPs per chromosome to simulate the approximate number of positions in the Affymetrix Human Origins array. For the IBD analyses, we simulated groups to have descended from an ancestral group 1,800 years ago with $N_e$=50,000 and to have formed two groups with $N_e$=25,000. These groups continued separately until 100 generations ago when they combined in equal proportions to form a group with $N_e$=50,000. The group then split into 3 separate groups 72 generations ago that have bottlenecks leading to $N_e$ of either 400, 800, or 1600. The 3 groups then exponentially expanded to a present size of $N_e$=50,000. We designed these simulations to capture important features of demographic history typical of Indian groups [4, 12]. We chose the bottleneck sizes because they represent founder events with approximately the strength of Finns (the bottleneck to 800), and twice as strong (400) and half as strong (1600) as that group. We then performed the IBD analyses described below with 99 individuals from the group with bottleneck strength similar to that of Finns (198 haploid individuals were simulated and merged to produce 99 diploid individuals) and different numbers of individuals from the other groups. These analyses demonstrate that with only 4-5 individuals we can accurately assess the strength of founder events in groups with strong founder events (Supplementary Figure 2 and Supplementary Table 2 of Nakatsuka *et al.* [17]). Weaker founder events are more difficult to assess, but these groups are of less interest for founder event disease mapping, so we aimed to sample ~5 individuals per group.

We wrote custom R scripts to simulate first and second cousin pairs. We took individuals from the bottleneck of size 800 and performed "matings" by taking 2 individuals and recombining their haploid chromosomes assuming a rate of $1*10^{-8}$ per bp per generation across the chromosome and combining one chromosome from each of these individuals to form a new diploid offspring. The matings were performed to achieve first and second cousins. We then placed these back into the group with group of size 800, and ran the relatedness filtering algorithms to evaluate whether they would identify these individuals.

**Phasing, IBD Detection, and IBD Score Algorithm:**

We phased all datasets using *Beagle 3.3.2* with the settings *missing=0; lowmem=true; gprobs=false; verbose=true* [32]. We left all other settings at default. We determined IBD segments using *GERMLINE* [29] with the parameters *-bits 75 -err_hom 0 -err_het 0 -min_m 3*. We used the genotype extension mode to minimize the effect of any possible phasing heterogeneity amongst the different groups and used the *HaploScore* algorithm to remove false positive IBD fragments with the recommended genotype error and switch error parameters of 0.0075 and 0.003 [33]. We chose a *HaploScore* threshold matrix based on calculations from Durand *et al.* [33] for a "mean overlap" of 0.8, which corresponds to a precision of approximately 0.9 for all genetic lengths from 2-10cM. It can sometimes be difficult to measure IBD in admixed populations due to differential proportions of the divergent ancestries amongst different individuals in the same group,

but we found that in both the simulated and real data we were able to detect IBD at the expected amounts.

In addition to the procedure we developed to remove close relatives (Quality Control section), we also removed segments longer than 20cM as simulations showed that this increased sensitivity of the analyses (Supplementary Table 2 of Nakatsuka *et al.* [17]). We computed "IBD score" as the total length of IBD segments between 3-20cM divided by $\left\{\binom{2n}{2} - n\right\}$ where n is the number of individuals in each group to normalize for sample size. We then expressed each group's score as a ratio of their IBD score to that of Finns and calculated standard errors for this score using a weighted Block Jackknife over each chromosome with 95% confidence intervals defined as IBD score ±1.96*s.e.

We repeated these analyses with *FastIBD* [34] for the Affymetrix 6.0 and Illumina datasets and observed that the results were highly correlated (r>0.96) (data not shown). We chose *GERMLINE* for our main analyses, however, because the *FastIBD* algorithm required us to split the datasets into different groups, since it adapts to the relationships between LD and genetic distance in the data, and these relationships differ across groups. We used data from several different Jewish groups and all twenty-six 1000 Genomes groups to improve phasing, but of these groups we only included results for Ashkenazi Jews and two outbred groups (CEU and YRI) in the final IBD score ranking.

**Disease patient analyses:**

We use Affymetrix Human Origins arrays to successfully genotype 12 patients with progressive pseudorheumatoid dysplasia (PPD) and six patients with mucopolysaccharidosis (MPS) type IVA, all of whom had disease mutations previously determined [20, 21, 35] (3 of the surveyed MPS patients are newly reported here). A total of six of the PPD patients had p.Cys78Tyr substitutions, six had p.Cys337Tyr substitutions (all six of the MPS patients had Cys78Arg mutations). We measured IBD as described above and also detected homozygous segments within each individual by using GERMLINE with the parameters *-bits 75 -err_hom 2 -err_het 0 -min_m 0.5 -homoz-only*.

Haplotype sharing was assessed by analyzing phased genotypes for each mutation group. At each SNP, we counted the number of identical genotypes for each allele and computed the fraction by dividing by the total number of possible haplotypes (2 times the number of individuals). We took the larger value of the two possible alleles (thus the fraction range was 0.5-1). We averaged these values over blocks of 10 or 25 SNPs and plotted the averages around the relevant mutation site.

**Between-Group IBD Calculations:**

We determined IBD using *GERMLINE* as above. We collapsed individuals into respective groups and normalized for between-group IBD by dividing all IBD from each group by $\left\{\binom{2n}{2}\right\}$ where n is the number of individuals in each group. We normalized for within-

group IBD as described above. We defined groups with high shared IBD as those with an IBD score greater than three times the founder event strength of CEU (and ~1/3 the event strength of Ashkenazi Jews).

**$f_3$-statistics:**

We used the $f_3$-statistic [11] $f_3(Test; Ref_1, Ref_2)$ to determine if there was evidence that the *Test* group was derived from admixture of groups related to $Ref_1$ and $Ref_2$. A significantly negative statistic provides unambiguous evidence of mixture in the Test group. We determined the significance of the $f_3$-statistic using a Block Jackknife and a block size of 5 cM. We considered statistics over 3 standard errors below zero to be significant.

**Computing Group Specific Drift:**

We used *qpGraph* [11] to model each Indian group on the cline as a mixture of ANI and ASI ancestry, using the model (YRI, (Indian group, (Georgians, ANI)), [(ASI, Onge])) proposed by Moorjani *et al*. [12] This approach provides estimates for post-admixture drift in each group (Supplementary Figure 5 of Nakatsuka *et al.* [17]), which is reflective of the strength of the founder event (high drift values imply stronger founder events). We only included groups on the Indian cline in this analysis, and we removed all groups with evidence of East Asian related admixture (Figure 1B and Supplementary Table 6 of Nakatsuka *et al.* [17]), because this admixture is not accommodated within the above model.

**PCA-Normalized $F_{ST}$ Calculations:**

As a third method to measure strength of founder events, we estimated the minimum $F_{ST}$ between each South Asian group (Supplementary Figure 6) and their closest clusters based on PCA (Supplementary Note of Nakatsuka *et al.* [17]) (the clusters were used to account for intermarriage across groups that would otherwise produce a downward bias in the minimum $F_{ST}$). For the Affymetrix 6.0, Illumina, and Illumina_Omni datasets, we split the Indian cline into two different clusters and combined the Austroasiatic speakers and those with ancestry related to Austroasiatic speakers (according to the PCA of Figure 1b) into one cluster for a total of three clusters (all other groups were ignored for this analysis). For the Human Origins data set we split the Indian cline into three different clusters and combined the groups with ancestry related to the main cluster of Austroasiatic speakers into one cluster for a total of four clusters (Khasi and Nicobarese were ignored in this analysis, because they do not cluster with the other Austroasiatic speaking groups). We then computed the $F_{ST}$ between each group and the rest of the individuals in their respective cluster based on EIGENSOFT *smartpca* with Inbreed set to YES to correct for inbreeding. For Ashkenazi Jews and Finns, we used the minimum $F_{ST}$ to other European groups.

**$F_{ST}$ Calculations to Determine Overlapping Groups:**

Overlapping groups between the datasets were determined in the first place based on anthropological information (Online Table 1 of Nakatsuka *et al.* [17]). We further tested empirically for overlap by computing $F_{ST}$ between different groups across all datasets for groups with significantly stronger IBD scores than those of Finns (we could not perform this analysis for groups with less strong founder events, because they would have low $F_{ST}$ to each other even if they were truly distinct groups). We considered pairs with $F_{ST}$ less than 0.004 to be overlapping. These included all groups known to be overlapping based on anthropological information as well as 3 additional pairs of groups that might be genetically similar due to recent mixing (e.g. Kanjars and Dharkar are distinct nomadic groups that live near each other but intermarry, leading to low $F_{ST}$ between them).

**Code Availability:**

Code for all calculations available upon request.

# Chapter 3: Using modern DNA to perform admixture mapping in African-Americans with and without Multiple Sclerosis

This section is based on the following paper I contributed to during my PhD:

> **Nakatsuka, N.**; Patterson, N.; Patsopoulos, N.; Altemose, N.; Tandon, A.; Beecham, A.; McCauley, N.; Isobe, N.; Hauser, S.; De Jager, P.; Hafler, D.; Oksenberg, J.; Reich, D. "Two Genetic Variants Explain the Association of European Ancestry to Multiple Sclerosis Risk in African Americans." In review at *Scientific Reports*.

**Overview:**

In Chapter 2, human history in the form of large founder events was examined in the DNA of present-day people for the purpose of disease gene mapping to improve human health. In this chapter, human history in the form of historical admixture between individuals of West African ancestry and individuals of European ancestry was examined for the purpose of mapping genetic variants underlying Multiple Sclerosis. The admixture in African-Americans was inferred to have occurred an average of approximately six generations ago, and the European ancestry was male-biased [1]. It is likely that a substantial proportion of this mixture occurred due to abusive power dynamics in the context of plantation slavery, leaving a large moral stain ingrained in genetic legacy. However, one hope of this study is to use this genetic history to improve health, particularly for African-Americans with Multiple Sclerosis.

**Ethics:**

This study was conducted solely in the United States, so the governmental regulations in the US were followed. In particular, institutional review board (IRB) approval was performed by the UCSF Human Research Protection Program Institutional Review Board (IRB) (10-05039), and full informed consent was obtained for each study subject. It is true that oftentimes African-Americans are under-represented in the local and federal government, as well as on institutional review board committees, and that this study could be sensitive due to it potentially evoking memories of past (and present) oppression and abuse. Thus, we needed to be respectful to principles of autonomy and non-maleficence. There is no institutional system clearly designated by African-Americans to represent them specifically as a group for these issues, so in this case we proceeded based on the permissions we obtained, recognizing that at least this shows that a significant fraction of the relevant group (1,305 African-Americans with Multiple Sclerosis) approved of the study. In addition, the principles of beneficence and restorative justice hopefully will be fulfilled as this study has potential implications for improving the health of Multiple Sclerosis patients, and particularly African-American patients who were the subject of this study. In addition, many have recognized the potential power of population genetics for healing in African-American communities [2].

# Two Genetic Variants Explain the Association of European Ancestry with Multiple Sclerosis Risk in African-Americans

Nathan Nakatsuka[1,2,*], Nick Patterson[3,4], Nikolaos A. Patsopoulos[4,5,6], Nicolas Altemose[7], Arti Tandon[1,4], Ashley H. Beecham[8], Jacob L. McCauley[8,9], Noriko Isobe[10], Stephen Hauser[10], Philip L. De Jager[4,11], David A. Hafler[4,12], Jorge R. Oksenberg[10], David Reich[1,3,4,13,*]

[1]Department of Genetics, Harvard Medical School, New Research Building, Boston, MA 02115, USA
[2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Human Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA
[4]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA
[5]Systems Biology and Computer Science Program, Ann Romney Center for Neurological Diseases, Department of Neurology, Brigham & Women's Hospital, Boston, MA 02115, USA
[6]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[7]Department of Bioengineering, University of California Berkeley, San Francisco, Berkeley, CA 94720
[8]John P. Hussman Institute for Human Genomics, Miller School of Medicine, University of Miami, Miami, FL 33136, USA
[9]Dr. John T. Macdonald Foundation Department of Human Genetics, Miller School of Medicine, University of Miami, Miami, FL 33136, USA
[10]Department of Neurology, University of California San Francisco School of Medicine, San Francisco, CA 94158, USA
[11]Center for Translational & Computational Neuroimmunology, Department of Neurology, Columbia University Irving Medical Center, New York, NY, 10032, USA
[12]Departments of Neurology and Immunobiology, Yale School of Medicine, New Haven, CT 06520, USA
[13]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

* Corresponding authors: Nathan Nakatsuka (nathan_nakatsuka@hms.harvard.edu) and David Reich (reich@genetics.med.harvard.edu)

**Keywords:**
Multiple sclerosis, admixture mapping, African-Americans

**Supplementary Material:** All supplementary material can be found in the supplement of Nakatsuka *et al.*, 2020 *Scientific Reports*, in review.

# Abstract:

Epidemiological studies have suggested differences in the rate of multiple sclerosis (MS) in individuals of European ancestry compared to African ancestry, motivating genetic scans to identify genetic factors that could contribute to such patterns. In a whole-genome scan in 899 African-American cases and 1,155 African-American controls, we confirm that African-Americans who inherit segments of the genome of European ancestry at a chromosome 1 locus are at increased risk for MS [logarithm of odds (LOD) = 9.8], although the signal weakens when adding an additional 406 cases, reflecting heterogeneity in the two sets of cases [logarithm of odds (LOD) = 2.7]. The association in the 899 individuals can be fully explained by two variants previously associated with MS in European ancestry individuals. These variants tag a MS susceptibility haplotype associated with decreased *CD58* gene expression (odds ratio of 1.37; frequency of 84% in Europeans and 22% in West Africans for the tagging variant) as well as another haplotype near the *FCRL3* gene (odds ratio of 1.07; frequency of 49% in Europeans and 8% in West Africans). Controlling for all other genetic and environmental factors, the two variants predict a 1.44-fold higher rate of MS in European-Americans compared to African-Americans.


# Background:

Admixed populations are formed when two populations with divergent ancestries have offspring. Admixture mapping is a method to screen through the genome, searching for loci where individuals with a disease from an admixed population tend to have a significantly different proportion of one ancestry than their population's genome-wide average. For example, admixture mapping in African-Americans involves searching genomes for areas where European ancestry deviates from the average, which is roughly 20% in African-Americans albeit with substantial variation across the United States [3]. These genomic loci of high deviation in local ancestry indicate the presence of disease risk variants that differ in frequency between the ancestral populations, which can then be followed up using fine-mapping approaches. To date, admixture mapping has successfully identified genomic loci that were then fine-mapped to identify genetic risk variants for several diseases, notably prostate cancer and end-stage renal disease [4-7].

The first successful admixture mapping study was for multiple sclerosis (MS) in African-Americans, a good candidate disease for this method because early prevalence studies suggested that individuals of European ancestry have significantly higher rates of MS than African-Americans (1.49- to 2.27-fold) [8-10]. However, more recent studies have disputed the epidemiological observation, with one showing African-American women having a 1.59-fold higher risk than European-American women and African-American men having the same risk as European-American men [11]. Another study suggested that African-Americans of both sexes have a 1.27-fold higher risk of MS than European-Americans [12]. In 2005, admixture mapping analysis led to the discovery of a genetic

risk factor for MS near the centromere of chromosome 1 estimated to lead to an approximately 1.44-fold increase in MS risk per allele of European ancestry relative to African ancestry [13]. Since the publication of that study, joint analyses of genome-wide association studies (GWAS) and targeted SNP genotyping, largely focused on people of European ancestry, have discovered 201 genetic risk variants for MS outside of the MHC (Major Histocompatibility Complex) region and 32 variants within the MHC region [14-19]. However, there is evidence for incomplete overlap between African-American and European MS risk variants [20], and the specific variants responsible for the 2005 admixture association in African-Americans have to date been unresolved.

We analyzed 1,305 African-American MS cases and 1,155 African-American controls and found that the 2005 admixture mapping signal could be fully explained by two variants that are strongly correlated with haplotypes previously linked to MS in people of European ancestry: one in the *CD58* gene located on the p-arm near the centromere and one in the *FCRL3* gene on the q-arm near the centromere. We suggest that the localization of the admixture association signal over the centromere of chromosome 1 is due to a combined association of risk factors straddling the centromere.

# Results:

**Replication of Chromosome 1 Signal:**

As part of a large replication study of multiple sclerosis risk variants, a custom genotyping array (MS Chip) was designed using the Illumina platform that included 321,105 single nucleotide polymorphisms (SNPs) across the genome that were genotyped in 24,770 cases with multiple sclerosis (of which 1,305 were African-American) and 23,193 controls without multiple sclerosis (of which 1,155 were African-American) [17]. A total of 9,014 SNPs were incorporated into the array specifically to study the chromosome 1 admixture mapping result published in 2005, including 4,014 Ancestry Informative Markers (AIMs, known to be highly differentiated between people of West-African and European ancestry) spread genome-wide, and 5,000 SNPs within the chromosome 1 admixture mapping peak, which we designed to provide dense coverage in both European and African haplotype backgrounds and to include SNPs in non-repetitive regions within the chromosome 1 centromere. Detailed design specifications of the SNP array are given elsewhere [17].

We initially analyzed all 1,305 African-American MS cases and 1,155 African-American controls together (see Supplementary Online Table 1 for sample details) using the ANCESTRYMAP software [3], which calculates LOD scores across the genome as a Bayesian likelihood ratio of the genetic site's likelihood of the data under the specified disease model divided by the likelihood of the data under no disease model. We found a signal near the centromere of chromosome 1 (maximum LOD score of 2.7) that was significantly weaker than the result of the 2005 scan that had obtained a score of 5.2 in 605 cases [13] (which met the published threshold for genome-wide significance of 5) (Table 1 and Supplementary Figure 1). We were perplexed by this result, as the

maximum LOD score at chromosome 1 had increased to 9.3 in a follow-up study in 2007 where the sample size was increased to 1,044 cases albeit using a less dense set of AIMs than in the study reported here and using a mixture of genotyping methods [21]. When we restricted the analysis of our new data to the 899 MS subjects that overlapped the 2007 study, the maximum LOD on chromosome 1 went up to 9.8 (Supplementary Figure 2) in a region overlapping the centromere (physical position 116 Mb-164 Mb in hg19 genomic coordinates). In addition, when we used a permutation test to determine an empirical p-value, this was significant at $P<<0.001$ (Table 1). This provided a technical validation of the 2007 result using a different genotyping platform on the same samples.

We separately analyzed the 406 new MS cases that did not overlap with the 2007 cases and found no significant signal at the chromosome 1 locus. Indeed, the 95% confidence interval for risk for MS per copy of European ancestry is significantly below that of African ancestry for the post-2007 cases (0.59-0.91-fold per copy of European ancestry), compared to a non-overlapping 95% confidence interval of 1.37-1.76-fold for the 2007 cases. This suggests an opposite effect of increased risk due to African ancestry at this locus, not increased risk due to European ancestry (Table 1, Supplementary Figure 3). Comparing the cases with data available to us in 2007 and the 406 post-2007 cases, we could not detect any difference in genome-wide ancestry proportions, sex proportions (ratio of females to males), quality control measures (genotyping error rate or PCA clustering, including after restricting to the region around the centromere of chromosome 1) (Supplementary Figure 4), origin of DNA (cell line vs. genomic DNA or location of sample collection), or Native American ancestral contribution (Supplementary Online Table 1). Study inclusion/exclusion criteria remained the same and were strictly implemented for all datasets. When we removed all samples known to be of Afro-Caribbean origin, the results did not change significantly. When we analyzed the HLA-DRB1 status (known to be associated strongly with MS [22]), the 2007 cases and the new cases did not show any significant difference in association to MS (Supplementary Online Table 1). The cases genotyped after 2007 had a lower average copy number for the HLA-DRB1*15:01/15:03 haplogroup (38%) relative to the cases available as of 2007 (42%), which meant that the elevation of the frequency of this haplogroup relative to the controls (34%) was significant for the cases available as of 2007 but not for the ones added afterward. Along with the attenuation of the admixture mapping signal, this result suggests that the cases collected up until 2007 may have been more enriched for people with genetic susceptibility to MS than the controls. We have no evidence that a sample mix-up occurred, and overall we could not find any significant difference between the 2007 cases and the new cases beyond the inhomogeneity in their contribution to the chromosome 1 admixture mapping peak. Nevertheless, motivated by the validation of the admixture signal in 2007 and access to high-density new genotyping data, we proceeded to use these data to understand the basis of this signal, restricting to the 899 cases that overlapped between the 2007 and 2016 studies as it was in this subset of cases that we had a strong ancestry association that we could parse in terms of specific variants contributing to it.

| Study | Cases | Controls | Highest LOD score on chr. 1 case-only | Highest LOD score on chr. 1 cases + controls | Highest case-control Z-score on chr. 1 | Empirical p-value for admixture association | 95% CI for risk per European chromosome |
|---|---|---|---|---|---|---|---|
| 2005 study | 605 | 1043 | 5.2 | 5.2 | 3.3 | N/A | 1.27-1.70 |
| 2007 follow-up | 1044 | 1161 | 9.2 | 9.3 | 4.2 | <<0.001 | 1.32-1.62 |
| Full new cohort | 1305 | 1155 | 3.9 | 2.7 | 3.4 | <<0.001 | 1.16-1.43 |
| 2007 subset of new cohort | 899 | 1155 | 9.8 | 9.3 | 4.5 | 0.114 | 1.37-1.76 |
| Samples added after 2007 | 406 | 1155 | -3.6 | -3.0 | 0.9 | 1 | 0.59-0.91 |

**Table 1. Admixture association scores on chromosome 1 in different sample sets.** The highest LOD score results are computed by ANCESTRYMAP based on a prior on relative risk per European ancestry allele of 1.5. The 95% confidence intervals for risk per European allele are obtained by running on a uniformly spaced grid of models from 0.5-2.0-fold per European allele, assuming an equal prior probability of risk for each, and then taking the LOD score to the power of 10 and normalizing to obtain a posterior. The LOD scores for the grid of models for the 2005 study are from the original publication [13]; all LOD scores are given in Supplementary Online Table 3. Empirical p-values were found through permutation analysis; see Methods (2005 study could not be done due to lack of the original data). P-values were listed as <<0.001 when 0 of the 1,000 permutations had a score at least as large as that of the LOD score of that data subset.

## Determining the Basis of the Chromosome 1 Signal:

We searched for potential causal variants in the chromosome 1 admixture mapping peak, which we defined for this analysis as 114-162 Mb on chromosome 1 in hg19 genome coordinates (that is, the area with LOD score>5 in the admixture analysis adding 2 Mb on either side; Figure 1). The African-American sample size was insufficient to obtain any genome-wide significant associations (Supplementary Figure 5) as suggested by power estimates (Supplementary Table 1). We thus began by narrowing our search to variants with some evidence for association with MS from a study of a much larger sample size of people of European ancestry who were genotyped on a custom SNP array designed to test for association to MS [17] (Supplementary Online Table 2; Supplementary Figure 6), making the plausible assumption that the variants associated with disease risk in European-Americans are also risk factors in African-Americans [20]. Seventy-nine variants in the region had a p-value $<10^{-5}$ in the European ancestry association study, and we focused our analysis on these 79 variants in all subsequent analyses. Within the admixture mapping peak, the European ancestry association study had identified seven independent genetic tag variants that together were sufficient to account for all the genome-wide significant association to MS detected in that study (that is, controlling for the allelic status of these variants, there was no additional genome-wide evidence of association to MS in the genomic region) [17]. However, we found that these seven tag SNPs were not sufficient to account for the association to MS detected in African-Americans, as if we condition on the allelic status of all seven following the procedure discussed in what follows, there is still a residual ancestry association signal in African-Americans with LOD=3.6.

In light of the fact that the tag SNPs from the European genetic map were insufficient to explain the admixture association signal in African-Americans, we turned to studying the 79 variants significant at $p<10^{-5}$ in the European ancestry association study and

evaluating their contribution to MS in African-Americans. Specifically, we used these variants to perform a logistic regression in the African-American data of SNP genotype association on case-control status while controlling for genome-wide and local European ancestry as covariates (Figure 1, Supplementary Online Table 3). Because we controlled for ancestry in this analysis, this procedure treated the SNP association signals completely independently from the ancestry association signals, allowing us to drive the analysis entirely by SNP association results and only afterward to determine whether the chosen SNPs explained the admixture mapping peak.

The SNP with the strongest association to MS in the African-American data was rs12025416 in the vicinity of the *CD58* gene. After using this SNP as a covariate in a logistic regression of European ancestry on case-control status with global ancestry also as a covariate, the significant ancestry association signal disappeared on the p-arm ($p>0.01$ for the ancestry association and $p>0.05$ for the genotype associations at all SNPs). However, an ancestry association signal on the q-arm remained ($p<0.01$). When we instead used the *CD58* SNP with the strongest signal in the European MS Chip association analysis (rs1335532) [17], the ancestry association signal on the p-arm was still present ($p<0.005$), confirming the value of guiding our analysis using the SNP associations measured directly in African-Americans; that is, on focusing on the most strongly associated SNP from the African-American scan rather than the most strongly associated SNP from the European scan (Supplementary Online Table 3). In the ancestry association tests controlling for SNP association here and in what follows, we also carried out follow-up analyses re-inferring ancestry after masking all positions within 1 centi-Morgan (cM) of the variants whose SNP associations we were using as covariates to address the potential pitfall that SNPs used in the admixture analysis might be in tight LD with these SNPs. These sub-analyses did not lead to any significant change in our results.

We next took the most highly associated variant after conditioning on allelic status of rs12025416, which was rs6681271 in the vicinity of the *FCRL3* gene (Figure 1). After using rs6681271 jointly with rs12025416 as covariates, there was no residual association in African-Americans across the entire admixture mapping peak ($p>0.05$ for both ancestry association and genotype association). When any significant variants (from the 79 above) on the q-arm were used as covariates instead of rs6681271, only one, rs7528684 also in the *FCRL3* gene, produced more of a decrease in the ancestry association signal. However, we cannot statistically distinguish between the signals and hence we have chosen to tag the *FCRL3* association using rs6681271, which we obtained through a formal procedure of using SNP associations entirely independently of ancestry associations (Supplementary Online Table 3). Similar to the *CD58* variant, when we instead conditioned on the strongest associated *FCRL3* variant from the European MS Chip association analysis (rs3761959) [17], some ancestry association signal on the q-arm remained ($p<0.005$) (Supplementary Online Table 3), reaffirming how the European MS Chip association cannot by itself identify variants that explain the signal of

admixture association in African-Americans. Thus, the African-American data adds additional valuable information about MS risk even with its much smaller sample size.

After controlling for the top two variants associated to MS risk in African-Americans, there is no significant MS risk at *CD58* (residual association of p=0.50 at the European study tag SNP rs1335532) or *FCRL3* (residual association of p=0.22 at the European study tag SNP rs3761959). In contrast, as discussed above, controlling for those top two variants from the European study and indeed all seven across the MS peak does not account for all evidence of association in African-Americans (after conditioning on all seven SNPs, there is a residual SNP association at rs12025416 in *CD58* of p=0.039 and at rs6681271 in *FCRL3* of p=0.0082; the residual admixture association has LOD=3.6).

Why is it that the seven variants that were the best tag SNPs in the region from the European genetic map cannot explain the admixture association in African-Americans, whereas the top two SNPs identified in an African-American association scan are able to explain it? A possible explanation is that shorter average linkage disequilibrium (LD) in genomic segments of African ancestry compared to European ancestry allow us to more finely map the true causal variants, which by implication would be different from the tag SNPs that emerged from the European ancestry association study. The variants we identify as associated to MS are in high LD with the SNPs in the same gene identified in the association study in people of European ancestry ($r^2$ = 0.73 between rs12025416 and rs1335532, and $r^2$ = 0.89 between rs6681271 and rs3761959, Supplementary Online Table 4; they are on the same risk haplotype) but less so in people of African ancestry ($r^2$ = 0.23 between rs12025416 and rs1335532, and $r^2$ = 0.20 between rs6681271 and rs3761959). While the higher frequencies of the risk alleles in people of European ancestry than of African ancestry for both tag SNPs means that the LD structure in Europeans is more relevant for these SNPs in African-Americans than might be expected from the overall proportion of European ancestry in the population, a substantial fraction of the risk alleles in African Americans are still coming from an African ancestry background given the overall very high proportion of African ancestry in African-Americans, and this decreases the relevance of the European tag SNPs.

In the 1000 Genomes Phase 3 data, the allele tagging the *CD58* risk haplotype (rs12025416) has a frequency of 83.8% in CEU (European-Americans of Northern European ancestry), 21.8% in YRI (Yoruba of Nigeria), and 31.9% in CHB (Han Chinese) and JPT (Japanese). In the European MS Chip association, it had a p-value of $3.32*10^{-32}$ and an effect size (odds ratio) of 1.37 (Table 2), one of the strongest risk alleles for MS outside of the MHC. The *FCRL3* risk variant (rs6681271) has a frequency of 49.0% in CEU, 7.9% in YRI, and 56.5% in CHB and JPT. In the European MS Chip association study, it had a p-value of $3.24*10^{-6}$ and odds ratio of 1.07. If one makes the simplifying assumption of a constant (fixed) effect size in all populations, the population attributable risk (PAR—the expected reduction in MS incidence if the risk alleles did not exist in the population) for both variants together was 45% for European-Americans, 15% for West Africans, and 21% for African-Americans. (For just the *CD58* variant, the PAR is 41% for European-

Americans, 14% for West Africans, and 20% for African-Americans. For just the *FCRL3* variant the numbers are 7% for people of European ancestry, 1% for West Africans, and 2% for African-Americans.) This implies a 1.44-fold higher risk for MS in Europeans Americans than in African-Americans, after controlling for all other genetic and environmental factors (1.44=(100%-21%)/(100%-45%)).

Another way to assess the epidemiological effect on relative risk for MS in European-Americans compared to African-Americans is to directly use the ancestry association. Focusing on the estimated risk per copy of European ancestry in the 899 cases available up until 2007 and included in most of the analyses in this study, we infer the increased risk per copy of European ancestry to be 1.54 per copy of European ancestry at the chromosome 1 locus (95% confidence interval of 1.37-1.76; Table 1). This implies that African-Americans with two copies of European ancestry at the locus (corresponding to roughly 4%=(20%)*(20%) of African-Americans) have 1.93-fold higher risk for MS than the average African-American (95% confidence interval of 1.67-2.33). This is higher than the increased risk in European-Americans of 1.44 computed just based on the two variants (above), and suggests the possibility that their effect size estimates in European-Americans may be underestimates of those in African-Americans, perhaps reflecting different gene-environment interactions. Alternatively, using the estimates of ancestry association in the full cohort of 1305 individuals (Table 1), we obtain an estimate of 1.47 increased risk in African-Americans carrying two European copies at the site relative to African-Americans (95% confidence interval 1.26-1.73), which is more in line with the SNP association results.

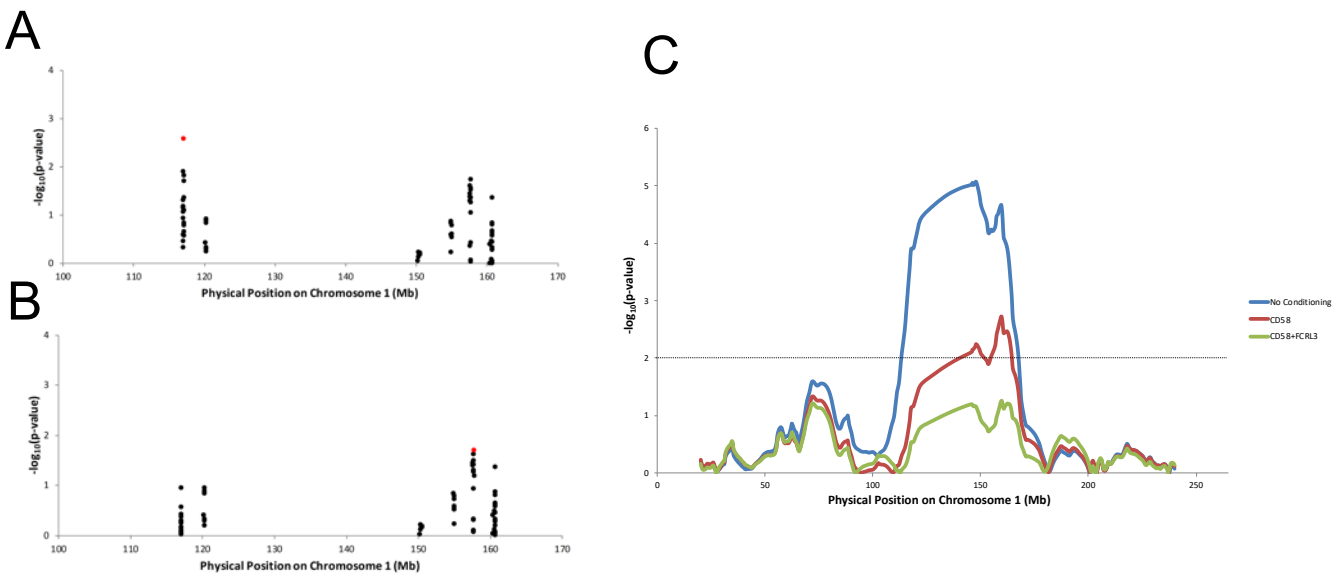


**Figure 1. Two variants are sufficient to explain the admixture association signal. (A-B)** Top GWAS variants in the region of the admixture association signal (red box) were taken and used in a logistic regression for genotype association on MS case-control status in African-American data after conditioning on global and local European ancestry. Y-axis is $-\log_{10}$(p-value) of association with MS case status. Shown in red are the most highly associated variants, rs12025416 in the *CD58* gene for panel A and rs6681271 in the *FCRL3* gene in panel B (after conditioning on the top variant in panel A). **(C)** Logistic regression of local European ancestry on case-control status in African-American data after controlling for global ancestry as a covariate as well as the top variants from panels A and B. No conditioning indicates only controlling for global ancestry. The dotted line indicates threshold for significance (this p-value threshold represents a lower bound on significance due to the fact that the peak can shift after conditioning). The African-American data used for all analyses was the 2007 subset of the new cohort.

| Dataset | p-value for association with MS status at top *CD58* variant | p-value for association with MS status at top *FCRL3* variant |
|---|---|---|
| Europeans | $3.32 * 10^{-32}$ | $3.24 * 10^{-6}$ |
| African-Americans before conditioning | $2.51 * 10^{-3}$ | $1.80 * 10^{-2}$ |
| African-Americans after conditioning on top *CD58* variant | $1.10 * 10^{-1}$ | $1.97 * 10^{-2}$ |
| African-Americans after conditioning on top *CD58* and *FCRL3* variants | $1.24 * 10^{-1}$ | $5.83 * 10^{-2}$ |

**Table 2. p-values for association with MS status in different datasets.** European dataset is the MS Chip association signal [17]. African-American dataset is the 2007 subset of the new cohort. The *CD58* variant is rs12025416, and the *FCRL3* variant is rs6681271.

# Discussion:

We sought to determine the specific genetic variants contributing to difference in risk for MS between Africans and Europeans attributable to the chromosome 1 centromere region. Using data from a custom-built genotyping platform designed specifically for studying multiple sclerosis risk including the chromosome 1 admixture association, we replicated the original admixture mapping signal on the centromere of chromosome 1 in the 899 cases that overlapped with a 2007 follow-up of the 2005 study. This demonstrates that the association found in 2005 was not likely to be due to an artifact of genotyping or a technical error in the admixture mapping software. We have not been able to provide an explanation for the post-2007 cases lacking the signal, as the signal in the post-2007 cases is significantly different from the signal in the cases available until 2007.

Nevertheless, we identified two variants that are sufficient to explain the ancestry association signal in the 899 case group, which are associated on a per-allele level to MS in a high-powered study of people of European ancestry. The variants are associated with the *CD58* and *FCRL3* genes [15]. The two variants together would be expected to lead to a 1.44-fold higher rate of MS in European-Americans compared to African-Americans, with the increase primarily driven by the *CD58* variant. The variants found in this study were not the most statistically significant effects found in the most recent genetic map of MS in Europeans [17], but they do reside in the same haplotype in European ancestry reference populations. In African ancestry reference data, the respective LD is notably smaller, driven by allele frequency differences, suggesting that the variants identified in the European admixture scan are tags and not the true variants, and demonstrating the added value for fine-mapping provided by data from people of African ancestry.

Given the current epidemiological uncertainty and disagreement in various studies over the prevalence of MS in European-Americans compared to African-Americans, our findings cannot fully account for reported epidemiological differences in prevalence between individuals of European compared to African ancestry, and indeed it is certainly the case that there are other important genetic and environmental effects that modulate MS prevalence and incidence rates across populations and in changing environments. Nevertheless, it is remarkable that these two genetic variants by themselves, irrespective of other factors, would be sufficient to predict an increased risk in Europeans above that of Africans approximating the range that has been

documented in some epidemiological studies (~1.49-2.27 fold). An admixture mapping study of individuals with MS [23] that used a subset of the samples in this study did not find any signal outside of the MHC region, except a minor one on chromosome 8 in Hispanic individuals. However, most of those data were genotyped without dense coverage near chromosome 1, so it is possible that this study lacked power to find the signal that we have reported here.

Additional fine-mapping of the regions identified in this study will be necessary to determine the true causal SNPs driving the association to MS. This ultimately will require experimental manipulation of candidate SNPs (after narrowing down to a smaller set of candidates *in silico* [24]) in relevant cell lines to determine whether they affect expression of the associated genes (or ones nearby, perhaps through affecting enhancer or transcription factor binding [25]). However, previously published experimental work has already begun to provide insight into the mechanism by which the CD58 gene variant contributes risk for MS. CD58, the protein coded for by the *CD58* gene, is a cell adhesion molecule present on antigen presenting cells (APCs) that binds to CD2 on T cells to both strengthen the adhesion between T cells and APCs and to enhance T cell activation [26]. *CD58* is also expressed in B cells, a key player in MS pathology, with higher levels of expression linked both to enhanced migration to inflammatory sites (for anti-inflammatory activity) and to CD2 ligation [27]. A past study showed that the protective allele of the rs2300474 variant, in strong LD with the protective allele of the rs12025416 variant, increases *CD58* levels [19] (in contrast, the MS susceptibility haplotype, which harbors the risk allele of this variant, decreases CD58 expression). This protective effect is supported by finding that *CD58* mRNA is higher in MS subjects during clinical remission [19]. Engagement of the CD58 receptor, CD2, up-regulates the expression of transcription factor FoxP3 leading to the enhanced function of $CD4^{+}CD25^{high}$ regulatory T cells [19] that are defective in subjects with MS [28]. In this regard, a CD58:IgG1 fusion protein (Alefacept) approved for the treatment of psoriasis has been shown to have agonistic properties [29] that might be useful for MS treatment. Moreover, the rs2300474 variant and rs1335532, both in high LD with rs12025416, have been associated with the autoimmune diseases neuromyelitis optica [30, 31] and primary biliary cholangitis [32]. Lastly, the rs12025416 risk allele found in this study (A/T allele) was found to have higher IL-6 and TNF-alpha macrophage response to *Candida* exposure [33], suggesting that another possible mechanism for increased MS risk may be related to a pro-inflammatory state induced by a stronger TNF-alpha response, possibly through decreased CD58 stimulation on macrophages leading them to take on a pro-inflammatory state, consistent with other MS susceptibility variants showing an enrichment in the TNF-alpha pathway [17, 18].

Surface expression of FCRL3, an immunoglobulin receptor, on B cells has been associated with increased risk for several autoimmune diseases, including systemic lupus erythematosus, autoimmune thyroid disease, Graves' disease and rheumatoid arthritis [34-41], though incongruously, the same allele (which is in high LD with the risk allele of the rs6681271 variant) associated with increased FCRL3 expression and risk for these diseases was found to be associated with protection from MS [42, 43], and the association with SLE was not present in African-American, Korean or European American groups [38, 44-46]. Nevertheless, FCRL3 stimulation via secretory IgA has recently been shown to promote a pro-inflammatory phenotype, in part by inhibiting the suppressive effects of regulatory T cells [47], as previously

shown [48], and the risk allele of the rs6681271 variant has been shown to increase FCRL3 expression [41, 49], consistent with an increased MS risk (assuming the increased expression promotes increased total FCRL3 stimulation on regulatory T cells). Similarly, expression of its homolog FCRL1 was higher in patients with MS [50]. Thus, FCRL3 (or secretory IgA) inhibition represents one potential strategy for an MS therapeutic.

In summary, two variants involved in regulation of immune responses predict a 1.44-fold increased risk of MS in African-Americans with two copies of European ancestry compared to baseline-risk African-Americans. It is likely that other genetic and environmental effects have major impacts on incidence in people of both ancestries, and future work is necessary to determine how these numerous factors interact to lead to the variable prevalence rates of MS observed in different populations today.

# References:


1. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D: **Methods for high-density admixture mapping of disease genes.** *The American Journal of Human Genetics* 2004, **74:**979-1000.
2. Nelson A: *The social life of DNA: Race, reparations, and reconciliation after the genome.* Beacon Press; 2016.
3. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, et al: **Methods for high-density admixture mapping of disease genes.** *Am J Hum Genet* 2004, **74:**979-1000.
4. Seldin MF, Pasaniuc B, Price AL: **New approaches to disease mapping in admixed populations.** *Nat Rev Genet* 2011, **12:**523-528.
5. Kao WH, Klag MJ, Meoni LA, Reich D, Berthier-Schaad Y, Li M, Coresh J, Patterson N, Tandon A, Powe NR, et al: **MYH9 is associated with nondiabetic end-stage renal disease in African Americans.** *Nat Genet* 2008, **40:**1185-1192.
6. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, Waliszewska A, Penney K, Steen RG, Ardlie K, John EM, et al: **Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men.** *Proc Natl Acad Sci U S A* 2006, **103:**14068-14073.
7. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, Waliszewska A, Neubauer J, Tandon A, Schirmer C, McDonald GJ, et al: **Multiple regions within 8q24 independently affect risk for prostate cancer.** *Nat Genet* 2007, **39:**638-644.
8. Kurtzke JF, Beebe GW, Norman JE, Jr.: **Epidemiology of multiple sclerosis in U.S. veterans: 1. Race, sex, and geographic distribution.** *Neurology* 1979, **29:**1228-1235.
9. Wallin MT, Page WF, Kurtzke JF: **Multiple sclerosis in US veterans of the Vietnam era and later military service: race, sex, and geography.** *Ann Neurol* 2004, **55:**65-71.
10. Oh SJ, Calhoun CL: **Multiple sclerosis in the negro.** *Journal of the National Medical Association* 1969, **61:**388.
11. Langer-Gould A, Brara SM, Beaber BE, Zhang JL: **Incidence of multiple sclerosis in multiple racial and ethnic groups.** *Neurology* 2013, **80:**1734-1739.
12. Wallin MT, Culpepper WJ, Coffman P, Pulaski S, Maloni H, Mahan CM, Haselkorn JK, Kurtzke JF, Veterans Affairs Multiple Sclerosis Centres of Excellence Epidemiology G: **The Gulf War era multiple sclerosis cohort: age and incidence rates by race, sex and service.** *Brain* 2012, **135:**1778-1785.
13. Reich D, Patterson N, De Jager PL, McDonald GJ, Waliszewska A, Tandon A, Lincoln RR, DeLoa C, Fruhan SA, Cabre P, et al: **A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility.** *Nat Genet* 2005, **37:**1113-1118.
14. International Multiple Sclerosis Genetics C, Wellcome Trust Case Control C, Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, et al: **Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis.** *Nature* 2011, **476:**214-219.
15. International Multiple Sclerosis Genetics C, Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, Cotsapas C, Shah TS, Spencer C, Booth D, et al: **Analysis of**

**immune-related loci identifies 48 new susceptibility variants for multiple sclerosis.** *Nat Genet* 2013, **45:**1353-1360.

16. Patsopoulos NA, Barcellos LF, Hintzen RQ, Schaefer C, van Duijn CM, Noble JA, Raj T, Imsgc, Anzgene, Gourraud PA, et al: **Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects.** *PLoS Genet* 2013, **9:**e1003926.
17. Patsopoulos NA, Baranzini SE, Santaniello A, Shoostari P, Cotsapas C, Wong G, Beecham AH, James T, Replogle J, Vlachos IS: **Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility.** *Science* 2019, **365**.
18. De Jager PL, Jia X, Wang J, De Bakker PI, Ottoboni L, Aggarwal NT, Piccio L, Raychaudhuri S, Tran D, Aubin C: **Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci.** *Nature genetics* 2009, **41:**776.
19. De Jager PL, Baecher-Allan C, Maier LM, Arthur AT, Ottoboni L, Barcellos L, McCauley JL, Sawcer S, Goris A, Saarela J, et al: **The role of the CD58 locus in multiple sclerosis.** *Proc Natl Acad Sci U S A* 2009, **106:**5264-5269.
20. Isobe N, Madireddy L, Khankhanian P, Matsushita T, Caillier SJ, More JM, Gourraud PA, McCauley JL, Beecham AH, International Multiple Sclerosis Genetics C, et al: **An ImmunoChip study of multiple sclerosis risk in African Americans.** *Brain* 2015, **138:**1518-1530.
21. Reich DP, N.; De Jager, P.L.; Tandon, A.; McCarroll; S.; Waliszewska, A.; Neubauer, J.; Schirmer, C.; Lincoln, R.R.; Poduslo, S.; Khan, O.; Hauser, S.L.; Oksenberg, J.R.; Hafler, D.A.: **Fine mapping of a risk gene for multiple sclerosis.** In *ASHG 2007 Annual Meeting*; 2007.
22. Ramagopalan SV, Knight JC, Ebers GC: **Multiple sclerosis and the major histocompatibility complex.** *Current opinion in neurology* 2009, **22:**219-225.
23. Chi C, Shao X, Rhead B, Gonzales E, Smith JB, Xiang AH, Graves J, Waldman A, Lotze T, Schreiner T: **Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry.** *PLoS genetics* 2019, **15:**e1007808.
24. Schaid DJ, Chen W, Larson NB: **From genome-wide associations to candidate causal variants by statistical fine-mapping.** *Nature Reviews Genetics* 2018, **19:**491-504.
25. Claussnitzer M, Dankel SN, Kim K-H, Quon G, Meuleman W, Haugen C, Glunk V, Sousa IS, Beaudry JL, Puviindran V: **FTO obesity variant circuitry and adipocyte browning in humans.** *New England Journal of Medicine* 2015, **373:**895-907.
26. Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun ZY, Wagner G, Reinherz EL: **Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors.** *Cell* 1999, **97:**791-803.
27. Mitkin NA, Muratova AM, Korneev KV, Pavshintsev VV, Rumyantsev KA, Vagida MS, Uvarova AN, Afanasyeva MA, Schwartz AM, Kuprash DV: **Protective C allele of the single-nucleotide polymorphism rs1335532 is associated with strong binding of Ascl2 transcription factor and elevated CD58 expression in B-cells.** *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 2018, **1864:**3211-3220.

28. Viglietta V, Baecher-Allan C, Weiner HL, Hafler DA: **Loss of functional suppression by CD4+ CD25+ regulatory T cells in patients with multiple sclerosis.** *The Journal of experimental medicine* 2004, **199:**971-979.
29. Haider AS, Lowes MA, Gardner H, Bandaru R, Darabi K, Chamian F, Kikuchi T, Gilleaudeau P, Whalen MS, Cardinale I, et al: **Novel insight into the agonistic mechanism of alefacept in vivo: differentially expressed genes may serve as biomarkers of response in psoriasis patients.** *J Immunol* 2007, **178:**7442-7449.
30. Kim JY, Bae JS, Kim HJ, Shin HD: **CD58 polymorphisms associated with the risk of neuromyelitis optica in a Korean population.** *BMC neurology* 2014, **14:**57.
31. Liu J, Shi Z, Lian Z, Chen H, Zhang Q, Feng H, Miao X, Du Q, Zhou H: **Association of CD58 gene polymorphisms with NMO spectrum disorders in a Han Chinese population.** *Journal of Neuroimmunology* 2017, **309:**23-30.
32. Qiu F, Tang R, Zuo X, Shi X, Wei Y, Zheng X, Dai Y, Gong Y, Wang L, Xu P: **A genome-wide association study identifies six novel risk loci for primary biliary cholangitis.** *Nature communications* 2017, **8:**1-8.
33. Kumar V, Cheng S-C, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, Karjalainen J, Franke L, Withoff S, Plantinga TS: **Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia.** *Nature communications* 2014, **5:**1-8.
34. Kochi Y, Myouzen K, Yamada R, Suzuki A, Kurosaki T, Nakamura Y, Yamamoto K: **FCRL3, an autoimmune susceptibility gene, has inhibitory potential on B-cell receptor-mediated signaling.** *The Journal of Immunology* 2009, **183:**5502-5510.
35. Chistiakov DA, Chistiakov AP: **Is FCRL3 a new general autoimmunity gene?** *Human immunology* 2007, **68:**375-383.
36. Capone M, Bryant JM, Sutkowski N, Haque A: **Fc receptor-like proteins in pathophysiology of B-cell disorder.** *Journal of clinical & cellular immunology* 2016, **7**.
37. Bajpai UD, Swainson LA, Mold JE, Graf JD, Imboden JB, McCune JM: **A functional variant in FCRL3 is associated with higher Fc receptor–like 3 expression on T cell subsets and rheumatoid arthritis disease activity.** *Arthritis & Rheumatism* 2012, **64:**2451-2459.
38. Gibson AW, Li FJ, Wu J, Edberg JC, Su K, Cafardi J, Wiener H, Tiwari H, Kimberly RP, Davis RS: **The FCRL3– 169CT promoter SNP, which is associated with SLE in Japanese, predicts receptor protein expression on CD19+ B cells.** *Arthritis and rheumatism* 2009, **60:**3510.
39. Kochi Y, Yamada R, Suzuki A, Harley JB, Shirasawa S, Sawada T, Bae S-C, Tokuhiro S, Chang X, Sekine A: **A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities.** *Nature genetics* 2005, **37:**478-485.
40. Thalayasingam N, Nair N, Skelton AJ, Massey J, Anderson AE, Clark AD, Diboll J, Lendrem DW, Reynard LN, Cordell HJ: **CD4+ and B lymphocyte expression quantitative traits at rheumatoid arthritis risk loci in patients with untreated early arthritis: implications for causal gene identification.** *Arthritis & Rheumatology* 2018, **70:**361-370.
41. Zhao S-X, Liu W, Zhan M, Song Z-Y, Yang S-Y, Xue L-Q, Pan C-M, Gu Z-H, Liu B-L, Wang H-N: **A refined study of FCRL genes from a genome-wide association study for Graves' disease.** *PloS one* 2013, **8:**e57758.

42. Matesanz F, Fernández O, Milne RL, Fedetz M, Leyva L, Guerrero M, Delgado C, Lucas M, Izquierdo G, Alcina A: **The high producer variant of the Fc-receptor like-3 (FCRL3) gene is involved in protection against multiple sclerosis.** *Journal of neuroimmunology* 2008, **195:**146-150.
43. Martínez A, Mas A, de las Heras V, Bartolomé M, Arroyo R, Fernández-Arquero M, Emilio G, Urcelay E: **FcRL3 and multiple sclerosis pathogenesis: role in autoimmunity?** *Journal of neuroimmunology* 2007, **189:**132-136.
44. You Y, Wang Z, Deng G, Hao F: **Lack of association between Fc receptor-like 3 gene polymorphisms and systemic lupus erythematosus in Chinese population.** *Journal of dermatological science* 2008, **52:**118-122.
45. Sanchez E, Callejas J, Sabio J, Camps M, García-Hernández F, Koeleman B, Martín J, González-Escribano M: **Polymorphisms of the FCRL3 gene in a Spanish population of systemic lupus erythematosus patients.** *Rheumatology (Oxford, England)* 2006, **45:**1044-1046.
46. Choi CB, Kang CP, Seong SS, Bae SC, Kang C: **The− 169C/T polymorphism in FCRL3 is not associated with susceptibility to rheumatoid arthritis or systemic lupus erythematosus in a case–control study of Koreans.** *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology* 2006, **54:**3838-3841.
47. Agarwal S, Kraus Z, Dement-Brown J, Alabi O, Starost K, Tolnay M: **Human Fc receptor-like 3 inhibits regulatory T cell function and binds secretory IgA.** *Cell Reports* 2020, **30:**1292-1299. e1293.
48. Swainson LA, Mold JE, Bajpai UD, McCune JM: **Expression of the autoimmune susceptibility gene FcRL3 on human regulatory T cells is associated with dysfunction and high levels of programmed cell death-1.** *The Journal of Immunology* 2010, **184:**3639-3647.
49. Sasayama D, Hori H, Nakamura S, Miyata R, Teraishi T, Hattori K, Ota M, Yamamoto N, Higuchi T, Amano N: **Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population.** *PLoS One* 2013, **8:**e54967.
50. Baranov KO, Volkova O, Mechetina LV, Chikaev NA, Reshetnikova ES, Nikulina GM, Taranin AV, Naiakshin AM: **[Expression of human B-cell specific receptor FCRL1 in normal individuals and in patients with autoimmune diseases].** *Mol Biol (Mosk)* 2012, **46:**500-507.
51. Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, Dauriz M, Hivert MF, Raghavan S, Lipovich L, et al: **Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility.** *Nat Commun* 2015, **6:**5897.
52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *The American journal of human genetics* 2007, **81:**559-575.
53. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**s13742-13015-10047-13748.

54. Tandon A, Patterson N, Reich D: **Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays.** *Genet Epidemiol* 2011, **35:**80-83.
55. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al: **An integrated map of structural variation in 2,504 human genomes.** *Nature* 2015, **526:**75-81.
56. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J: **10 years of GWAS discovery: biology, function, and translation.** *The American Journal of Human Genetics* 2017, **101:**5-22.

# Materials and Methods:

## Data Sets:

All individuals participating in the study provided full informed consent, and all experimental assays were performed in accordance with the relevant guidelines and regulations as determined by ethical review from the UCSF Human Research Protection Program Institutional Review Board (IRB) (10-05039) Protocol Title: Genetic and non-genetic risk factors for MS. We used data from samples genotyped on a specially designed Multiple Sclerosis SNP array (MS Chip), which included SNPs from the Illumina HumanExome BeadChip [51], ancestry informative markers (AIMS), GWAS catalog SNPs for MS, and additional SNPs near and on the centromere of chromosome 1. We excluded individuals based on low call rate (average of <98%), discrepancies between reported sex and genetically inferred sex, high autosomal heterozygosity (>3.5 standard deviations above the mean), principal components analysis (PCA) outliers, and high relatedness with other samples (PLINK PI_HAT>0.2) [52, 53]. We excluded SNPs based on low call rate (<98%), discordance with plate controls, Hardy-Weinberg disequilibrium ($p<0.00001$), and differential missingness between cases and controls ($p<0.001$). After QC exclusions, there was a total of 2,460 individuals genotyped at 300,287 SNPs (from a start of 2,630 individuals and 321,105 SNPs) available for analysis (1,305 cases and 1,155 controls). We also used summary statistics from MS Chip data [17] in a set of 39,238 European samples from 7 different European populations (European ancestry individuals in Australia, Denmark, Italy, Greece, Sweden, UK, and US for a total of 20,282 cases and 18,956 controls).

## Admixture Mapping:

An AIM panel was constructed based on information from Tandon *et al.* [54] as well as additional SNPs near the centromere of chromosome 1 with high allele frequency differentiation (>50% difference in minor allele frequency) between YRI and CEU populations in the 1000 Genomes Phase 3 data [55]. SNPs were LD-pruned by removing one SNP per pair with $r^2>0.2$ in the CEU and YRI populations and one SNP per pair with genetic distance<0.005 Morgans. SNPs with estimated frequencies that do not match the parental frequencies were removed as suggested by Patterson *et al*. [3] ANCESTRYMAP was used to analyze the data with LOD score for association defined as the ratio of the likelihood of the data under a disease model divided by the likelihood of the data under no disease model. Local and global estimates of ancestry were also obtained with the ANCESTRYMAP software. For most runs,

ANCESTRYMAP was set to have risk = 1.5 and the following parameters: splittau: YES, numburn: 100, numiters: 200, emitter: 30, cleaninit: YES, resitter: 5, with the rest of the parameters set to default. The risk was set to different numbers in the calculation of a 95% confidence interval as described below. Empirical p-values were determined by permuting Case and Control status and running ANCESTRYMAP for each different subset (all cases and all controls, 2007 cases and all controls, non-2007 cases and all controls, and old 2007 cases and controls from the previous genotyping platform) 1,000 times and determining the maximum LOD score across all chromosomes, then comparing these scores with the maximum LOD score of the different data subsets. P-values were calculated as the proportion of times (out of 1,000) that the permuted run produced a maximum LOD score greater than the maximum LOD score of that data subset. In these analyses there is no multiple hypothesis testing so a statistical significance of $p<0.05$ can be used.

**Tests for statistical significance of alleles:**

We ran logistic regressions in R of genotype on case-control status as the main tests of association while controlling for local and genome-wide estimates of European ancestry (calculated within the ANCESTRYMAP software during the admixture mapping runs for these samples) as covariates. For these analyses we used the command glm(CaseControl~Genotype+GlobalEuropeanAncestry+LocalEuropeanAncestry, family=binomial(logit)). We also ran the reverse test regressing local European ancestry onto case-control status and controlling for the genotypes of the top variants from the above association as a covariate (glm(CaseControl~ LocalEuropeanAncestry+GlobalEuropeanAncestry+TopVariant, family=binomial(logit))). For this second analysis we calculated local European ancestry using ANCESTRYMAP using interpolated scores at 1 cM intervals across the admixture association peak. To calculate the 95% confidence intervals, we ran ANCESTRYMAP with the different data subsets of Table 1 at risk models from 0.50 to 2.0 in intervals of 0.01 and found the top LOD score on Chromosome 1 for each model. We then obtained the raw scores as 10 to the power of the (LOD score), normalized the scores to sum to 1 within each data subset, and calculated the confidence intervals as the intervals that contain 95% of the normalized score, with the outer edges each summing to 2.5% (this assumes a prior with equal weight on each of the risk models). LD between different variants ($r^2$) was calculated using PLINK version 1.90 [52, 53] using the --ld command. Power analyses of SNP association in the African-American data were calculated using the calculator provided at: https://github.com/kaustubhad/gwas-power (accessed July 26, 2020) derived from Appendix 1 of Visscher *et al*., 2017 [56]. The power_beta_maf function was used with beta values varying from 0.1 to 0.4, maf (minor allele frequency) varying from 0.05 to 0.5, n=899, and pval=5E-8. The power for the rs12025416 and rs6681271 variants were calculated using odds ratios of 1.37 and 1.07, minor allele frequencies of 0.162 and 0.490, n=899, and pval=5E-8.

**Population Attributable Risk:**

The population attributable risk (PAR) was calculated as 1-1/(total reduced risk), where the total reduced risk was: $\alpha^2*p^2 + 2*\alpha*(p)*(q) + q^2$, where p is the frequency of the variant, q=1-p, and α is the odds ratio of the variant's effect in the population. For two variants, the total reduced risk was calculated as:

$(\alpha_1^2*p_1^2*\alpha_2^2*p_2^2)+(2*\alpha_1^2*p_1^2*\alpha_2*p_2*q_2)+(\alpha_1^2*p_1^2*q_2^2)+(2*\alpha_1*p_1*q_1*\alpha_2^2*p_2^2)+(4*\alpha_1*p_1*q_1*\alpha_2*p_2*q_2)+(2*\alpha_1*p_1*q_1*q_2^2)+(q_1^2*\alpha_2^2*p_2^2)+(2*q_1^2*\ \alpha_2*p_2*q_2)+(q_1^2*q_2^2)$, where the subscripts indicate the variant. The risk for African-Americans was calculated assuming a model with 20% European ancestry and 80% African ancestry (global ancestry proportions estimated in ANCESTRMAP for each individual are provided in Supplementary Online Table 1).

# Declarations

**Ethics Approval and Consent to Participate**

All individuals provided full informed consent and ethical review was provided by the UCSF Human Research Protection Program Institutional Review Board (IRB) (10-05039) Protocol Title: Genetic and non-genetic risk factors for MS.

**Consent for publication**

Not applicable.

**Availability of Data:**

All data analyzed in this article are available upon request for MS research only per the informed consent and IRB approval (contact Jorge Oksenberg: Jorge.Oksenberg@ucsf.edu).

**Competing interests:**

The authors declare that they have no competing interests.

**Funding:**

Funding was provided by an NIGMS (GM007753) fellowship to NN. DR is an Investigator of the Howard Hughes Medical Institute and this work was supported by NIH grant HG006399 supporting methods for disease-gene mapping in multi-ethnic populations, and grants from the NIH (NS046630), the Wadsworth Foundation, and the National Multiple Sclerosis Society (NMSS) to support the identification of multiple sclerosis risk variants in African-Americans. N.A.P was supported by the NMSS (RG-1707-28657 and JF-1808-32223).

**Authors' Contributions:**

N.N., N.Patt., P.D.J., S.H., D.A.H., J.O. and D.R. conceived the study. N.N., N.Patt., N.A., N.I., A.T., A.B., J.M., N.A.P., and D.R.. performed analysis. N.N. and D.R. wrote the manuscript with the help of all co-authors.

**Acknowledgements:**

We thank Alkes Price for helpful discussions. We thank the Biorepository Facility and the Center for Genome Technology laboratory personnel (specifically Patrice Whitehead and Anna Konidari) within the John P. Hussman Institute for Human Genomics at the University of Miami for centralized DNA handling and genotyping of the MS Chip array for this project. We are deeply grateful to the IMSGC (International Multiple Sclerosis Genetics Consortium) for generating the data and providing access to it.

# Chapter 4: Estimating contamination in ancient nuclear DNA using linkage disequilibrium

This section is based on the following paper I contributed to during my PhD:

> **Nakatsuka, N.***; Harney, E.*; Mallick, S..; Mah, M.; Patterson, N.; Reich, D. "Estimation of Ancient Nuclear DNA Contamination Using Breakdown of Linkage Disequilibrium." *Genome Biology* (Software). 2020 Aug. 10; 21:199. https://doi.org/10.1186/s13059-020-02111-2.

**Overview:**

Chapters 2 and 3 focused on analyses of present-day DNA, while the rest of this thesis focuses on ancient DNA (aDNA) analyses. One important aspect of aDNA studies is proper quality control, especially to determine if the DNA is truly from the ancient individual of interest rather than from contamination. Apart from *in silico* tools, it is usually easier to determine if a sample of present-day DNA has been contaminated, because one can go back to the original individual (or to the immortalized cell lines) and compare with a new analysis. In contrast, there is usually a more limited supply of the ancient skeletal material, and this material itself might already be contaminated. This chapter focuses on an *in silico* tool we developed to estimate the level of contamination in aDNA samples.

**Ethics:**

All samples used in this study were previously published data, and the statements made in this study were almost solely based on the level of contamination in the sample rather than anything in particular about the ancient individuals themselves. One could argue that some of the samples might have come from studies that did not fully follow the ethical framework of Chapter 1.2, but in these past studies there did not appear to be a breach of any previously established guidelines and there was not (to our knowledge) any widespread opposition by the relevant communities to the use of the skeletal material in the past studies of the samples used in this work. Thus, this study's use of the previously published samples is in line with the general framework of Chapter 1.2, especially with regards to non-maleficence (due to its focus on the sample characteristics rather than statements about the ancient individuals themselves) and beneficence (by providing a tool that can aid the field in preventing inaccurate population genetic inferences).

# ContamLD: Estimation of Ancient Nuclear DNA Contamination Using Breakdown of Linkage Disequilibrium

Nathan Nakatsuka[1,2,3,*,†], Éadaoin Harney[1,3,4,*,†], Swapan Mallick[1,3], Matthew Mah[1,3], Nick Patterson[3], and David Reich[1,3,5,6,†]

[1]Department of Genetics, Harvard Medical School, New Research Building, 77 Ave. Louis Pasteur, Boston, MA 02115, USA
[2]Harvard-MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[3]Department of Human Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA
[4]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave., Cambridge, MA 02138, USA
[5]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA 02141, USA
[6]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA

*co-first authors

[†] Corresponding authors: Nathan Nakatsuka (nathan_nakatsuka@hms.harvard.edu), Éadaoin Harney (harney@g.harvard.edu), and David Reich (reich@genetics.med.harvard.edu)

**Keywords:** Ancient DNA, linkage disequilibrium, contamination, autosomal DNA, nuclear DNA

**Supplementary Material:** All supplementary material can be found in the supplement of Nakatsuka *et al*., 2020 *BioRxiv*.

# Abstract

We report a method, *ContamLD*, for estimating autosomal ancient DNA (aDNA) contamination by measuring the breakdown of linkage disequilibrium in a sequenced individual due to the introduction of contaminant DNA, leveraging the idea that contaminants should have haplotypes uncorrelated to those of the studied individual. Using simulated data, we confirm that *ContamLD* accurately infers contamination rates with low standard errors (e.g. less than 1.5% standard error in cases with <10% contamination and data from at least 500,000 sequences covering SNPs). This method is optimized for application to aDNA, leveraging characteristic aDNA damage patterns to provide calibrated contamination estimates. Availability: https://github.com/nathan-nakatsuka/ContamLD.


# Background

Ancient DNA (aDNA) data has emerged as a powerful tool for learning about ancient population history, allowing direct study of the genomes of individuals who lived thousands of years in the past [1-3]. Unfortunately, these inferences can be distorted by contamination during the excavation and storage of skeletal material, as well as the intensive processing required to extract the DNA and convert it into a form that can be sequenced.

Accurate measurement of the proportion of contamination in ancient DNA data is important, because it can provide guidance about whether analysis should be restricted to sequences that show the characteristic pattern of C-to-T mismatch to the reference genome of authentic aDNA (if contamination is high) [4], or carried out at all. When analysis is restricted to focus only on sequences showing evidence of characteristic ancient DNA damage, the substantial majority of authentic sequences are usually removed from the analysis dataset, as only a fraction of genuinely ancient sequences typically carry characteristic damage. In addition, if a sample is contaminated by another individual with damaged DNA—which can arise for example as a result of cross-contamination from other specimens handled in the same ancient DNA laboratory—it is impossible to distinguish authentic sequences from contaminating ones based on the presence or absence of characteristic ancient DNA damage.

Current methods for estimating contamination have significant limitations. Methods based on testing for heterogeneity in mitochondrial DNA sequences (which are almost always homogeneous in an uncontaminated individual) can be biased, because there are several orders of magnitude of variation in the ratio of the mitochondrial to nuclear DNA copy number across samples. Thus, samples that have evidence of mitochondrial contamination can be nearly uncontaminated in their nuclear DNA, while samples that have no evidence of mitochondrial contamination can have high nuclear contamination [5]. Another reliable set of methods for estimating rates of contamination in ancient DNA leverage polymorphism on the X chromosome in males, including the popular *ANGSD* method [6-9] and an improved methodology that enhances power for low-coverage samples [10]. However, these methods do not work in females.

Several methods for estimating contamination rates in nuclear DNA from modern genomes have been published, including *ContEst* [11] and *ContaminationDetection* [12]. However, these methods generally rely on access to uncontaminated genotype data from the

individual of interest or access to all possible contaminating individuals, neither of which is typically available for aDNA. Another method estimated modern human autosomal contamination in aDNA from archaic Denisovans [13] and Neanderthals [14] by producing maximum likelihood co-estimation of sequence error, contamination, and parameters correlated with divergence and heterozygosity. However, this method heavily relies on the significant divergence between archaic and modern humans. A similar method, *DICE*, expanded on this method and jointly estimates contamination rate and error rate along with demographic history based on allele frequency correlation patterns [15]. However, this method requires both explicit demographic modeling and high genome coverage. While this may be effective for estimation of contamination in archaic genomes like Neanderthals and Denisovans that are highly genetically diverged from likely contaminant individuals, it is not optimized for study of contamination among closely related present-day human groups with complex demographic relationships relative to each other, or contamination from individuals of the same population. In Racimo *et al.* 2016 [15], *DICE* required over 3x genome sequence coverage and solved the distinctive problem of measuring contamination of present-day humans in a Neanderthal genome.

We report a method for estimating autosomal aDNA contamination using patterns of linkage disequilibrium (LD) within a sample. This approach, implemented in our software *ContamLD*, is based on the idea that when sequences from one or more contaminating individuals are present in a sample, LD among sequences derived from that sample is expected to be diminished, because the contaminant DNA derives from different haplotypes and therefore should have no LD with the authentic DNA of the ancient individual of interest. Thus, the goal of the algorithm is to determine the LD pattern the ancient individual would have had without contamination and compare it to the LD pattern found in the sample. The LD patterns of ancient individuals are determined using reference panels from 1000 Genomes Project populations to compute approximate background haplotype frequencies, where haplotypes are defined as pairs of SNPs with high correlation to each other. Contamination is then estimated by fitting a maximum likelihood model of a mixture of haplotypes from an uncontaminated individual and a proportion of contamination (to be estimated from the data) from an unrelated individual. *ContamLD* corrects for mismatch of the ancestry of the ancient individual with the reference panels using two different user-specified options. In the first option, mismatch is corrected using estimates from damaged sequences (which, ideally, lack present-day contaminants). In the second option, *ContamLD* performs an "external" correction by subtracting the sample's contamination estimate from estimates for individuals of the same population believed to have negligible contamination (the user could obtain this value from a *ContamLD* calculation on a male individual with a very low estimate of contamination based on *ANGSD*). The second option has more power than the first option and allows detection of cross-contamination by other ancient samples, but it could be biased if a reliable estimate from an un-contaminated individual from the same population is not available for the external correction.

We show that *ContamLD* accurately infers contamination in both ancient and present-day individuals of widely divergent ancestries with simulated contamination coming from individuals of different ancestries. The contamination estimates are highly correlated with estimates based on X chromosome analysis in ancient samples that are male, as assessed using

*ANGSD* [16]. *ContamLD* run with the first option has standard errors less than 1.5% in samples with at least 500,000 sequences covering SNPs (~0.5x coverage for data produced by in-solution enrichment for ~1.2 million SNPs [2, 17], or ~0.1x coverage for data produced using whole-genome shotgun sequences). With the second option, *ContamLD* has standard errors less than 0.5% in these situations, allowing users to detect samples with 5% or more contamination with high confidence so they can be removed from subsequent analyses.

# Results

*Simulations of Contamination in Present-Day Individuals:*

To test the performance of *ContamLD*, we simulated sequence level genetic data. For our first simulations, each uncontaminated individual was simulated based on genotype calls from a present-day individual from the 1000 Genomes Project dataset. To determine the sequence coverage at each site, we used data from an ancient individual for which we had data at 1.02x coverage and in each case generated the same number of simulated sequences at each site, with the allele drawn from the present-day individual e.g. if the present day individual is homozygous for the reference allele at a site, all simulated alleles are of the reference type, while if the present day individual is heterozygous, simulated alleles are either of the reference or alternative variant, with 50% probability of each). The damage status (i.e. whether it carries the characteristic C-to-T damage often observed in ancient DNA sequences) of each sequence was also determined based on the status of the ancient reference individual. Contaminating sequences were then "spiked-in" at varying proportions (0 to 40%), using an additional present-day individual from the 1000 Genomes Project to determine the contaminating allele type (see Methods). All contaminating sequences were defined to be undamaged, as would be expected if the contamination came from a non-ancient source.

For most of the analyses reported in this study, we simulate data for SNP sites targeted in the 1.24 million SNP capture reagent [2, 17] that intersect with 1000 Genomes sites, after removing sites on the X and Y chromosomes (this leaves ~1.1 million SNPs). The *ContamLD* software also allows users to make panels based on their own SNP sets, and in a later section we report results from a larger panel (~5.6 million SNPs) provided with the software that we recommend for shotgun sequenced samples, and which provides more power to measure contamination.

We first analyzed data generated using a reference individual from the 1000 Genomes CEU population (Utah Residents (CEPH) with Northern and Western European Ancestry) and the SNP coverage profile of a 1.02x coverage ancient individual of West Eurasian ancestry (Iberian Bronze individual I3756 who lived 2014-1781 calBCE; see Methods). Supplementary Figure 1 illustrates the distribution of LOD (logarithm of the odds) scores generated when *ContamLD* is run on samples with 0%, 7% and 15% simulated contamination. Supplementary Figure 2 shows the contamination rate estimates generated for data with simulated contamination rates between 0 to 40%. At very high contamination (above 15%) *ContamLD* often overestimates contamination, but in practice samples with above 10% contamination are generally removed from population genetic analyses, so inaccuracies in the estimates at these levels are not a concern in our view (the importance of a contamination estimate in many cases is to flag problematic samples, not to accurately estimate the contamination proportion).

*ContamLD* assumes that the individual making up the majority of the sequences is the base individual, so we do not explore contamination rates greater than 50% in these simulation studies.

We observe a linear shift in the contamination estimates such that most estimates are biased to be slightly higher than the actual value, with even greater overestimates occurring at higher contamination rates (Supplementary Figure 2). This is likely due to the difference between the haplotype distribution of the test individual and that of the haplotype panel, since the magnitude of this shift increases as the test individual increases in genetic distance from the haplotype panel. Even in cases where the test individual is of the same ancestry as the haplotype panel (as in Supplementary Figure 2) there is expected to be a shift, because the test individual's haplotypes are a particular sampling of the population's haplotypes, and the difference between having only frequencies of the haplotype panel and a particular instantiation of those frequencies in the test individual will lead to the artificial need for an external source ("contaminant") to fit the model properly.

In contrast to the upward bias in contamination estimates due to mismatch of the individual's haplotypes with the reference panel haplotype frequencies, we observe negative shifts for inbred individuals, as expected because *ContamLD* assumes the paternal and maternal copy of a chromosome are unrelated. In contrast, if the two chromosomes are related, extra LD will be induced and more contamination will be necessary to produce the expected LD pattern. In principle, this inbreeding effect could be corrected explicitly by estimating the total amount of ROH in each individual and applying this as a correction, although we do not provide such functionality as part of our software. If a reliable methodology for quantifying the proportion of the genome that is affected by inbreeding in ancient individuals becomes available, *ContamLD* could be further improved by using this information as an input parameter.

A final type of bias could be expected to arise if the contamination comes from an individual related to the target individual. In this case the true contamination rate is expected to be under-estimated, because *ContamLD* only detects contamination where the contaminant sequence differs from the target individual's sequence. If the contaminant carries the same haplotypes as the target individual, in the most extreme case as expected for an identical twin, then the existence of contamination will be missed altogether. In general, contamination from closely related individuals is unlikely to be a concern for many population genetic analyses, as close relatives usually (but with important exceptions) have very similar ancestry.

In our implementation, we correct for these systematic biases in two ways, implemented as different options in *ContamLD*.

The first option leverages information from sequences that contain evidence of the C-to-T damage that is characteristic of ancient sequences. This option assumes these sequences are authentically ancient and not derived from a contaminating source (assumed to be from present-day individuals), so the *ContamLD* estimate based on undamaged sequences is corrected by estimates based on the damaged sequences (see Methods for more details). In the second option, we allow the user to subtract the contamination estimate from the estimate of an individual of the same ancestry assumed to be uncontaminated. An advantage of the second option compared to the first is that it has smaller standard errors (Figure 1), reflecting the fact that it does not rely on estimates from damaged sequences (reliance on damaged sequences reduces power since it often reflects a very small subset of the data). A second advantage of

the second option is that it allows estimation of contamination in cases where the source of contamination is also ancient in origin, as would be expected if the contamination occurred anciently or due to cross contamination with other ancient samples (the first option would be expected to produce an underestimate of contamination in such cases, since it assumes that sequences that contain C-to-T damage are not contaminated). On the other hand, a drawback of the second option is that it requires users to identify a relatively high coverage, uncontaminated, ancestry-matched sampled for benchmarking purposes; the method is also only expected to work if there is minimal inbreeding in either the sample of interest or the matched sample. Identifying such benchmarking samples may be impossible when analyzing samples from previously unsampled contexts (e.g. early modern humans), and indeed verifying that a benchmarking sample is uncontaminated is very difficult if it is female (if it is male a method like *ANGSD* can be used).

In what follows, we report results of analyses based on the first option, but *ContamLD* includes both methods as options. The uncorrected score also forms the basis for warning output by the software, namely high contamination or possible contamination with another ancient sample leading to an inaccurate damage correction estimate.

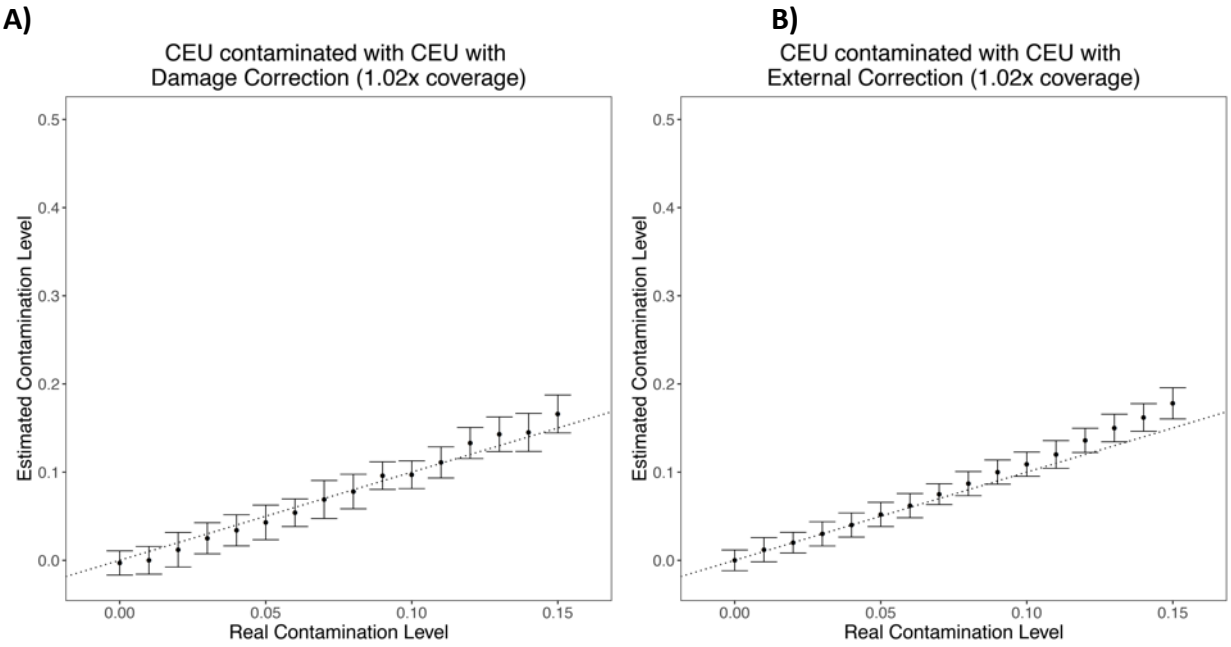


**Figure 1. *ContamLD* estimates when the target individual, contaminant, and haplotype panel are from the CEU population.** Contamination estimates when the simulated contamination rate is between 0.00-0.15. **A)** Estimates with damage restricted correction (option 1). **B)** Estimates with external correction from an uncontaminated sample (option 2). The black dotted line is y=x, which would correspond to a perfect estimate of contamination. Error bars are 1.96*standard error (95% confidence interval determined via jackknife resampling across chromosomes).

*Simulated Contamination of Ancient Samples with Present-Day Samples:*

*ContamLD* is designed to work on ancient individuals, so we simulated contamination of real ancient individuals with present-day individuals from the 1000 Genomes Project, a scenario

that would occur when skeletal material from ancient individuals is contaminated by present-day individuals during excavation or at some point during the processing of the material. We used data from male individuals selected due to very low X chromosome contamination estimates (less than 1%) based on *ANGSD* [16] (developed first in Rasmussen *et al*. [9]; we used method 1 of that software). (We subtracted the *ANGSD* estimates from the *ContamLD* estimates to correct for any residual contamination.) Figure 2A shows results from the Iberian Bronze Age sample [18] (I3756) with 1.02x coverage at the targeted ~1.24 million SNP positions, demonstrating that *ContamLD* produces highly accurate contamination estimates for this simulation.

*Effect of Different Haplotype Panels*

There are many potential cases in which ancient individuals can come from populations with very different genetic profiles compared to present-day 1000 Genomes populations, leading to an ancestry mismatch to the haplotype reference panels. *ContamLD* provides panels from all 1000 Genomes populations as well as tools to identify the panel most closely matching to the ancestry of their ancient individual (based on outgroup-$f_3$ statistics [19] to determine the most shared genetic drift), which they can then select for the analysis. However, due to the potential for ancestry mismatch to still occur, we tested the effect of choosing haplotype panels that are genetically diverged from the individual of interest (Figure 2A). For the ancient Iberian sample, the CEU and TSI (Toscani in Italia) panels—representing northern and southern European ancestry, respectively—yielded contamination estimates that are close to the true contamination rate, especially for rates below 5%. However, *ContamLD* underestimates contamination by ~2% when the CHB (Han Chinese in Beijing, China) and YRI (Yoruba in Ibadan, Nigeria) panels were used instead (though we view these as very pessimistic cases, because the user should usually be able to choose a panel more closely related to their ancient individual than these scenarios). We thus recommend that users take care to choose an appropriate panel that is within the same continental ancestry as their ancient individual. Nevertheless, we note that we were able to obtain reasonably accurate estimates for Upper Paleolithic European hunter-gatherers, such as the Kostenki14 individual [20], who is ~37,470 years old, even when using present-day European panels that have significantly different ancestry from the hunter-gatherers (Supplementary Figure 3).

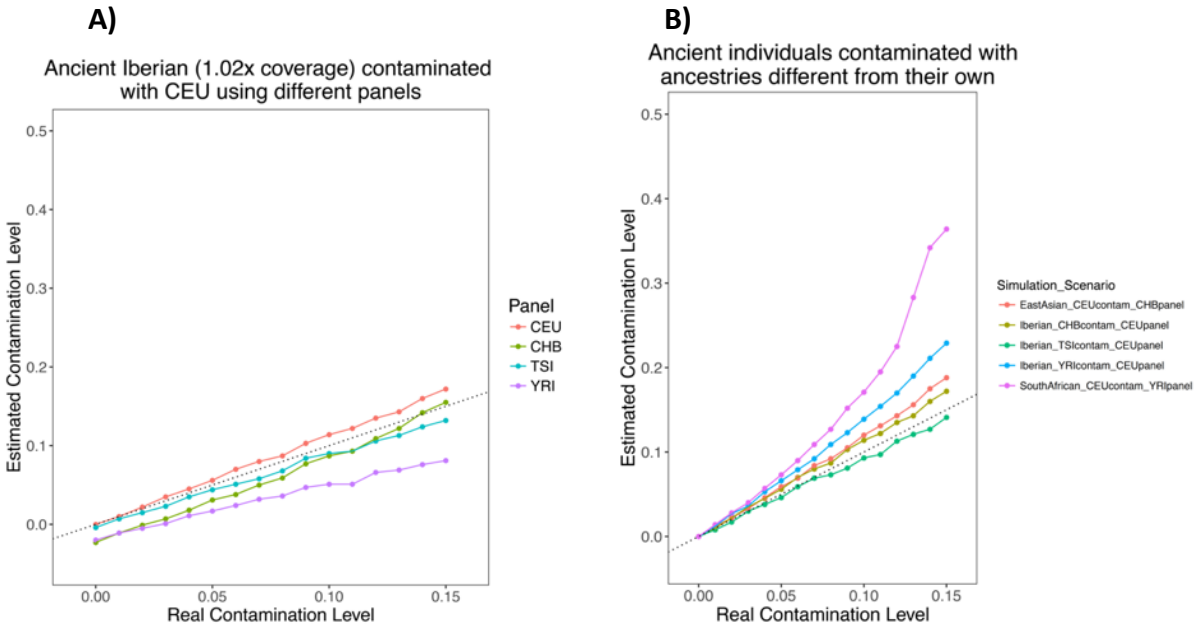


**Figure 2. Genetic divergence between uncontaminated individual and contamination sources or haplotype panels impacts *ContamLD* estimates. A)** Ancient Iberian (I3756, 1.02x coverage) contaminated with CEU with haplotype panels generated from CEU, TSI, CHB, and YRI populations. **B)** Contamination estimates from the same ancient Iberian contaminated with TSI, CHB, or YRI and analyzed with a CEU panel; from an ancient East Asian (DA362.SG, 1.10x coverage) contaminated with CEU and analyzed with a CHB panel; and from an ancient South African (I9028.SG, 1.21x coverage) contaminated with CEU and analyzed with a YRI panel. The black dotted line is y=x, corresponding to a perfect estimate of contamination. All estimates use the damage restricted correction (option 1).

*Effect of Mismatch Between the Ancestry of the True Sample and Contaminating Individual*

Contamination can come from a wide variety of sources, including, but not limited to, members of the archaeological excavation team, the aDNA laboratory, or residual human DNA on the plastic and glassware or in laboratory reagents. Thus, we sought to understand the effect of mismatch in the ancestry of the true sample and the contaminating individual in our contamination estimates. We found that as the ancestry of the two diverged, *ContamLD* over-estimated contamination (Figure 2B and Supplementary Figure 4). This occurred when we tested an ancient European with different contaminant ancestries and when we tested ancient East Asian [21] and ancient South African [22] samples contaminated with European DNA. Nevertheless, the over-estimation was not severe at contamination levels below 5 percent, and samples above this proportion would likely be flagged as problematic. We also explored scenarios where the ancestry of the panel matches the contaminant rather than the true sample (Supplementary Figure 4) and found a ~2% under-estimate at low levels of contamination and an over-estimate at high levels of contamination; these are modest effects and are unlikely to change our qualitative assessment. When we tested the effect of having multiple contaminant individuals (Supplementary Figure 5), we found only a slight over-estimate at higher levels of contamination, as expected given *ContamLD* normally assumes contamination from a single individual where the haplotypes are re-formed if they are created

from two contaminant reads (which will happen at lower rates with more contaminant individuals).

*Estimating Contamination in Admixed Individuals*

*ContamLD* relies on measuring the difference between the LD pattern of the sample and that expected from an uncontaminated individual. However, individuals from groups recently admixed between two highly divergent ancestral groups have LD patterns that are similar in some ways to that of an unadmixed individual with contamination from a group with ancestry diverged from that of the individual of interest. To understand how this would impact *ContamLD*, we ran the software on an ASW (Americans of African Ancestry in Southwest USA) individual with different levels of added CEU contamination. When we ran *ContamLD* with a YRI panel and no correction on an individual with no contamination, the individual was inferred to have a contamination of ~20% (likely because the individual had ~15% European ancestry, and this was interpreted by the software as contamination). Using an ASW panel did not perform any better. However, the concerns were mostly addressed by the damage-restricted correction (option 1) at low contamination levels (Supplementary Figure 6). The simulation with African-Americans represents an extreme of difficulty, because the individual is from a group with very recent admixture (~6 generations [23]) of ancestries highly divergent from each other with one of the ancestries very genetically similar to the reference panel. It highlights how the damage-restricted correction is still able to produce accurate estimates in these difficult cases.

*Effect of Coverage:*

We tested the power of our procedure at different coverages with simulations of ancient West Eurasian ancestry individuals contaminated with CEU on the 1240K SNP set (Figure 3). We found that while our estimates were not biased to produce estimates consistently above or below the true value, the standard errors increased significantly at lower coverages, as expected for the decreased power for accurate estimation in these scenarios. We provide a much larger panel with ~5.6 million SNPs (vs. ~1.1 million for the 1240K panel) that usually decreases standard errors for samples that are shotgun sequenced (Supplementary Figure 7). This panel increases *ContamLD*'s compute time and memory requirements, so we recommend that it only be used for individuals with lower than 0.5x coverage. As an additional feature, we provide users tools to create their own panels to meet their specific needs.

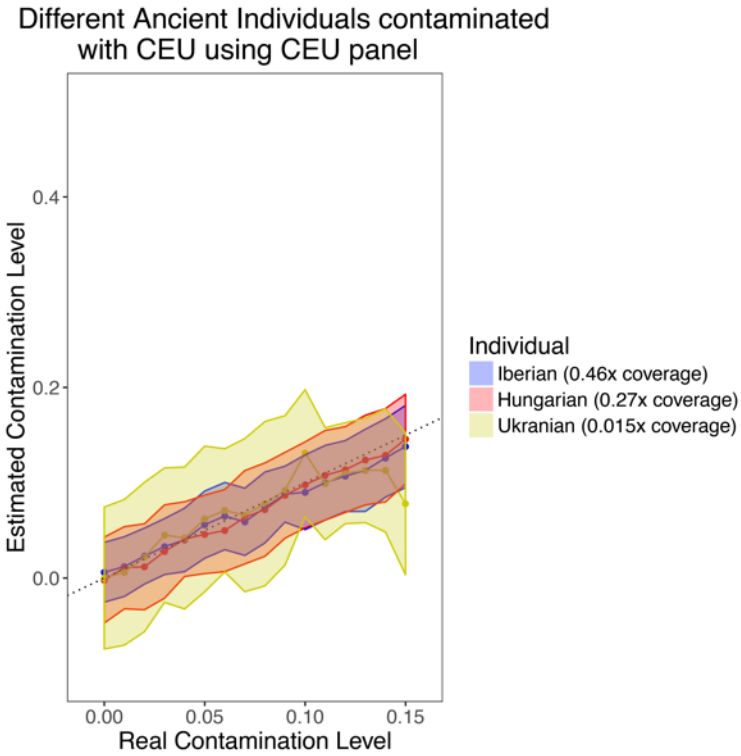


**Figure 3. *ContamLD* estimates for ancient European samples of different coverages after damage restricted correction (option 1).** An ancient Iberian of 0.46x coverage, an ancient Hungarian of 0.27x coverage, and an ancient Ukranian of 0.015x coverage (~16,000 snps) were contaminated with CEU and analyzed using a CEU panel with *ContamLD* option 1 (damage restricted correction). The black dotted line is y=x. Error shading is 1.96*standard error (95% confidence interval).

*Simulations to Compare ContamLD to ANGSD X Chromosome Estimates*

We performed simulations where we randomly added sequences at increasing levels from 0 to 15% from an ancient West Eurasian individual (I10895) into the BAM files of 65 ancient male individuals of variable ancestries and ages (we set the damaged sequences to be only from the non-contaminant individual; see Methods). We chose ancient male individuals that had average coverage over 0.5X and X chromosome contamination estimates under 2% (using method 1 of *ANGSD*) when no artificial contamination was added (and also corrected even for this baseline contamination by setting damaged reads to be a 5% down-sampling of the files that had no artificial contamination; see Methods). We then analyzed the individuals with *ContamLD* and *ANGSD* and found that compared to *ANGSD*, *ContamLD* consistently had similar errors relative to the true contamination level (Figure 4, Supplementary Online Table 2).

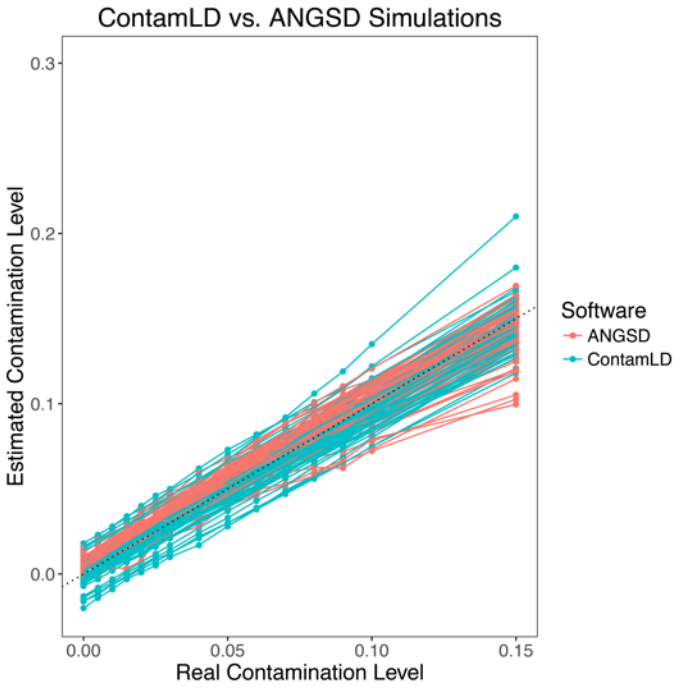


**Figure 4. Contamination estimates with *ContamLD* and *ANGSD* for ancient individuals with different levels of contamination added.** 65 ancient individuals with average coverage over 0.5X had increasing levels of artificial contamination added in (from I10895, an ~1200BP ancient West Eurasian individual) and were then analyzed with *ContamLD* (with panels most genetically similar to the ancient individual and using damage restricted correction, option 1) and *ANGSD*. Details of all estimates (including standard errors) are provided in Supplementary Online Table 2. The black dotted line is y=x, which would correspond to a perfect estimate of the contamination.

*Comparing ContamLD, ANGSD, and Mitochondrial Estimates (ContamMix) in Ancient Individuals without Added Contamination*

We tested 439 ancient males with *ContamLD*, *ANGSD* (X chromosome contamination estimates), and *ContamMix* (mitochondrial contamination estimates) without adding additional contamination. For this analysis, we included published data generated with the ~1.24 million SNP enrichment reagent, as well as data from libraries that failed quality control due to evidence of contamination (Supplementary Online Table 3). Similar to prior studies [5], the mitochondrial estimates often differed from the nuclear (*ANGSD* and *ContamLD*) estimates, showing high contamination in some libraries with low nuclear contamination, and low mitochondrial contamination in some libraries with high nuclear contamination (Figure 5A). In contrast, *ANGSD* and *ContamLD* had better concordance. However, we observed that some of the samples with high contamination estimates based on *ANGSD* had much lower *ContamLD* estimates, reflecting over-correction from analyzing the damaged sequences, perhaps because the contamination was actually cross-contamination from other ancient individuals, violating the assumptions of our damage-correction (Figure 5B). This problem was mitigated in part, however, because *ContamLD* produces a warning of "Very_High_Contamination" if the uncorrected estimate is above 15% (even in cases where the corrected estimate is very low), and all samples with X chromosome estimates over 5% were flagged with this warning and/or had estimates of over 5% contamination with *ContamLD* (all samples with less than 5% contamination in *ANGSD* had lower than 5% contamination with *ContamLD*). It is unfortunately

not possible to know the true contamination of the samples we tested in Figure 5, but the fact that our software produced results with good correlation to X chromosome estimates shows that it works well in real ancient data.

It is possible for there to be samples with moderately high contamination from another ancient individual but both a low damage restricted correction estimate and no warning generated, because these would have high uncorrected estimates, yet not high enough to reach the threshold required for the warning. These samples would have to be identified with an external correction. Lowering the threshold for the "Very_High_Contamination" warning would produce too many false positives, because there are many cases with high uncorrected estimates that have low corrected estimates that are likely not contaminated (e.g. due to ancestry mismatches of the panel and the test individual). To understand these issues better, we performed a simulation in which an ancient Iberian (I3756) was contaminated with another ancient West Eurasian individual (I10895) and the damaged sequences were set to be a 5% down-sampling of the set of contaminated sequences (thus simulating a case in which all of the contamination is from another ancient individual who has the same damage proportion as the ancient individual of interest). We found that, as expected, the contamination from the ancient individual was not detected (the contamination estimates were always near 0%) by the damage restricted correction version of *ContamLD* until the contamination reached 15% at which point the "Very_High_Contamination" flag came up (Supplementary Figure 9). The contamination would have been detected with the external correction version of *ContamLD* (since the damage restricted correction continued to go up with increasing contamination; see Supplementary Online Table 4), but without an uncontaminated ancient individual of the same group as the target individual, this would be difficult to do without bias in the contamination estimate.

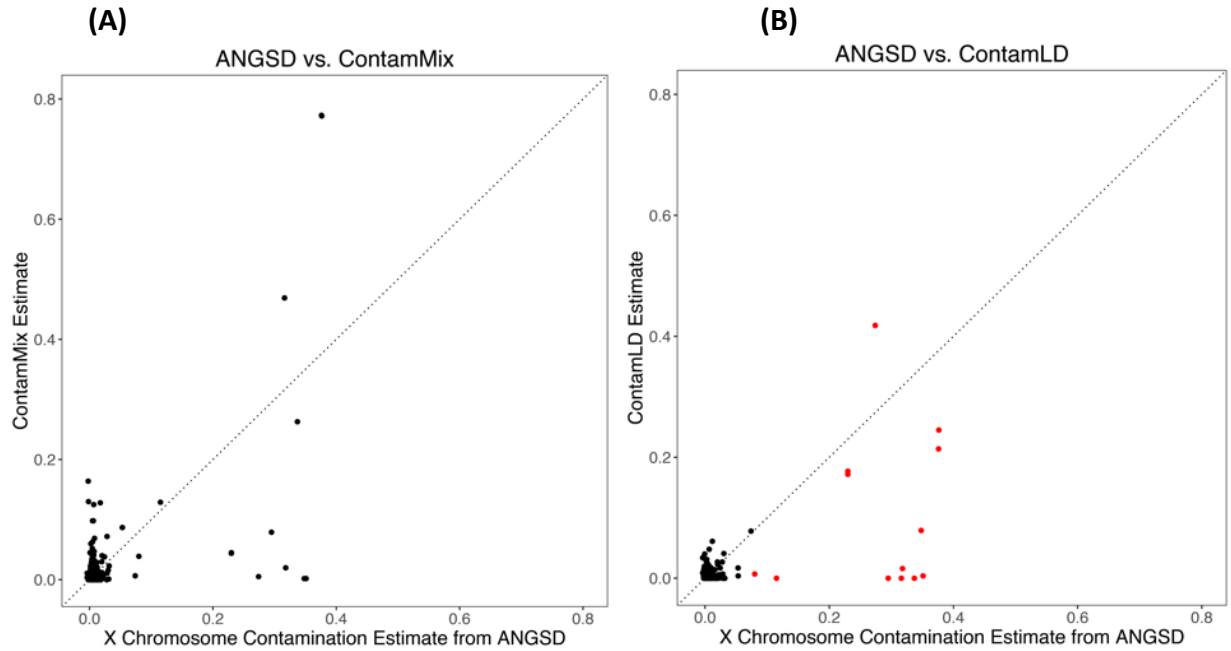


**Figure 5. Contamination estimates from *ContamLD*, *ANGSD*, and *ContamMix* in 439 ancient individuals of variable ancestry.** *ANGSD* estimates (method 1) are plotted on the X-axis, and on the Y-axis are either **(A)** *ContamMix* or **(B)** *ContamLD* estimates. In red are samples that were flagged in *ContamLD* as "Very_High_Contamination" based on having uncorrected estimates over 15%. All *ContamLD* estimates below 0 were set to 0.

# Discussion and Conclusion

We have presented a tool, *ContamLD*, for estimating rates of autosomal DNA contamination in aDNA samples. *ContamLD* is able to measure contamination accurately in both male and female individuals, with standard errors less than 1.5% for individuals with coverage above 0.5X on the 1240K SNP set (for contamination levels less than 10%) for the damage restricted correction method (option 1). On the shotgun panel we provide, standard errors are less than 1.5% for coverages above 0.1x. *ContamLD* is best suited to scenarios in which the contaminant and the ancient individual of interest are similar ancestry, which is useful, because *DICE* [15] and many population genetic tools (e.g. PCA or ADMIXTURE [24]) are better suited for detecting cases where the contaminant is of very different ancestry from the ancient individual of interest. *ContamLD* works even for recently admixed individuals. Lastly, *ContamLD* can detect cases of contamination from other ancient individuals, though this works best if it is large amounts of contamination that can reach the threshold required for the "Very_High_Contamination" flag.

We tested *ContamLD* in multiple simulation scenarios to determine when bias or less reliable results could be expected. When applied to the situation with a test individual (ancient or present-day), contaminant, and haplotype reference panel all from the same continental ancestry, *ContamLD* provides an accurate, unbiased estimate of contamination. When the contaminant comes from a population that is of a different continental ancestry from the population used for the base and haplotype panel, the contamination appears to be slightly overestimated, particularly for higher contamination. This should not be a large problem in analyses of real (i.e. non-simulated) data, because the effect is small at the contamination levels of interest (<5%). When we varied haplotype panels, we found that the estimator is robust when applied to simulated datasets using haplotype panels that are moderately divergent from the base sample (within-continent levels of variation). We provide users tools for automatically determining the panel that shared the most genetic drift with the sample so that the user can select the panel most closely related to the sample. In other simulations, we found that the performance of the algorithm declines as the coverage of the sample decreases. The estimates are not biased, the standard errors substantially increase when fewer than 300,000 sequences are available. In these cases, if the individual was shotgun sequenced, we recommend that users choose the shotgun panel, which will substantially increase power for the analyses.

We applied the algorithm to estimate contamination levels in dozens of ancient samples and compared them to X chromosome-based contamination estimates. There was generally good correlation with the X chromosome estimates, except that when the true contamination was very high, the LD based estimates were sometimes estimated incorrectly, likely because the contamination was due to cross-contamination from another ancient individual and there was over-correction from the damage estimates. This problem is mitigated, however, because the software indicates if the uncorrected estimate is very high so users can identify highly contaminated samples and remove them from further analyses. A difficult case for the software is if there is contamination in part from another ancient sample. This can cause an over-correction and lead to an under-estimate of the contamination. The

“Very_High_Contamination” warning catches very high contamination from other ancient samples, but it will miss cases of moderate levels of contamination from other ancient samples, because it will not reach the threshold required for the warning. In theory, the user can determine the true contamination in these cases using the external correction, but the external correction can be difficult if the user does not have an adequate sample to correct the estimate of the sample of interest. The damage correction of the software also does not work if the samples have undergone full UDG treatment (no damaged sequences), and for this case, the external correction is the only option.

The software run-time is dependent on SNP coverage. If ~1,000,000 SNPs are covered (the depth of the coverage on each SNP does not affect run-time), the analysis takes approximately 2 hours if 3 cores are available on CentOS 7.2.15 Linux machines (~25 GB of memory). The software is designed for samples to be run in parallel, so the total time for analysis even for large numbers of samples is often not much greater than the time for a single sample.

In summary, *ContamLD* is able to estimate autosomal nuclear contamination in ancient DNA accurately with standard errors that depend on the coverage of the sample. This will be particularly useful for female samples where X chromosome estimates are not possible. As a general recommendation for users, we believe in most cases all samples with a contamination estimate that is greater than 0.05 (5%) should be removed from further analyses, or the contamination should be explicitly modeled in population genetic analyses.

# Supplementary Data:

The supplement of this study includes 4 Excel spreadsheets detailing all ancient samples used and the contamination estimates for this algorithm. Also included are 10 supplementary figures.

# Materials and Methods

**Datasets:**

*Present-day samples:*
Genome wide datasets from individuals that were part of the 1000 Genomes Project [25] were used as present-day reference samples. We restricted to autosomal sites included in the aDNA ~1.24 million SNP capture reagent [2, 17] and to SNPs at greater than 10% minor allele frequency in the pooled 1000 Genomes Project dataset [25]. However, the software allows users to make panels based on their own SNP set. In the analyses presented here, we filtered for SNPs that were present in the 1000 Genomes dataset and also removed all sex chromosome SNPs leading to 1,085,678 SNPs in the final 1240K dataset and 5,633,773 SNPs in the final shotgun dataset.

*Ancient data set:*
We analyzed mitochondrial and X chromosome contamination estimates [16, 26] from ancient individuals from previous studies generated by shotgun sequencing or targeted enrichment with 1.24 million SNP enrichment, including many samples that failed quality control due to contamination but were from the same archaeological sites [2, 22, 27-33]. Information about the ancient individual data are detailed in Supplementary Online Table 1 and below.

*Obtaining sequence information:*
For each ancient individual, we generated the sequence-depth data from the sample bam file, counting the number of reference and alternative alleles at each SNP site in the analysis dataset. Damage-restricted data was generated by restricting to sequences with PMD scores greater than or equal to 3 [4]. Our software can accommodate both genotype call data as well as sequence data (the sequence data adds additional power to the analyses), but all analyses were performed using the sequence-based method. We provide users with tools to pull down read count data from BAM files in the format required for *ContamLD*.

**Haplotype Calculation**

To create haplotype panels, we obtained all SNP pairs in high LD for each 1000 Genomes population using PLINK version 1.9 [34] with $r^2$ cut-off of 0.2. (Users can increase power slightly at the expense of increased computational time by creating their own haplotype panel with a lower $r^2$ cutoff.) We then calculated the frequencies of each SNP in all of these pairs as well as the haplotype frequencies at each of these pairs while holding out the present-day individuals used for contamination simulation.

**Algorithm to Estimate Contamination**

*Overview:*

Our goal is to estimate α, the level of contamination, by examining the frequencies of allele pairs that should be in LD (we term this two-allele pair a haplotype) and determining how much their frequencies differ from what would be expected under no contamination. To estimate this, we need both the distribution underlying the haplotypes (*q*) that an uncontaminated test sample should have as well as the distribution of "unrelated haplotypes" ($\tilde{h}$) that would form by chance from background allele frequencies. Here and below, "distribution" refers to the set of frequencies of the different possible haplotypes (all possible combinations of ancestral and derived alleles at the SNP pair) across all haplotypes in the genome. Supplementary Figure 10 is a schematic of the algorithm.

*Determining Haplotype Distributions Based on Reference Panels:*

To determine *q* we must account for the fact that the test individual's genotypes do not have diploid calls and are not phased. Due to the low sequence depths at each SNP in many ancient DNA datasets, it is difficult to make confident heterozygous calls, so instead we create pseudo-haploid calls by randomly choosing a sequence to represent the genotype at that position (this holds when we are using genotype calls or the sequence information directly, and when multiple sequences cover the same SNP, we use all of them and treat them as independent). Thus, for this analysis, when examining a pair of SNPs, it is equally likely for the SNP pair to have been formed from the true haplotype (if the same parental chromosome is sampled from in both SNPs of the haplotype) or the background distribution (if the opposite parental chromosome is sampled from). We therefore can estimate *q* as:

$$q = h/2 + \tilde{h}/2$$

where *h* is the distribution of true haplotypes and $\tilde{h}$ is the distribution of unrelated haplotypes that would form by chance from background allele frequencies. For inbred samples, the weight on *h* is more than 1/2, because the two parental chromosomes are more related, but this can generally be corrected (see below).

$\tilde{h}$ can be determined by multiplying the SNP frequencies to obtain the haplotype frequencies that would form after randomly pairing SNPs of unrelated individuals. *h* can be estimated from an external reference panel using a maximum likelihood estimator (MLE) to obtain haplotypes frequencies in the population from the counts (necessary because the panels are not phased). The MLE set-up is:

$$log(L(h|c)) = \sum_{j=1}^{n} \sum_{i=1}^{4} c_{ij} log(P(i,j|h))$$

with:

$$P(i,j|h) = \sum_{a_1,a_2,b_1,b_2=0,1;a,b\rightarrow(i,j)} h_{(a_1,b_1)} * h_{(a_2,b_2)}$$

where $P(i, j|h)$ is the (unknown) diploid count distribution of the haplotypes of the population the test individual is from (approximated by the external panel), $n$ is the number of SNP pairs, $c$ is the vector of observed haplotypes in the diploid count panel (from 1000 Genomes), i sums over all 4 haplotype possibilities, $h_{(a,b)}$ are the (also unknown) haplotype distributions of the parents of the test individual (the haploid chromosomes they pass on to their child), and $a, b \rightarrow (i, j)$ implies that $a_1 + a_2 = i$ and $b_1 + b_2 = j$, meaning that one adds up all cases where the haplotype combination would lead to a particular diploid count (e.g. in the notation, for example, 01,11 means the first parent contributes a haplotype that has 0 alternative alleles at the first SNP and 1 alternative allele at the second SNP, and the second parent contributes a haplotype where both SNPs have the alternative allele. The test individual with these parents would then have a 12 diploid count, which means at the first SNP the individual has 1 alternative allele and at the second SNP the individual has 2 alternative alleles. Since our observed data are not phased, both 01,11 and 11,01 would lead to a 12 diploid count). This assumes independence of SNP pairs, which is not true, but because our standard errors are based on jackknife resampling across chromosomes, correlation among SNP pairs is corrected for in our error estimates.

The MLE would be computationally intractable to solve due to our lack of knowledge of which parent contributed to each count, so we instead used an EM algorithm to obtain $h$, where knowledge of the parents' contribution is the unobserved latent variable. The algorithm involves an expectation step of:

$$n_1 = \frac{C_{(i,j)} * \sum_{a,b \rightarrow (i,j)} h_{(a,b)} * h_{(a_2,b_2)}}{P(i, j|h)}$$

where $n_1$ is the expected number of times that the $(a, b)$ configuration of the father's chromosome contributed to a particular diploid count (this is the same value for the mother, $n_2$, because they are assumed to be from the same haplotype distribution). In other words, given the observed haplotype counts in the reference panel, how many times would it be expected that a particular haplotype configuration (e.g. ancestral at SNP1, derived at SNP2) in one of the parents contributed to those counts?

Once the counts ($n_1$ and $n_2$) of the haploid parents are obtained, they are added together to produce the diploid individual (i.e. the expected number of all possible haplotype configurations). Then the expected value of the haplotype distribution can be maximized by averaging over the possible haplotype distributions. Thus, the maximization step is:

$$D_{(a,b)} = \sum_{(i,j)} C_{(i,j)} * [n_1 + n_2]$$

$$\hat{h_{(a,b)}} = \frac{D_{(a,b)}}{\sum_{a,b} D_{(a,b)}}$$

where $D(a,b)$ is the sum of the probabilities of a particular haplotype configuration over all diploid count configurations.

We initially set all $h(a,b)$ to be 0.25 and then iterated through the algorithm until convergence (using a squared distance summed over all SNPs and a threshold of 0.001). We then used this estimate of $h$ to get an estimate of $q$ (based on the first equation above).

*Estimating Contamination Based on Haplotype Distributions and Test Individual's Haplotypes:*
To estimate α, we used the equation:

$$T = (1-2\alpha'+2\alpha'^2)q+2\alpha'(1-\alpha')\tilde{h}$$

Here $T$ is the expected distribution underlying the observed haplotypes of the sample, which is a mix of the test individual and contaminant. This means that assuming the test individual comes from a population with a haplotype distribution (frequency of the different haplotype possibilities at each SNP pair throughout the genome) that can be approximated by the chosen reference panel (and estimated as above), $T$ is the haplotype distribution expected for the sample given a particular amount of contamination ($\alpha'$, where ' is used to indicate that this is an estimate of the real $\alpha$). $q$ is the haplotype distribution for an uncontaminated sample. A fraction $(1-\alpha')^2 + \alpha'^2$ of the distribution should look like this, where $(1-\alpha')^2$ is the probability that two uncontaminated sequences form the SNP pair and $\alpha'^2$ is the probability that two contaminated sequences form the SNP pair, assuming the contaminating sequences are from a single individual, which would "re-form" a SNP pair with LD (note: this also makes the simplifying assumption that the contaminant and the test individual have the same background haplotype and SNP distribution). $\tilde{h}$ is the distribution of unrelated "haplotypes" that would form by chance from background allele frequencies in the population. Contamination would form these unrelated haplotypes by breaking up LD, so a fraction $2\alpha'(1-\alpha')$ of the distribution should look like this (the probability that the SNP pair is formed from a contaminated sequence and an uncontaminated sequence).

This expression can be used to solve for α' by maximizing the LOD (log of the odds) scores under the null hypothesis that $\alpha' = 0$ and the alternative hypotheses of different $\alpha'$. A LOD score is assigned to each estimate of the contamination rate ($\alpha$) between -0.1 to 0.5 (negative scores are included to allow correction for inbreeding). The grid of $\alpha'$ is scaled by intervals of 0.0001. The $\alpha'$ with the highest LOD score is the best estimate of $\alpha$, and is returned. When we have multiple sequences on the same SNP we assume independence of the sequences, which provides additional power. The assumption of independence does not bias the error estimation for the same reason as explained above for independence of SNP pairs.

*Correcting for Bias in Contamination Estimates:*
In practice, the $\alpha'$ we obtain is not equal to the true $\alpha$, because the reference panel does not perfectly capture the SNP and haplotype frequencies of the test sample. We found that this

difference causes a linear shift in contamination estimate where the mismatch between the sample individual and the reference panel leads to a positive shift while inbreeding leads to a negative shift. These biases can be addressed in either of two ways.

First, for the "damage correction" approach, we performed an $\alpha$' estimate only on alleles from sequences with evidence of damage characteristic of ancient samples. Under the assumption that these sequences are not affected by present-day contamination, the inferred $\alpha$' would be an estimate of the bias, which can be subtracted out from the estimate based on all sites. We separately analyzed the following pairs of SNPs: UU (both SNPs at undamaged sequences), DU (one site damaged and the other undamaged), and DD (both SNPs at damaged sequences). For the UU pairs, the value we calculate would be $\alpha$ + k, where k is the linear shift. For DU pairs the value calculated would be $\alpha/2$ + k, and for DD pairs the value calculated would be k. We added the likelihoods for these pairs and maximized the likelihood to solve for $\alpha$ and k. After solving for $\alpha$, we multiply by (1-damage rate) to obtain the contamination level across all sequences, because $\alpha$ is the contamination rate at undamaged sequences.

Second, for the "external correction" approach, we took individuals from the test individual's population that were high coverage and samples we believed had very low contamination (based on X chromosome estimates with *ANGSD* using method 1 as developed first by Rasmussen *et al*. [9]) and measured $\alpha$'. We assumed a true contamination of 0 for these samples and thus subtracted this $\alpha$' from all other contamination estimates. We caution that this method does not correct for uncertainty in the contamination estimate in the external sample used for benchmarking.

*Comparison to a Similar Method:*
The approach of *ContamLD* is similar to that of Vohr *et al*. [35] except the two have the opposite goals. Vohr *et al*. searches for LD in reads from two different samples in an attempt to determine whether the two samples are from the same individual (or closely related individuals), using a reference panel to determine LD patterns. In contrast, *ContamLD* searches for breaks in LD in the sequences of a single sample to determine if sequences from other individuals are present in the sample.

**Data simulation:**

To test the accuracy of the algorithm, we applied it to a variety of scenarios with both present-day DNA as well as real aDNA samples that had simulated present-day DNA contamination. In all our simulations with 1000 Genomes individuals, we removed the individual being used from our haplotype panel before performing the analyses.

*Simulating Contamination of Present-day Individuals:*
We first simulated contamination of present-day individuals with other present-day individuals as contaminants (this allowed us to be sure that there was no baseline contamination). In order to best approximate the distribution of both the damaged and undamaged sequences that is

characteristic of aDNA data, we used sequence-depth information from an ancient individual as a reference. At each SNP, the total number of simulated "damaged" and "undamaged" sequences was determined based on the number of damaged and undamaged sequences at the SNP in the reference ancient individual. The identity of each allele for the present-day "base" sample was randomly chosen based on the genotype of the "base" present-day 1000 Genomes individual at each SNP, as described above for the contamination. The addition of contaminant sequences to the dataset was performed using the method described above. In order to reduce bias caused by the damage correction procedure, the damage restricted dataset was generated only once for each simulation type (which included multiple simulations across varying contamination rates) and combined with the undamaged dataset to produce the overall dataset. This method was used to generate a simulated individual using present-day CEU (NA06985) or ASW (NA19625) from the 1000 Genomes dataset as the "base" sample from the sequence distributions of a 1.02x coverage ancient Iberian individual (I3756) (the "reference") [18]. The CEU (NA06984) individual was used as "contaminant" in each case.

We generated simulated data with contamination from multiple sources by adjusting the present-day contamination simulation method to randomly sample from two or more present-day source contaminant genomes with equal probability. In each case, a 1000 Genomes Project CEU individual (NA06985) was used as a "base" genome with the sequence distribution of I3756 (the "reference"). In the case of 2 sources of contamination (Supplementary Figure 5), two CEU individuals from the 1000 Genomes Project dataset (NA06984 and NA06986) were used as contamination sources, and in the case of three contamination sources, an additional CEU individual was used (NA06989). Data was generated for all combinations of undamaged contamination rates, α, from 0-15%.

*Simulated contamination of ancient individuals:*
We performed two sets of simulations contaminating different ancient individuals. In both cases we selected ancient male individuals with minimal contamination (as assessed by X chromosome contamination levels from *ANGSD* [16]) to act as the "base" uncontaminated genome. In the first simulation set, we tested *ContamLD*'s performance with different ancient individuals and different present-day contaminant individuals from the 1000 Genomes dataset [25] to assess the impact of contaminant ancestry and coverage of the ancient individual. In this case we were only using *ContamLD* and thus we performed the simulated contamination on the genotype level. In the second simulation set, we compared *ContamLD* to *ANGSD* and used a ~1200BP ancient West Eurasian individual (I10895) to contaminate the BAM files directly.

In the first simulation set, we assumed that sequences with C-to-T damage are highly unlikely to be the product of contamination (this assumption would be falsified in the context of cross-contamination by another ancient DNA sample). Thus, we exclusively added contamination to the "undamaged" fraction of sequences. At each SNP site, we classified sequences present in the damage-restricted dataset as "damaged" and added to the simulated data. We classified all other sequences as "undamaged" and also added them to the simulated data, but for each "undamaged sequence" we added a contaminant sequence to the simulated SNP data with probability $\alpha/(1-\alpha)$, where $\alpha$ is equal to the contamination rate (since the added sequences

contribute to the total number of sequences, we needed to add a higher proportion than the contamination rate to obtain our desired contamination rate). The identity of the added contaminant allele was randomly chosen based on the genotype of the chosen "contaminant" present-day genome at the site (i.e. if the contaminant individual was homozygous at the site, the allele it possesses would be added to the simulated individual, while if it were heterozygous at the site, either the reference or alternative allele would be selected randomly and added to the simulated individual). This method maintains the underlying distribution of "uncontaminated" reference and alternative alleles at each SNP site, while adding additional "contaminant" alleles to each site, producing an overall contamination rate of $\alpha$ in the undamaged sequences.
For each simulation, we generated two output files: (1) a file reporting the total number of sequences carrying reference and alternative alleles at each SNP and (2) a damage restricted file reporting the total number of damaged sequences carrying reference and alternative alleles at each SNP. We used a 1.02x coverage ancient Iberian individual (I3756) (Supplementary Online Table 1) with contamination from either the 1000 Genomes CEU individual NA06984, the TSI individual NA20502, the CHB individual NA18525, or the YRI individual NA18486. We also used 5 other ancient individuals: I1845 (an ancient Iberian sample of 0.46x coverage) [18], I2743 (an ancient Hungarian of 0.27x coverage) [30], I5891 (a Neolithic Ukrainian individual of 0.016x coverage) [36], DA362.SG (a Russian early Neolithic Shamanka East Asian individual of 1.10x coverage) [21], and I9028.SG (a South African individual of 1.21x coverage) [22]. In each case, we simulated individuals with 0-15% contamination.

For the second simulation set, we analyzed 65 ancient individuals of average coverage over 0.5X and baseline *ANGSD* estimates under 2% (Supplementary Online Table 2). In these cases, we added artificial contamination with sequences from a ~1200BP ancient West Eurasian individual (I10895) into the BAM files at the amounts: (0.000, 0.005, 0.010, 0.020, 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, 0.150). We removed two base pairs from the end of each sequence of partial UDG treated samples and ten nucleotides for non-UDG treated samples and pulled down the genotypes by randomly selecting a single sequence at each site covered by at least one sequence in each individual to represent the individual's genotype at that position ("pseudo-haploid" genotyping). To ensure that the damage sequences were only from the non-contaminant individual (so that we could use the damage restricted correction mode, option 1, of *ContamLD* without bias), we created the "damaged" sequence set as a randomly chosen 5% of the sequences from the non-contaminant individual. We then analyzed the data with *ContamLD* (damage restricted correction version, option 1) and *ANGSD* using default settings (Method 1). We also performed simulations with a 1.0x coverage ancient West Eurasian ancestry individual (DA57.SG, an ancient Krgyzstanian individual) [37] down-sampled to 0.5x coverage and contaminated with I10895. To simulate different damage rates, we varied the damage rate to the proportions (0.005, 0.01, 0.02, 0.03, 0.04, 0.05, 0.075) by setting the amount of "damaged" sequences to be those proportions.

As a last simulation, we examined the case of an ancient individual contaminating another ancient individual where some of the damaged sequences would also come from the contaminating individual. In this simulation, we analyzed a 1.02x coverage ancient Iberian

individual (I3756) and contaminated the BAM with sequences from a ~1200BP ancient West Eurasian individual (I10895) in the proportions (0.000, 0.005, 0.010, 0.020, 0.025, 0.030, 0.040, 0.050, 0.060, 0.070, 0.080, 0.090, 0.100, 0.150, 0.200, 0.300). We then down-sampled the BAM, taking a random 5% of the sequences of these contaminated BAM files to act as the "damaged" sequences, because this would correct for any baseline contamination in the I3756 individual yet would simulate additional contamination of I3756 by an ancient individual with the same damage rate as I3756 (i.e. if there is 5% contamination, then also 5% of the damaged sequences would be from the contaminant individual in this simulation). We then performed the standard processing of both the full contaminated BAMs and the 5% down-sampled BAMs (simulated to be "damaged" sequences), removing two base pairs from the end of each sequence and carrying out a "pseudo-haploid" genotype pulldown. We ran *ContamLD* on the resulting data with damage restricted correction, option 1.

*Direct Analyses of Contamination Levels in Ancient Individuals:*
As our last set of analyses, we directly measured contamination levels in ancient individuals without simulated contamination. We used *ContamLD* to examine shotgun sequenced individuals analyzed at the1240K SNP set and the large 5.6 million SNP shotgun panel. The ancient shotgun sequenced individuals were of 0.1-0.5x coverage from Allentoft *et al*., 2015 [31], Damgaard *et al.*, Nature 2018 [37], and Damgaard *et al*., Science 2018 [21]. In addition, we analyzed 439 individuals from a variety of ancestries with *ContamLD* (damage corrected version), *ANGSD* [16, 38] using default settings (we report the results from Method 1), and *contamMix* [39] with the settings: down-sampling to 50X for samples above that coverage, --trimBases X (2 bases for UDG-half samples and 10 bases for UDG-minus samples), 8 threads, 4 chains, and 2 copies, taking the first one that finishes. Supplementary Online Table 1 includes all information from these individuals.

# Declarations

**Ethics Approval and Consent to Participate**
Not applicable (all samples were from previously published studies).

**Consent for publication**
Not applicable.

**Availability of Data and Materials:**
All data analyzed in this article are available in [2, 21, 22, 27-33, 37]. The software is available at: https://github.com/nathan-nakatsuka/ContamLD. It requires Python 3 and R (any version should suffice). Archived version (1.0) used for analyses in this manuscript: https://zenodo.org/record/3736774#.XoTbj257mgQ (DOI: 10.5281/zenodo.3736774)
Scripts for data simulations are available in the Github folder "Simulation_Scripts".

**Competing interests:**
The authors declare that they have no competing interests.

**Funding:**
Funding was provided by an NIGMS (GM007753) fellowship to NN and a MHAAM fellowship to EH. DR is an Investigator of the Howard Hughes Medical Institute this work was supported by grants HG006399 and GM100233 from the National Institutes of Health, by an Allen Discovery Center grant from the Paul Allen Foundation, and by grant 61220 from the John Templeton Foundation.

**Author Contributions:**
N.N., E.H., N.P., and D.R. conceived the study. N.N., E.H., and S.M. performed analysis. N.N., E.H., and D.R., wrote the manuscript with the help of all co-authors.

**Acknowledgements:**
We thank Iosif Lazaridis and Mark Lipson for helpful discussions.

# References


1. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.
2. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al: **Massive migration from the steppe was a source for Indo-European languages in Europe.** *Nature* 2015, **522:**207-211.
3. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D: **Partial uracil-DNA-glycosylase treatment for screening of ancient DNA.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20130624.
4. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Paabo S, Krause J, Jakobsson M: **Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal.** *Proc Natl Acad Sci U S A* 2014, **111:**2229-2234.
5. Sawyer S, Renaud G, Viola B, Hublin JJ, Gansauge MT, Shunkov MV, Derevianko AP, Prufer K, Kelso J, Paabo S: **Nuclear and mitochondrial DNA sequences from two Denisovan individuals.** *Proc Natl Acad Sci U S A* 2015, **112:**15696-15700.
6. Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S: **DNA analysis of an early modern human from Tianyuan Cave, China.** *Proc Natl Acad Sci U S A* 2013, **110:**2223-2227.
7. Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, Good JM, Maricic T, Stenzel U, et al: **A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing.** *Cell* 2008, **134:**416-426.
8. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH, et al: **A draft sequence of the Neandertal genome.** *Science* 2010, **328:**710-722.
9. Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T: **An Aboriginal Australian genome reveals separate human dispersals into Asia.** *Science* 2011, **334:**94-98.
10. Moreno-Mayar JV, Korneliussen TS, Dalal J, Renaud G, Albrechtsen A, Nielsen R, Malaspinas A-S: **A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data.** *Bioinformatics* 2020, **36:**828-841.
11. Cibulskis K, McKenna A, Fennell T, Banks E, DePristo M, Getz G: **ContEst: estimating cross-contamination of human samples in next-generation sequencing data.** *Bioinformatics* 2011, **27:**2601-2602.
12. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM: **Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.** *Am J Hum Genet* 2012, **91:**839-848.
13. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, De Filippo C: **A high-coverage genome sequence from an archaic Denisovan individual.** *Science* 2012, **338:**222-226.

14. Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C: **The complete genome sequence of a Neanderthal from the Altai Mountains.** *Nature* 2014, **505:**43-49.
15. Racimo F, Renaud G, Slatkin M: **Joint estimation of contamination, error and demography for nuclear DNA from ancient humans.** *PLoS genetics* 2016, **12**.
16. Korneliussen TS, Albrechtsen A, Nielsen R: **ANGSD: Analysis of Next Generation Sequencing Data.** *BMC Bioinformatics* 2014, **15:**356.
17. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al: **Genome-wide patterns of selection in 230 ancient Eurasians.** *Nature* 2015, **528:**499-503.
18. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M: **The genomic history of the Iberian Peninsula over the past 8000 years.** *Science* 2019, **363:**1230-1234.
19. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
20. Seguin-Orlando A, Korneliussen TS, Sikora M, Malaspinas A-S, Manica A, Moltke I, Albrechtsen A, Ko A, Margaryan A, Moiseyev V: **Genomic structure in Europeans dating back at least 36,200 years.** *Science* 2014, **346:**1113-1118.
21. de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, Barjamovic G, Rasmussen S, Zacho C, Baimukhanov N: **The first horse herders and the impact of early Bronze Age steppe expansions into Asia.** *Science* 2018, **360:**eaar7711.
22. Skoglund P, Thompson JC, Prendergast ME, Mittnik A, Sirak K, Hajdinjak M, Salie T, Rohland N, Mallick S, Peltzer A, et al: **Reconstructing Prehistoric African Population Structure.** *Cell* 2017, **171:**59-71 e21.
23. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D: **Methods for high-density admixture mapping of disease genes.** *The American Journal of Human Genetics* 2004, **74:**979-1000.
24. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
25. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.
26. Renaud G, Slon V, Duggan AT, Kelso J: **Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA.** *Genome Biol* 2015, **16:**224.
27. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al: **Ancient human genomes suggest three ancestral populations for present-day Europeans.** *Nature* 2014, **513:**409-413.
28. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al: **Genomic insights into the origin of farming in the ancient Near East.** *Nature* 2016, **536:**419-424.
29. Lazaridis I, Mittnik A, Patterson N, Mallick S, Rohland N, Pfrengle S, Furtwangler A, Peltzer A, Posth C, Vasilakis A, et al: **Genetic origins of the Minoans and Mycenaeans.** *Nature* 2017, **548:**214-218.

30. Lipson M, Szecsenyi-Nagy A, Mallick S, Posa A, Stegmar B, Keerl V, Rohland N, Stewardson K, Ferry M, Michel M, et al: **Parallel palaeogenomic transects reveal complex genetic history of early European farmers.** *Nature* 2017, **551:**368-372.
31. Allentoft ME, Sikora M, Sjogren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlstrom T, Vinner L, et al: **Population genomics of Bronze Age Eurasia.** *Nature* 2015, **522:**167-172.
32. Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, Leidinger P, Backes C, Khairat R, Forster M, et al: **New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing.** *Nat Commun* 2012, **3:**698.
33. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A: **The Beaker phenomenon and the genomic transformation of northwest Europe.** *Nature* 2018, **555:**190.
34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81:**559-575.
35. Vohr SH, Najar CFBA, Shapiro B, Green RE: **A method for positive forensic identification of samples from extremely low-coverage sequence data.** *BMC genomics* 2015, **16:**1034.
36. Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O: **The genomic history of southeastern Europe.** *Nature* 2018, **555:**197.
37. de Barros Damgaard P, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, Moreno-Mayar JV, Pedersen MW, Goldberg A, Usmanova E: **137 ancient human genomes from across the Eurasian steppes.** *Nature* 2018, **557:**369.
38. Durvasula A, Hoffman PJ, Kent TV, Liu C, Kono TJ, Morrell PL, Ross-Ibarra J: **angsd-wrapper: utilities for analysing next-generation sequencing data.** *Mol Ecol Resour* 2016, **16:**1449-1454.
39. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PL, Aximu-Petri A, Prufer K, de Filippo C, et al: **Genome sequence of a 45,000-year-old modern human from western Siberia.** *Nature* 2014, **514:**445-449.

# Chapter 5: Using ancient and modern DNA to infer the history of Central and South America

This section is based on the following papers I contributed to during my PhD:

Bongers, J.L.; **Nakatsuka, N.**; O'Shea, C.; Harper, T.; Tantaleán Inga, H.; Stanish, C.; Fehren-Schmitz, L. "Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca state-sponsored resettlement along the Peruvian Coast." *PNAS*. 2020 July 13; 202005965.

**Nakatsuka, N.***; Luisi, P.*; Motti, J.; Salemme, M.; Santiago, F.; del Campo, M.; Vecchi, R.; Espinosa-Parrilla, Y.; Prieto, A.; Adamski, N.; Lawson, A.M.; Harper, T.; Culleton, B.; Kennett, D.; Lalueza-Fox, C.; Mallick, S.; Rohland, N.; Guichon, R.; Cabana, G.; Nores, R.*; Reich, D.* "Ancient genomes in South Patagonia reveal population movements associated with technological shifts and geography." *Nature Communications*. 2020 Aug. 3; 11(1): 3867.

**Nakatsuka, N.**; Lazaridis, I., Barbieri, C.; Skoglund, P.; Rohland, N.; Mallick, S.; Harkins-Kinkaid, K.; Ferry, M.; Harney, E.; Michel, M.; Stewardson, K.; Novak-Forst, J.; Posth, C.; Durruty, M.; Álvarez, K.; Beresford-Jones, D.; Burger, R.; Cadwallader, L.; Capriles, J.; Fujita, R.; Isla, J.; Lau, G.; Aguirre, C.; LeBlanc, S.; Maldonado, S.; Meddens, F.; Messineo, P.; Culleton, B.; Harper, T.; Quilter, J.; Politis, G.; Reindel, M.; Rivera, M.; Salazar, L.; Sandoval, J.; Santoro, C.; Scheifler, N.; Standen, V.; Barreto, M.; Espinoza, I.; Tomasto-Cagigao, E.; Valverde, G.; Kennett, D.; Cooper, A.; Krause, J.; Haak, W.; Llamas, B.; Reich, D.*; Fehren-Schmitz, L.*. "A Paleogenomic Reconstruction of the Deep Population History of the Andes. *Cell*. 28 May 2020. 181(5): 1131-1145.e21.

Posth, C.*; **Nakatsuka, N.***; Lazaridis, I.; Skoglund, P.; Mallick, S.; Lamnidis, T.; Rohland, N.; Nagele, K.; Adamski, N.; Bertolini, E.; Broomandkhoshbacht, N.; Cooper, A.; Culleton, B.; Ferraz, T.; Ferry, M.; Furtwangler, A.; Haak, W.; Harkins, K.; Harper, T.; Hunemeier, T.; Lawson, A.; Llamas, B.; Michel, M.; Nelson, E.; Oppenhemier, J.; Patterson, N.; Schiffels, S.; Sedig, J.; Stewardson, K.; Talamo, S.; Wang, C.; Hublin, J.; Hubbe, M.; Havarti, K.; Nuevo Delaunay, A.; Beier, J.; Francken, M.; Kaulicke, P.; Reyes-Centeno, H.; Rademaker, K.; Trask, W.; Robinson, M.; Gutierrez, S.; Prufer, K.; Salazar-Garcia, D.; Chim, D.; Muller Plumm Gomes, L; Alves, M.; Liryo, A.; Inglez, M.; Oliveira, R.; Bernardo, D.; Barioni, A.; Wesolowski, V.; Scheifler, N.; Rivera, M.; Plens, C.; Messineo, P.; Figuti, L.; Corach, D.; Scabuzzo, C.; Eggers, S.; DeBlasis, P.; Reindel, M.; Mendez, C.; Politis, G.; Tomasto-Cagigao, E.; Kennett, D.*; Strauss, A.*; Fehren-Schmitz, L.*; Krause, J.*; Reich, D.*. "Reconstructing the Deep Population History of Central and South America." *Cell*. 15 Nov. 2018. 175(5):1185-1197.

**Overview:**

These four studies show the use of aDNA to infer human history as well as the integration of genetics with archaeology and linguistics (particularly, in the case of the Patagonia study) to obtain novel insights about the past. The Chincha valley Peruvian coast study also integrated strontium isotope and historical records. The first study (Posth*, Nakatsuka* *et al.*, 2018) in Chapter 5.1 focused on the earliest history of Central and South America (~11,000 BP to 4,000 BP). The next three studies were more regionally focused. Nakatsuka *et al.*, *Cell* 2020 in Chapter 5.2 focused on the changes in genetic structure in the Andes from 9,000 BP to the present, including an examination of the genetics of individuals from the major Andean cultures, such as the Moche, Wari, Tiwanaku, and Inca. Bongers *et al.*, 2020 is briefly summarized in Chapter 5.2; it focused in individuals from the Inca time period in Chincha Valley and showed that they had the same genetic profile as some Inca period

individuals from Torontoy, Cusco and an Inca sacrificial victim in Argentina, pointing to Inca state sponsored movement of people away from their original home locations. Nakatsuka *et al*., *Nature Communications* 2020 in Chapter 5.3 focused on Southern Patagonia and the migrations and admixtures that led to the genetic structure of the groups along the coast, in particular differentiating the maritime diet groups from the terrestrial diet groups.

**Ethics:**

The ethical approaches to these studies were specific for each region. The specific governmental and institutional permits for exportation and study of the skeletal material were followed in all cases, but additional measures were taken depending on the particular context. For example, in Belize the relevant permits were obtained, but in line with the principles of beneficence, non-maleficence, and autonomy, the research was done with prior review by the administration and scientific staff of the Ya'axché Conservation Trust (YCT), a local NGO that co-manages the Bladen Nature Reserve with the government of Belize. In addition, multiple public consultation presentations were done with YCT and other interested community members where the results of this study were presented and the public could give feedback. Similarly, in Argentina, the Communidad Mapuche-Tehuelche Cacique Pincen were consulted throughout the project with regards to the ancient individuals from near their land, and they participated in the rescue excavation. In addition, the results were communicated to the group through the leader of the community (Beatriz Araujo Pincen), who also co-authored the presentation about the site delivered in the Congreso de Arquelogia Pampeana.

In Patagonia, numerous meetings were held with different members of the local indigenous groups with regards to the work, and educational activities were done in local schools as well. The work followed the interest of some members of the native communities in establishing the relationship of the ancient humans to present-day Indigenous people, and the findings were distributed after translation to Spanish. The work also had elements of restorative justice as it helped to improve the conditions of the Museo del Fin del Mundo and preserve the skeletal material from being lost due to environmental disturbance.

In the Central Andes, particularly Peru, the cultural and historical context is different due to the strong history of indigenism, especially in the Peruvian government. Thus, in this case we did our primary consultation with the provincial and state-based offices of the responsible institutions, and in particular had the study approved by the Ministry of Culture of Peru, which was originally created to revalue indigenous culture. At the same time, however, there is a complex relationship between the mestizo individuals, who have high governmental representation, and the comunidades campesinas (peasant communities), many of whom are groups perceived as having greater ties to their indigenous culture (e.g. primarily speaking Quechua, Aymara, or other indigenous languages) but have less governmental representation [1-3]. Thus, in addition to the governmental permissions, substantial engagement with local communities was performed in Peru and Bolivia both before and throughout the study. In Chile, the local heritage institutions gave permission, but there were no local indigenous communities living near the site who were available to consult. The findings were translated to Spanish to increase accessibility in line with the principle of beneficence. Lastly, the study of the San Sebastian samples were from a US collection as part of a repatriation effort, and they are now curated in Cusco, in line with the principle of restorative justice.

# Chapter 5.1: Reconstructing the Deep Population History of Central and South America

Cosimo Posth[1,2,45,*], Nathan Nakatsuka[3,4,45,*], Iosif Lazaridis[3], Pontus Skoglund[3,5], Swapan Mallick[3,6,7], Thiseas C. Lamnidis[1], Nadin Rohland[3], Kathrin Nägele[1], Nicole Adamski[3,7], Emilie Bertolini[8], Nasreen Broomandkhoshbacht[3,7], Alan Cooper[9], Brendan J. Culleton[10,11], Tiago Ferraz[1,12], Matthew Ferry[3,7], Anja Furtwängler[2], Wolfgang Haak[1], Kelly Harkins[13], Thomas K. Harper[10], Tábita Hünemeier[12], Ann Marie Lawson[3,7], Bastien Llamas[9], Megan Michel[3,7], Elizabeth Nelson[1,2], Jonas Oppenheimer[3,7], Nick Patterson[6], Stephan Schiffels[1], Jakob Sedig[3], Kristin Stewardson[3,7], Sahra Talamo[14], Chuan-Chao Wang[1,15], Jean-Jacques Hublin[14], Mark Hubbe[16,17], Katerina Harvati[18,19], Amalia Nuevo Delaunay[20], Judith Beier[18], Michael Francken[18], Peter Kaulicke[21], Hugo Reyes-Centeno[18,19], Kurt Rademaker[22], Willa R. Trask[23], Mark Robinson[24], Said M. Gutierrez[25], Keith M. Prufer[26,27], Domingo C. Salazar-Garcia[14,28], Eliane Nunes Chim[29], Lisiane Müller Plumm Gomes[12], Marcony Lopes Alves[29], Andersen Liryo[30], Mariana Inglez[12], Rodrigo Elias Oliveira[12,31], Danilo V. Bernardo[32], Alberto Barioni[33], Veronica Wesolowski[29], Nahuel A. Scheifler[34], Mario A. Rivera[35,36,37], Claudia R. Plens[38], Pablo G. Messineo[34], Levy Figuti[29] , Daniel Corach[39], Clara Scabuzzo[40], Sabine Eggers[12,41], Paulo DeBlasis[29], Markus Reindel[42], César Méndez[20], Gustavo Politis[34], Elsa Tomasto-Cagigao[21], Douglas J. Kennett[10,11,46], André Strauss[12,18,29,43,46], Lars Fehren-Schmitz[13,44,46], Johannes Krause[1,2,46], David Reich[3,6,7,46,47,*]

[1] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena 07745, Germany
[2] Institute for Archaeological Sciences, Archaeo- and Palaeogenetics, University of Tübingen, Tübingen 72070, Germany
[3] Department of Genetics, Harvard Medical School, New Research Building, Boston, MA 02115, USA
[4] Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02115, USA
[5] Francis Crick Institute, London NW1 1AT, UK
[6] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[7] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02115, USA
[8] Dipartimento di Biologia e Biotecnologie, Università di Pavia, Pavia 27100, Italy
[9] Australian Centre for Ancient DNA, School of Biological Sciences and The Environment Institute, Adelaide University, Adelaide, SA 5005, Australia
[10] Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA
[11] Institutes for Energy and the Environment, The Pennsylvania State University, University Park, PA 16802, USA
[12] Departamento de Genética e Biologia Evolutiva, Universidade de São Paulo, São Paulo 05508-090, Brazil
[13] UCSC Paleogenomics, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[14] Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany
[15] Department of Anthropology and Ethnology, Xiamen University, Xiamen 361005, China
[16] Department of Anthropology, The Ohio State University, Columbus, OH 43210, USA
[17] Instituto de Arqueología y Antropología, Universidad Católica del Norte, San Pedro de Atacama, Región de Antofagasta, Antofagasta CP 1410000, Chile
[18] Institute for Archaeological Sciences, Palaeoanthropology and Senckenberg Centre for Human Evolution and Palaeoenvironment, University of Tuebingen, Tübingen 72070, Germany
[19] DFG Center for Advanced Studies, "Words, Bones, Genes, Tools," University of Tübingen, Tübingen 72070, Germany
[20] Centro de Investigación en Ecosistemas de la Patagonia, Coyhaique 5951601, Chile
[21] Pontifical Catholic University of Peru, San Miguel, Lima 32, Peru
[22] Department of Anthropology, Michigan State University, East Lansing, MI 48824, USA

[23] SNA International supporting the Defense POW/MIA Accounting Agency, Department of Defense, Joint Base Pearl Harbor-Hickam, HI 96853, USA
[24] Department of Archaeology, Exeter University, Exeter EX4 4QJ, UK
[25] Ya'axché Conservation Trust, Punta Gorda Town, Belize
[26] Department of Anthropology, University of New Mexico, Albuquerque, NM 87131, USA
[27] Center for Stable Isotopes, University of New Mexico, Albuquerque, NM 87131, USA
[28] Grupo de Investigación en Prehistoria, IKERBASQUE-Basque Foundation for Science, Bilbao 48013, Bizkaia, Spain
[29] Museu de Arqueologia e Etnologia, Universidade de São Paulo, São Paulo 05508-070, Brazil
[30] Museu Nacional da Universidade Federal do Rio de Janeiro, Rio de Janeiro 20940-040, Brazil
[31] Departamento de Estomatologia, Faculdade de Odontologia, Universidade de São Paulo, São Paulo 05508-000, Brazil
[32] Laboratório de Estudos em Antropologia Biológica, Bioarqueologia e Evolução Humana, Instituto de Ciências Humanas e da Informação, Universidade Federal do Rio Grande, Rio Grande do Sul 96203-900, Brazil,
[33] Faculdade de Filosofia Ciencias e Letras, Universidade de São Paulo, São Paulo 05508-080, Brazil
[34] INCUAPA-CONICET, Facultad de Ciencias Sociales, Universidad Nacional del Centro de la Provincia de Buenos Aires, Olavarría 7400, Argentina
[35] Comité Chileno del Consejo Internacional de Monumentos y Sitios, Santiago 8320000, Chile
[36] Field Museum of Natural History, Chicago, Illinois 60605, USA
[37] Universidad de Magallanes, Punta Arenas, 6200000, Chile
[38] Escola De Filosofia, Letras E Ciências Humanas, Universidade Federal de São Paulo, São Paulo 07252-312, Brazil
[39] Servicio de Huellas Digitales Genéticas, School of Pharmacy and Biochemistry, Universidad de Buenos Aires y CONICET, Ciudad Autonoma de Buenos Aires, Junin 954, Argentina
[40] CONICET- División Arqueología, Facultad de Ciencias Naturales y Museo, La Plata 1900, Argentina
[41] Naturhistorisches Museum Wien, Vienna 1010, Austria
[42] German Archaeological Institute, Commission for Archaeology of Non-European Cultures, Bonn D-53173, Germany
[43] Centro de Arqueologia Annette Laming Emperaire, Miguel A Salomão, Lagoa Santa, MG 33400-000, Brazil
[44] UCSC Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA
[45] These authors contributed equally
[46] Senior author
[47] Lead contact
* Corresponding author

**Keywords:**
South America, Central America, Population Genetics, Archaeology, Anthropology

**Correspondence**
posth@shh.mpg.de (C.P.), nathan_nakatsuka@hms.harvard.edu (N.N.) and reich@genetics.med.harvard.edu (D.R.)

**Supplementary Material:** All supplementary material can be found in the supplement of Posth*, Nakatsuka*, *et al.*, 2018 *Cell*.

# Abstract

We report genome-wide ancient DNA from 49 individuals forming four parallel time transects in Belize, Brazil, the Central Andes, and the Southern Cone, each dating to at least ~9,000 years. The common ancestral population radiated rapidly from just one of the two early branches that contributed to Native Americans today. We document two previously unappreciated streams of gene flow between North and South America. One affected the Central Andes by ~4,200 years ago, while the other explains an affinity between the oldest North American genome associated with the Clovis culture and the oldest Central and South Americans from Chile, Brazil and Belize. However, this was not the primary source for later South Americans, as the other ancient individuals derive from lineages without specific affinity to the Clovis associated genome, suggesting a population replacement that began at least 9000 years ago and was followed by substantial population continuity in multiple regions.


# Introduction

Genetic studies of present-day and ancient Native Americans have revealed that the great majority of ancestry in indigenous people in non-Arctic America derives from a homogeneous ancestral population. This population was inferred to have diversified 17500-14600 calendar years before present (BP) [4] into two branches that have been called "*Southern Native American*" or "*Ancestral A*" (*ANC-A*) and "*Northern Native American*" or "*Ancestral B*" (*ANC-B*) [4-8]. An individual dating to ~12900-12700 BP from the Anzick site in Montana and associated with the Clovis culture was on the *ANC-A* lineage, which is also heavily represented in present-day Central and South Americans and in ancient Californians. In contrast, *ANC-B* ancestry is heavily represented in eastern North Americans and in ancient people from southwest Ontario [8]. The original studies that documented these two deep lineages fit models in which Central and South Americans were of entirely *ANC-A* ancestry [6, 7]. However, Scheib *et al.* 2018 suggested that all Central and South Americans harbor substantial proportions of both ancestries (at least ~30% of each).

Recent analyses have also shown that some groups in Brazil share more alleles with Australasians (indigenous New Guineans, Australians and Andaman Islanders) [5, 9] and a ~40000 BP individual from northern China [10] than do other Central and South Americans. Such patterns suggest that these groups do not entirely descend from a single homogeneous population and instead derive from a mixture of populations, one of which, *Population Y,* bore a distinctive affinity to Australasians.

Prior to this study, published data from Central and South America older than the last millennium was limited to two low coverage genomes [5]. Here we report genome-wide data from 49 individuals from Belize, Brazil, Peru, and the Southern Cone (Chile and Argentina), 41 older than 1000 years, with each time transect starting between 10900-8600 BP (Figure 1 and Online Table 1). To obtain these data, we worked with government agencies and indigenous peoples to identify samples, prepared powder for skeletal material, extracted DNA [11], and generated single and double stranded DNA libraries most of which we treated with the enzyme uracil-DNA glycosylase (UDG) to reduce characteristic errors of ancient DNA [12, 13]. We enriched for mitochondrial DNA (mtDNA) and ~1.2 million single nucleotide polymorphisms

(SNPs) [14], and sequenced the enriched libraries on Illumina instruments (Methods and Online Table 1). We combined ancient and present-day data to study genetic changes over the last 11000 years.

### *Ethics Statement*

Genetic studies of human history shed light on how ancient and present-day people are biologically related, and it is therefore important to be attentive not just to scientific issues but also to perspectives of indigenous communities when carrying out this work [15]. We took a case-by-case approach in each region we studied. In Peru and in some other countries in Central and South America, there is a strong tradition of indigenism in state policy, and governmental officials are recognized as representatives of indigenous perspectives [[16, 17]; Ley General del Patrimonio Cultural de la Nación (Law No. 28296).] We therefore consulted with provincial and state-based offices of the Ministry of Culture to obtain permission for analysis, and also incorporated feedback from local community archaeologists to represent indigenous perspectives; permission for sampling was obtained under Resolución Directoral Nacional No. 1346, 545-2011, and RDN No. 092-2016. In Brazil, we obtained research permits from IPHAN (the National Institute of Historical and Artistic Heritage). In Chile and Argentina, in addition to obtaining permits from the local heritage institutions, we sought to determine if any local indigenous group considered the skeletons we analyzed to be ancestors. For most samples, no indigenous community lived near the sites or indicated a connection to the analyzed skeletons, with the exception of a community living near the site of Laguna Chica in Argentina, which approved the study after consultation and participated in the rescue excavation. In Belize, we obtained permission from the National Institute of Culture and History and the Institute of Archaeology, the legal entities responsible for issuing research permits, and we carried out public consultation with local collaborators and communities (see the archaeological site information section in the Supplemental Information for additional details).

## Results and Discussion

### *Authenticity of Ancient DNA*

We evaluated the authenticity of the isolated DNA based on its harboring: 1) characteristic cytosine-to-thymine mismatches to the reference genome at the ends of the sequenced fragments, 2) point estimates of contamination in mtDNA below 5% [18], 3) point estimates of X chromosome contamination in males below 3% [19], and 4) point estimates of genome-wide contamination below 5% (Nakatsuka, Harney *et al.*, in preparation). We removed from analysis two individuals that we genetically determined to be first degree relatives of other individuals with higher DNA yields within the dataset but fully report the data for both here (Methods and Online Table 1).

### *Long-Standing Population Continuity in Multiple Regions of South America*

We grouped ancient individuals by location, date range, and genetic similarity, for the most part using italicized labels like *Argentina_ArroyoSeco2_7700BP* (“country” followed by

"site" followed by a "date" which for us is the average of the midpoint of the date ranges for the individuals in the grouping rounded to the nearest hundred) [20]. These groupings sometimes span an extensive period of time; for example, the eight Arroyo Seco 2 date estimates range from 8570 to 7160 BP. For some analyses we also lumped individuals into larger clusters, for example grouping individuals from the Andes before and after ~4200 BP into "*Early Andes*" and "*Late Central Andes*" based on qualitatively different affinities to other individuals in the dataset (Methods).

To obtain an understanding of how the ancient individuals relate to present-day ones, we computed $f_3$- and $f_4$-statistics, which estimate allele sharing between samples in a way that is unbiased by population-specific drift [21].

The oldest individuals in the data set show little specific allele sharing with present-day people. For example, a ~10900 BP individual from Chile (from the site of Los Rieles) shows only slight excess affinity to later Southern Cone individuals. In Belize, individuals from two sites dating to ~9300 and ~7400 BP (Mayahak Cab Pek and Saki Tzul) do not share significantly more alleles with present-day people from the region near Belize than they do with present-day groups elsewhere in Central and South America. In Brazil, genetic data from sites dating to ~9600 BP (Lapa do Santo) and ~6700 BP (Laranjal) show no distinctive shared ancestry with present-day Brazilians (Figure 2, Figure S1, Table S1), although the Laranjal individuals do show evidence of shared ancestry with a ~5800 BP individual from Moraes (Online Table 2), confirmed by the statistic $f_4$(*Mbuti*, *Brazil_Laranjal_6700BP*; *Brazil_LapaDoSanto_9600BP*, *Brazil_Moraes_5800BP*), which is Z=7.7 standard errors from zero.

We detect long-standing continuity between ancient and present-day Native Americans in each of the regions of South America we analyzed beginning at least ~5800 BP, a pattern that is evident in heatmaps, trees, and multi-dimensional scaling plots computed on outgroup-$f_3$ statistics (Figure 2, Table S1, Figure S1, Figure S2). In Peru, the most ancient individuals dating up to ~9000 BP from Cuncaicha and Lauricocha share alleles at the highest rate with present-day indigenous groups living in the Central Andes [22, 23]. Individuals dating up to ~8600 BP from Arroyo Seco 2 and Laguna Chica also show the strongest allele sharing with some present-day indigenous people in the Southern Cone. In Brazil the evidence of continuity with present-day indigenous people begins with the Moraes individual at ~5800 BP. A striking pattern of continuity with present-day people is also observed in the ~2000 BP Jabuticabeira 2 individuals who were part of the Sambaqui shell-mound building tradition that was spread along the south Brazilian coast from around 8000-1000 BP. The Jabuticabeira 2 individuals share significantly more alleles with some Ge-speaking groups than they do with some Tupi-Guarani speaking groups who have been predominant on the coast during the post-Colonial period (Figure S3, Table S1). This supports the theory of shared ancestry between the makers of the Sambaqui culture and the speakers of proto-Ge who are hypothesized to have arrived ~2000 BP [24]. These findings also support the theory of coastal replacement of Ge speakers by Tupi-Guarani speakers after ~1000 BP [25] (Methods).

In Belize, none of the samples shows evidence of specific affinity to present-day people. A time transect study from after ~7400 BP would reveal how the types of ancestry in diverse present-day Maya language speakers began to be established in the region.

***Evidence for at Least Four Genetic Exchanges Between North and South America***

Figure 1 plots the excess rate of allele sharing of ancient Central and South Americans with the ~12800 BP *Anzick-1* individual from Montana compared to the ~11500 BP *USR1* individual from Alaska, an Ancient Beringian who derives from a lineage that split from the one leading to all other known Native Americans before they separated from each other [4] (Online Table 2). The distribution of this statistic $f_4$*(Mbuti, Test; USR1, Anzick-1)* confirms previous findings that *Anzick-1* relatedness is greatest in Central and South Americans and lowest in North American groups (Online Table 2) [6], with the exception of the California Channel Islands, where the earliest individuals from San Nicolas Island around 4900 BP show some of the highest *Anzick-1* relatedness, consistent with an early spread of *Anzick-1*-related people to these islands followed by local isolation [8] (Figure S2D).

More careful examination reveals significant ancestry variability in the ancient South Americans. The ~10900BP Los Rieles individual from Chile, the ~9600 BP individuals from Lapa do Santo in Brazil, and individuals from southern Peru and northern Chile dating to ~4200 BP and later ("Late Central Andes" from Cuncaicha, Laramate and Pica Ocho), share more alleles with *Anzick-1* than do other South Americans (Figure 1, Online Table 2). Many of these signals of asymmetrical relationship to *Anzick-1* are significant as assessed by statistics of the form $f_4$*(Mbuti, Anzick-1; Test*$_1$*, Test*$_2$*)*: Z-score for deviation from zero as high as 3.4 for the *(Test*$_1$*, Test*$_2$*)* pair *(Early Andes, Chile_LosRieles_10900BP)*, 3.1 for the pair *(Early Andes, Brazil_LapaDoSanto_9600BP)* and 3.0 for the pair *(Early Andes, Late Central Andes)* (Table S2). We confirmed these findings using *qpWave* [7], which evaluates the minimum number of sources of ancestry that must have contributed to a test set of groups relative to a set of outgroups (Methods). We tested all possible pairs of populations and found that none of the three combinations are consistent with being derived from a homogeneous ancestral population: P=0.0023 for (*Early Andes, Brazil_LapaDoSanto_9600BP*), P=0.0007 for (*Early Andes, Late Central Andes*), and P=0.0000004 for (*Brazil_LapaDoSanto_9600BP, Late Central Andes*). We obtained qualitatively similar results replacing *Brazil_LapaDoSanto_9600BP* with *Chile_LosRieles_10900BP* (Figure S4 and Online Table 3). We also obtained similar results for subsets of individuals in each group. Our power to reject models of just two sources of ancestry for the ancient South American individuals depends critically on the use of *Anzick-1* as an outgroup, as when we remove this individual from the outgroup set there is no evidence of a third source of ancestry contributing to *Brazil_LapaDoSanto_9600BP* (P=0.11) or *Chile_LosRieles_10900BP* (P=0.35). It also depends critically on the use of California Channel Islands individuals, as when we remove them as outgroups there is no evidence for a third source of ancestry contributing to *Late Central Andes* groups (P=0.12).

The fact that the three pairs each require two different sources of ancestry in order to produce a model fit could mean that they descend from a total of three (or more) distinct sources of ancestry differentially related to groups outside South America or alternatively that they are mixtures in different proportions of only two sources. To distinguish these possibilities, we used *qpWave*'s ability to test for consistency with the hypothesis that sets of three populations *(Test*$_1$*, Test*$_2$*, Test*$_3$*)* derive from just two populations relative to the same set of outgroups. *qpWave* rejects the hypothesis of two sources (P=0.0022), a result that is unlikely to be due to backflow from South America into Central America as the signal persists when we remove present-day Mexicans from the outgroup set (P=0.001) (Online Table 3). Further evidence for the robustness

of the finding of three source populations comes from the fact that the signal remains significant when we restrict to transversion polymorphisms that are not affected by cytosine-to-thymine errors (P=0.01). We caution that we did not find significant signals of ancestry heterogeneity relative to North American outgroups when repeating the *qpWave* tests on pairs of present-day populations. We speculate that this may reflect more recent homogenization leading to variation in ancestry proportions too subtle for our methods to detect.

When we add present-day *Surui* individuals into the analysis, there is evidence for a fourth source of ancestry (P=0.03) (Online Table 3), likely reflecting the same signal that led to finding "*Population Y*" ancestry in this group [5, 9].

***Modeling the Deep History of Central and South America***

We modeled the relationships among diverse ancient Americans using *qpGraph,* which evaluates whether a model of population splitting and admixture is consistent with all *f*-statistics relating pairs, triples and quadruples of groups [21].

We were able to fit genome-wide data from nine ancient North, Central and South American groups (not including *Anzick-1*) as a star-like radiation from a single source population with negligible admixture between the *ANC-A* and *ANC-B* lineages after their initial bifurcation (maximum |Z|-score for a difference between the observed and expected statistics of 2.9 (Figure 3) and 3.2 (Figure S5A); we represent *ANC-B* by the Ancient Southern Ontario population *Canada_Lucier_4800BP-500BP*). This model is not what would be expected based on the claim of a recent study [8] that major *ANC-A / ANC-B* admixture (at least ~30% of each) is necessary to model Central and South Americans. While we confirmed that the model proposed in [8] fits the data when restricting to a subset of the populations they analyzed, when we added into the model non-American populations with previously established relationships to Native Americans, the model failed (Methods). To more directly explore whether there is evidence of widespread *ANC-B* ancestry in South America, we tested whether *Canada_Lucier_4800BP-500BP* shares more alleles with a range of Central and South American *Test* populations than with *Anzick-1,* but find no evidence for a statistically significant skew (Online Table 2). Indeed, the supplementary materials of the previously reported study (Figure S13 of [8]) show that a model such as the one we favor—without widespread *ANC-B* admixture in South America—fits the data with no differences between observed and expected *f*-statistics greater than Z>2. We also find that when we explicitly model *ANC-B* admixture into the ancestors of South Americans, the inferred genetic drift specific to *Canada_Lucier_4800BP-500BP* is not significantly different from 0, providing evidence against specific affinity to *ANC-B* in South Americans (Figure S6, Methods).

To fit the *Anzick-1* genome associated with the Clovis culture into the admixture graph, we needed to specify additional admixture events. We identified a range of fits for the data. In Figure 4 we show a model obtained by manually exploring models guided by common sense principles (geography, time, and archaeology) as well as the genetic data. In Figure 5 we show a model obtained by a semi-automated procedure constrained only by the fit to the genetic data [26]. The most important difference between the two models concerns the question of how the Clovis culture associated *Anzick-1* genome relates to ancient Central and South Americans. Figure 4, which models the lineage leading to *Anzick-1* as unadmixed, seems most plausible because it is natural to expect that the oldest individuals will be least admixed, and because it is simple to explain this model via North-to-South spreads. Figure 5 models some of the ancestry of the Clovis

associated genome as deriving from within the radiation of lineages represented in South America, which if true would require a more complex history.

We highlight four points of agreement between the two admixture graphs.

First, both graphs imply a minimum of four genetic exchanges between South Americans and non-South Americans consistent with the *qpWave* results in the previous section. This includes: (a) a primary source of *ANC-A* ancestry in all South Americans; (b) an *ANC-A* lineage with distinct affinity to *Anzick-1* in *Chile_LosRieles_10900BP*, *Brazil_LapaDoSanto_9600BP*, and some early Southern Cone populations; and (c) *ANC-A* ancestry with a distinctive affinity to ancient individuals from the California Channel Islands (*USA_SanNicolas_4900BP*) present in the Central Andes by ~4200 BP (Figures S5B and S5C). In Figures 4 and 5 we do not include the *Surui* but do show such models in Figures S5G-I where *Surui* can only be fit by proposing some ancestry differently related to Eurasians than is the case for other Native Americans (as expected if there is *Population Y* ancestry in the *Surui*).

Second, both graphs specify minimal *ANC-B* ancestry in South Americans. While we do find significant allele sharing with a representative *ANC-B* population (*Canada_Lucier_4800BP-500BP*) in people from the Central Andes after ~4200 years ago—as reflected in significantly positive (2<Z<4) statistics of the form $f_4$*(Mbuti, Canada_Lucier_4800BP-500BP; Brazil_LapaDoSanto_9600BP* or *Brazil_Laranjal_6700BP, Late Central Andes or present-day Aymara and Quechua from Peru)* (Table S2, Online Table 2)—when we fit admixture graph models specifying an *ANC-B* contribution to *Late Central Andes* groups, the *ANC-B* proportion is never more than 2% (Figure S5D-F).

Third, both graphs infer little genetic drift separating the lineages leading to the different ancient groups in each major region of South America. This can be seen in our inferred five-way split whose order we cannot resolve involving lineages leading to: (a) the early Belizeans, (b) early Peruvians, (c) early Southern Cone populations, (d) the main lineage leading to the Brazilian *LapaDoSanto_9600BP*, and (e) the lineage leading to *Chile_LosRieles_10900BP* (Figure S5A). This suggests rapid human radiation upon reaching South America [5, 7].

Fourth, both graphs agree that there is distinctive shared ancestry between the Clovis culture associated *Anzick*-1 and the earliest South America individuals from Lapa do Santo in Brazil and Los Rieles in Chile. We also detect evidence of ancestry related to *Anzick-1* in the oldest Central American genome, as the most ancient individual from Belize has evidence of more *Anzick-1* relatedness than later Belize individuals as reflected in the weakly significant statistic $f_4$*(Mbuti, Anzick-1; Belize_SakiTzul_7400BP, Belize_MayahakCabPek_9300BP)* (Z=2.1). Taken together, these results support the hypothesis that an expansion of a group associated with the Clovis culture left an impact far beyond the geographic region in which this culture was spread [27]. At the same time, both classes of models provide evidence against a stronger version of this hypothesis, which is that an expansion of a homogeneous population associated with the Clovis culture was the primary source of the ancestry of later Central and South Americans. Specifically, both models find that the overwhelming majority of the ancestry of most Central and South Americans derives from one or more lineages without the *Anzick-1* affinities present at Lapa do Santo. Thus, a different *ANC-A* lineage from the one represented in *Anzick-1* made the most important contribution to South Americans, and there must have been a population turnover in the mid-Holocene that largely replaced groups such as the ones represented by the ~10900 BP individual at Los Rieles in Chile and the ~9600 BP individuals at Lapa do Santo in Brazil. This

genetic evidence of a major population turnover correlates with the findings from morphological studies of a population in Brazil around this time [28].

It is tempting to hypothesize that the early branching *ANC-A* lineages that we have shown contributed most of the ancestry of Central and South Americans today—and that harbor no specific *Anzick-1* association—contributed to the people who lived at the site of Monte Verde in southern Chile and whose material artifacts have been dated to a pre-Clovis period at least ~14500 BP [29]. However, since all the earliest Central and South American individuals show affinities to *Anzick-1*, our results could also be consistent with a scenario in which nearly all the ancestry of the South Americans genomes derives from population movements from North America that began no earlier than the Clovis period. In either case, we demonstrate that the non-*Anzick-1* associated ancestry type began to spread in South America by at least 9000 BP, the date of the oldest genomes that have no specific Anzick-1 affinity (from Cuncaicha and Lauricocha in the Central Andes).

***All the Ancient South Americans Descend from the Same Eurasian Source Population***

Previous studies have suggested that the present-day groups like *Surui* from Amazonia harbor ancestry from a source termed "*Population Y*" [5, 9], which shared alleles at an elevated rate with Australasian groups (*Onge*, *Papuan*, and *Australians*) as well as the ~40000 BP Tianyuan individual from China [10]. We tested for this signal in the ancient South American individuals with statistics of the form $f_4$*(Mbuti, Australasian; X, Mixe or ancient South American),* and while we replicated the originally reported signal when *X* was present-day *Karitiana* or *Surui*, we could not detect a signal when *X* was any of the ancient South Americans (Online Table 4). We also studied the statistic $f_4$*(Mbuti, Tianyuan; Ancient*$_1$*, Ancient*$_2$*)* to test if any ancient individual is differentially related to Tianyuan [10], but no statistic was significant (Online Table 4). We finally applied *qpWave* to all pairs of South American groups, testing whether they were homogeneously related to a set of diverse non-Native American outgroups (*Mbuti*, *Han*, *Onge*, *French*, and *Papuan*) and found no pair of ancient South Americans that consistently gave significant signals ($p<0.01$), as expected if all the ancient South Americans we analyzed derived from the same stem Native American population (Online Table 4). Our failure to find significant evidence of Australasian or Paleolithic East Asian affinities in any of the ancient Central and South American individuals raises the question of what ancient populations could have contributed the *Population Y* signal in *Surui* and other Amazonian groups and increases the previously small chance that this signal—despite the strong statistical evidence for it—was a false positive. A priority is to search for the *Population Y* signal in additional ancient genomes.

Our finding of no excess allele sharing with non-Native American populations in the ancient samples is also striking as many of these individuals—including those at Lapa do Santo—have a "Paleoamerican" cranial morphology that has been suggested to be evidence of the spread of a substructured population of at least two different Native American source populations from Asia to the Americas [30]. Our finding that early Holocene individuals with such a morphology are consistent with deriving all their ancestry from the same homogeneous ancestral population as other Native Americans extends the finding of Raghavan *et al.* 2015 who came to a similar conclusion after analyzing Native Americans inferred to have Paleoamerican morphology who lived within the last millennium.

***Single Locus Analysis***

The D4h3a mtDNA haplogroup has been hypothesized to be a marker for an early expansion into the Americas along the Pacific coast [31]. However, its presence in two Lapa do Santo individuals and *Anzick-1* [6] makes this hypothesis highly unlikely (Figure S7, Methods, Online Table 1).

The patterns we observe on the Y chromosome also force us to revise our understanding of the origins of present-day variation. Our ancient DNA analysis shows that the Q1a2a1b-CTS1780 haplogroup, which is currently rare, was present in a third of the ancient South Americas. In addition, our observation of the currently extremely rare C2b haplogroup at Lapa do Santo disproves the suggestion that it was introduced after 6000 BP [32].

The patterns of variation at phenotypically significant variants are also notable. Our data show that a variant in *EDAR* that affects tooth shape, hair follicles and thickness, sweat, and mammary gland ductal branching and that occurs at nearly 100% frequency in present day Native Americans and East Asians [33] was not fixed in *USR1*, *Anzick-1*, a *Brazil_LapaDoSanto_9600BP* individual and a *Brazil_Laranjal_6700BP* individual, all of whom carry the ancestral allele (Online Table 5). Thus, the derived allele rose in frequency in parallel in both East Asians and in Native Americans. In contrast at *FADS2*, one of the variants at a polymorphism (rs174570) associated with fatty acid desaturase 2 levels is derived in all the ancient individuals, supporting the hypothesis that the selective sweep that drove it to near fixation was complete prior to the peopling of the Americas [34].

# Discussion

Our finding of two previously undocumented genetic exchanges between North and South America has significant implications for models of the peopling of the Americas.

Most important, our discovery that the Clovis-associated *Anzick-1* genome at ~12800 BP shares distinctive ancestry with the oldest Chilean, Brazilian and Belizean individuals supports the hypothesis that the expansion of people who spread the Clovis culture in North America also affected Central and South America, as expected if the spread of the Fishtail Complex in Central and South America and the Clovis Complex in North America were part of the same early Paleoindian phenomenon (direct confirmation would require ancient DNA from a Fishtail-context) [35]. However, the fact that the great majority of ancestry of later South Americans lacks specific affinity to *Anzick-1* rules out the hypothesis of a homogeneous founding population. Thus, if Clovis-related expansions were responsible for the peopling of South America, it must have been a complex scenario involving arrival in the Americas of sub-structured lineages with and without specific *Anzick-1* affinity, with the one with *Anzick-1* affinity making a minimal long-term contribution. While we cannot at present determine when the non-*Anzick*-1 associated lineages first arrived in South America, we can place an upper bound on the date of the spread to South America of all the lineages represented in our sampled ancient genomes as all are *ANC-A* and thus must have diversified after the *ANC-A / ANC-B* split estimated to have occurred ~17500-14600 BP [4].

A second notable finding of this study is our evidence that the ancient individuals from the California Channel Islands have distinctive and significant allele sharing with groups that became widespread over the Central Andes after ~4200 BP. There is no archaeological evidence

of large-scale cultural exchange between North and South America around this time, but it is important to recognize that ~4200 BP is a minimum date for the exchange between North and South American that drove this pattern; the gene flow itself could have occurred thousands of years before and the ancestry deriving from it could have persisted in a region of South America not yet sampled with ancient DNA. The evidence of an expansion of this ancestry type in the Central Andes by ~4200 BP is notable in light of the increasing density of sites in this region at approximately this time, a pattern that is consistent with a demographic expansion of a previously more restricted population [36].

We conclude by highlighting several limitations of this study. First, all the individuals we newly report have a date <11000 BP and thus we could not directly probe the initial movements of people into Central and South America. Second, from the period between 11000-3000 BP that includes most of our individuals, we lacked ancient data from Amazonia, northern South America, and the Caribbean and thus cannot determine how individuals from these regions relate to the ones we analyzed. Third, because we reported few individuals from after 3000 BP, this study provides just a glimpse of the power of this type of analysis to reveal more recent events. Regionally focused studies with large sample sizes are needed to realize the potential of ancient DNA to reveal how the human diversity of this region came to be the way it is today.

## Acknowledgements:

In Belize we thank the Institute of Archaeology, our Maya collaborators, and the Ya'axché Conservation Trust for providing opportunities to consult with their community. In Argentina we thank the Comunidad Indígena Mapuche-Tehuelche Cacique Pincen for supporting sampling of skeletal material from Laguna Chica. In Peru we thank Johny Isla-Cuadrado and the National Museum of Archaeology, Anthropology and History for permission to sample skeletal material. In Brazil we thank all of the individuals who participated in the Lagoa Santa studies; J. Hein, R. Tavares de Oliveira, J. Bárbara Filh; and institutional support from the IEF, IPHAN IBAMA, and the cities of Matozinhos, Lagoa Santa and Pedro Leopoldo.

We thank S. Fiedel, Q. Fu, C. Jeong, T. Kivisild, M. Lipson, V. Narasimhan, I. Olalde, B. Potter, A. Scally, C. Scheib, O. Semino, and T. Stafford for critical comments. We are grateful to M. O'Reilly for graphic design help. We thank the wet laboratory and computational teams at MPI-SHH and at Harvard Medical School.

L. Fe. was supported by the Wenner-Gren Foundation (SC-14-62), a Hellman Foundation fellowship, and the NSF (A15-0187-001). N.N. was supported by the NIGMS (GM007753). C.M. and A.N.D. were supported by FONDECYT #1170408 and CONICYT R17A10002. Archaeological research in Argentina was funded by the National Geographic Society (9773-15), ANPCYT (PICT 2015-2070) and CONICET (PIP Nº 0414). B.L., W.H., and A.C. were supported by the ARC, and received sequencing support funding from The Environment Institute at Adelaide University. Funding for the Belize research to K.M.P. and D.J.K. came from the NSF (BCS1632061, BCS1632144), to K.M.P. from the Alphawood Foundation, and to D.J.K. and B.J.C. from the NSF Archaeometry program (BCS-1460369). The Cuncaicha work was supported by an Alexander von Humboldt Foundation fellowship (K.R.), the DFG (FOR 2237, INST 37/706, FOR 2237), the Pontifical Catholic University of Lima, and Northern Illinois University. D.C. is Superior Researcher of CONICET. Financial support for research in Brazil was provided by FAPESP (99/12684-2,

04/01321-6, 04/11038-0, and 2017/16451-2). D.R. was supported by the NSF (BCS-1032255), the NIGMS (GM100233), by an Allen Discovery Center grant, and is an Investigator of the Howard Hughes Medical Institute.

This manuscript is dedicated to the National Museum of Brazil whose irreplaceable collections of natural and cultural material were lost in the fire of September 2, 2018.

## Author Contributions:

Conceptualization, C.P., N.N., I.L., P.S., A.C., T.F., T.H., N.P., S.S., J.S., M.H., A.S., L.Fe., J.K. and D.R.; Formal Analysis, C.P., N.N., I.L., P.S., T.C.L., E.B. and C.C.; Investigation, C.P., N.R., K.N., N.A., N.B., B.J.C., M.Fe., A.F., W.H., Ke.H., T.K.H., A.M.L., B.L., M.M., E.N., J.O., K.S., S.T. and L.Fe.; Resources, J.H., Ka.H., A.N.D., J.B., M.Fr., P.K., H.R., K.R., W.R.T., M.Ro., S.M.G., K.M.P., D.C.S., E.N.C., L.M.P.G., M.L.A., A.L., M.I., R.E.O., D.B., A.B., V.W., N.A.S., M.A.R, C.R.P., P.G.M., L.Fi., D.C., C.S., S.E., P.D., M.Re., C.M., G.P., E.T., D.J.K. and A.S.; Data Curation, C.P., N.N., S.M. and D.R.; Writing, C.P., N.N., A.S., L.Fe. and D.R.; Supervision, D.J.K., A.S., L.Fe., J.K. and D.R.

## Declaration of Interests:

The authors declare no competing interests.

## References:

1. Del Castillo L: **Property rights in peasant communities in Peru.** In *Peruvian Center of Social Studies-Centro Peruano de Estudios Sociales-CEPES Colloque international "Les frontières de la question foncière–At the frontier of land issues", Montpellier*. 2006
2. Knapp G: **Linguistic and cultural geography of contemporary Peru.** 1987.
3. Mallon FE: *Peasant and nation: the making of postcolonial Mexico and Peru.* Univ of California Press; 1995.
4. Moreno-Mayar JV, Potter BA, Vinner L, Steinrucken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas AS, Sikora M, et al: **Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans.** *Nature* 2018, **553:**203-207.
5. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas AS, et al: **POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans.** *Science* 2015, **349:**aab3884.
6. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al: **The genome of a Late Pleistocene human from a Clovis burial site in western Montana.** *Nature* 2014, **506:**225-229.
7. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al: **Reconstructing Native American population history.** *Nature* 2012, **488:**370-374.

8. Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Morseburg A, Johnson JR, Potter A, et al: **Ancient human parallel lineages within North America contributed to a coastal expansion.** *Science* 2018, **360:**1024-1027.
9. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D: **Genetic evidence for two founding populations of the Americas.** *Nature* 2015, **525:**104-108.
10. Yang MA, Gao X, Theunert C, Tong H, Aximu-Petri A, Nickel B, Slatkin M, Meyer M, Paabo S, Kelso J, Fu Q: **40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia.** *Curr Biol* 2017, **27:**3202-3208 e3209.
11. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.
12. Gansauge M-T, Meyer M: **Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA.** *Nature protocols* 2013, **8:**737.
13. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D: **Partial uracil-DNA-glycosylase treatment for screening of ancient DNA.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20130624.
14. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al: **An early modern human from Romania with a recent Neanderthal ancestor.** *Nature* 2015, **524:**216-219.
15. Bardill J, Bader AC, Garrison NA, Bolnick DA, Raff JA, Walker A, Malhi RS, Summer internship for IpiGC: **Advancing the ethics of paleogenomics.** *Science* 2018, **360:**384-385.
16. Herrera A: *Indigenous Archaeology…in Peru?* New York: Routledge; 2011.
17. Silverman H: **Cultural Resource Management and Heritage Stewardship in Peru.** *CRM: Journal of Heritage Stewardship* 2006, **3:**57-72.
18. Renaud G, Slon V, Duggan AT, Kelso J: **Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA.** *Genome Biol* 2015, **16:**224.
19. Korneliussen TS, Albrechtsen A, Nielsen R: **ANGSD: Analysis of Next Generation Sequencing Data.** *BMC Bioinformatics* 2014, **15:**356.
20. Eisenmann S, Bánffy E, van Dommelen P, Hofmann KP, Maran J, Lazaridis I, Mittnik A, McCormick M, Krause J, Reich D: **Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis.** *Scientific reports* 2018, **8:**13003.
21. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
22. Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A, et al: **Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas.** *Sci Adv* 2016, **2:**e1501385.
23. Lindo Jea: **The genetic prehistory of the Andean highlands 7,000 Years BP through European contact.** 2018.

24. Iriarte J, DeBlasis P, De Souza JG, Corteletti R: **Emergent complexity, changing landscapes, and spheres of interaction in southeastern South America during the middle and late Holocene.** *Journal of Archaeological Research* 2017, **25:**251-313.
25. Hubbe M, Neves WA, de Oliveira EC, Strauss A: **Postmarital residence practice in southern Brazilian coastal groups: Continuity and change.** *Latin American Antiquity* 2009, **20:**267-278.
26. Lazaridis Iea: **Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry.** 2018.
27. Fiedel SJ: **The Anzick genome proves Clovis is first, after all.** *Quaternary International* 2017, **444:**4-9.
28. Hubbe M, Okumura M, Bernardo DV, Neves WA: **Cranial morphological diversity of early, middle, and late H olocene B razilian groups: Implications for human dispersion in B razil.** *American journal of physical anthropology* 2014, **155:**546-558.
29. Dillehay TD, Ramirez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD: **Monte Verde: seaweed, food, medicine, and the peopling of South America.** *Science* 2008, **320:**784-786.
30. von Cramon-Taubadel N, Strauss A, Hubbe M: **Evolutionary population history of early Paleoamerican cranial morphology.** *Science advances* 2017, **3:**e1602289.
31. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Hooshiar Kashani B, Ritchie KH, Scozzari R, Kong QP, et al: **Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups.** *Curr Biol* 2009, **19:**1-8.
32. Roewer L, Nothnagel M, Gusmão L, Gomes V, González M, Corach D, Sala A, Alechine E, Palha T, Santos N: **Continent-wide decoupling of Y-chromosomal genetic variation from language and geography in native South Americans.** *PLoS genetics* 2013, **9:**e1003460.
33. Kamberov YG, Wang S, Tan J, Gerbault P, Wark A, Tan L, Yang Y, Li S, Tang K, Chen H, et al: **Modeling recent human evolution in mice by expression of a selected EDAR variant.** *Cell* 2013, **152:**691-702.
34. Amorim CEG, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hünemeier T: **Genetic signature of natural selection in first Americans.** *Proceedings of the National Academy of Sciences* 2017, **114:**2195-2199.
35. Pearson GA: **Bridging the Gap: An Updated Overview of Clovis across Middle America and its Techno-Cultural Relation with Fluted Point Assemblages from South America.** *PaleoAmerica* 2017, **3:**203-230.
36. Goldberg A, Mychajliw AM, Hadly EA: **Post-invasion demography of prehistoric humans in South America.** *Nature* 2016, **532:**232-235.
37. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E: **Reconstructing the deep population history of Central and South America.** *Cell* 2018, **175:**1185-1197. e1122.
38. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, Rohland N, Mallick S, Stewardson K, Kistler L, et al: **Archaeogenomic evidence reveals prehistoric matrilineal dynasty.** *Nat Commun* 2017, **8:**14115.

39. Lohse JC, Madsen DB, Culleton BJ, Kennett DJ: **Isotope paleoecology of episodic mid-to-late Holocene bison population expansions in the Southern Plains, USA.** *Quaternary Science Reviews* 2014, **102:**14-26.
40. Talamo S, Richards M: **A comparison of bone pretreatment methods for AMS dating of samples> 30,000 BP.** *Radiocarbon* 2011, **53:**443-449.
41. Korlević P, Talamo S, Meyer M: **A combined method for DNA analysis and radiocarbon dating from a single sample.** *Scientific reports* 2018, **8:**4127.
42. Ramsey CB, Lee S: **Recent and planned developments of the program OxCal.** *Radiocarbon* 2013, **55:**720-730.
43. Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Ramsey CB, Buck CE, Cheng H, Edwards RL, Friedrich M: **IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1869-1887.
44. Hogg AG, Hua Q, Blackwell PG, Niu M, Buck CE, Guilderson TP, Heaton TJ, Palmer JG, Reimer PJ, Reimer RW: **SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1889-1903.
45. Reimer PJ, Reimer RW: **A marine reservoir correction database and on-line interface.** *Radiocarbon* 2001, **43:**461-463.
46. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S: **Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA.** *Nucleic acids research* 2009, **38:**e87-e87.
47. https://github.com/jstjohn/SeqPrep.
48. Schubert M, Lindgreen S, Orlando L: **AdapterRemoval v2: rapid adapter trimming, identification, and read merging.** *BMC research notes* 2016, **9:**88.
49. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
50. Stenzel U: **https://bitbucket.org/ustenzel/biohazard.** 2018.
51. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K: **EAGER: efficient ancient genome reconstruction.** *Genome biology* 2016, **17:**60.
52. Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L: **mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters.** *Bioinformatics* 2013, **29:**1682-1684.
53. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M: **Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal.** *Proceedings of the National Academy of Sciences* 2014, **111:**2229-2234.
54. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al: **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.** *Nature* 2016, **538:**201-206.
55. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al: **Ancient human genomes suggest three ancestral populations for present-day Europeans.** *Nature* 2014, **513:**409-413.
56. Jostins L, Xu Y, McCarthy S, Ayub Q, Durbin R, Barrett J, Tyler-Smith C: **YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data.** *arXiv preprint arXiv:14077988* 2014.

57. Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C: **HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment.** *Hum Mutat* 2013, **34:**1189-1194.
58. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic acids research* 2004, **32:**1792-1797.
59. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: molecular evolutionary genetics analysis version 6.0.** *Molecular biology and evolution* 2013, **30:**2725-2729.
60. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, et al: **Beringian standstill and spread of Native American founders.** *PLoS One* 2007, **2:**e829.
61. Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, Mitchell J, Worl R, Dixon EJ, Fifield TE, et al: **Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity.** *Proc Natl Acad Sci U S A* 2017, **114:**4093-4098.
62. Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, et al: **Do the four clades of the mtDNA haplogroup L2 evolve at different rates?** *Am J Hum Genet* 2001, **69:**1348-1356.
63. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
64. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
65. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.
66. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
67. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39:**W475-478.
68. Lipson M, Reich D: **A working model of the deep relationships of diverse modern human genetic lineages outside of Africa.** *Molecular biology and evolution* 2017, **34:**889-902.
69. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K: **Genomic insights into the origin of farming in the ancient Near East.** *Nature* 2016, **536:**419.
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
71. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987-2993.
72. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sanchez E, Nielsen R: **Archaic Adaptive Introgression in TBX15/WARS2.** *Mol Biol Evol* 2017, **34:**509-524.

73. Consortium STD, Williams AL, Jacobs SB, Moreno-Macias H, Huerta-Chagoya A, Churchhouse C, Marquez-Luna C, Garcia-Ortiz H, Gomez-Vazquez MJ, Burtt NP, et al: **Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico.** *Nature* 2014, **506:**97-101.
74. Fehren-Schmitz L, Georges L: **Ancient DNA reveals selection acting on genes associated with hypoxia response in pre-Columbian Peruvian Highlanders in the last 8500 years.** *Sci Rep* 2016, **6:**23485.
75. Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA, et al: **Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans.** *Am J Hum Genet* 2017, **101:**752-767.
76. Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, Korneliussen TS, Gerbault P, Skotte L, Linneberg A: **Greenlandic Inuit show genetic signatures of diet and climate adaptation.** *Science* 2015, **349:**1343-1347.

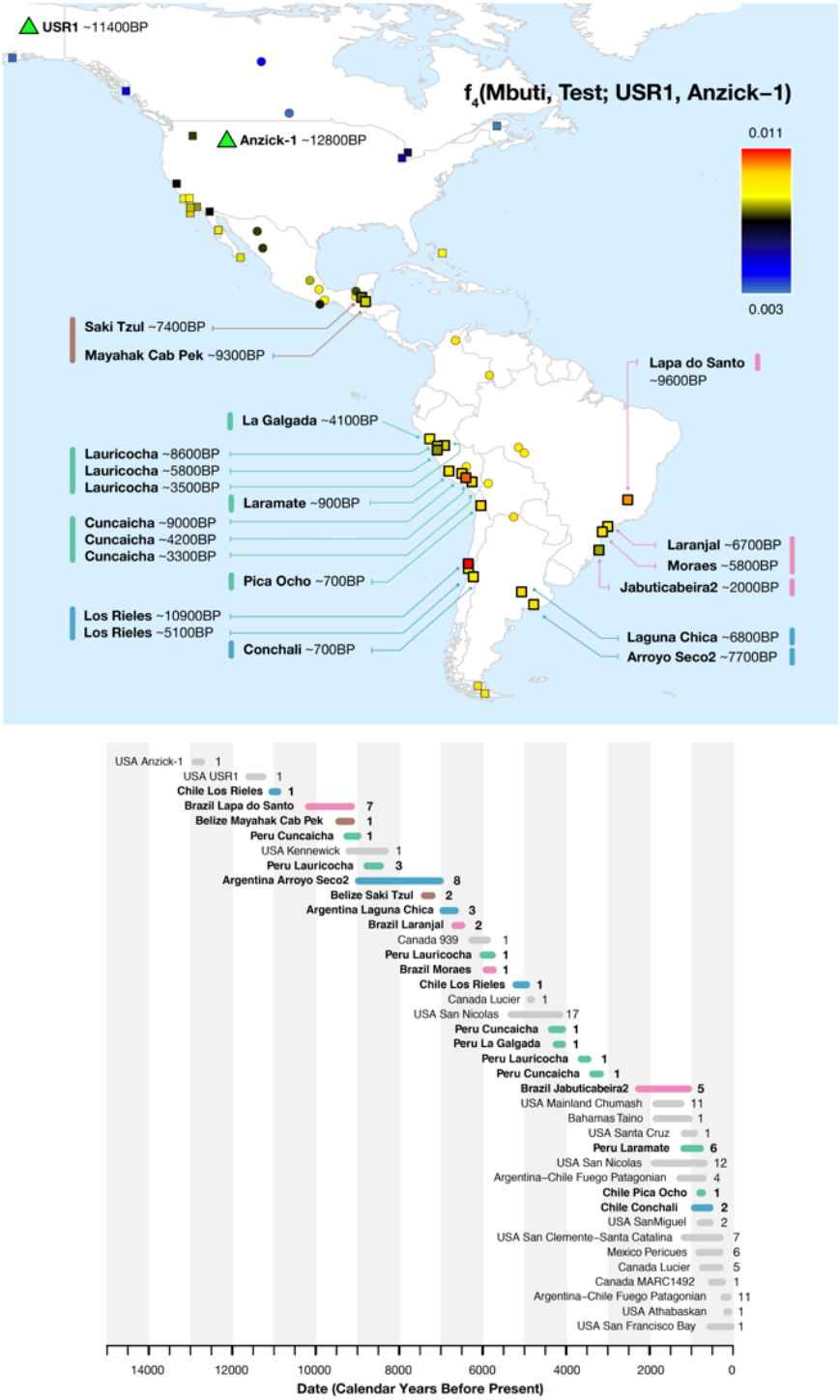


**Figure 1. Geographic locations and time ranges**. **A**) Color coding is based on the value of $f_4$*(Mbuti, Test; USR1, Anzick-1)*, which measures the degree of allele sharing of each "Test" population with *Anzick-1* compared to the Ancient Beringian *USR1* (the latter two plotted as green triangles). All values and standard errors are listed in Online Table 2. Present-day individuals are circles and ancient individuals are squares (the newly reported individuals are indicated with a thick black outline). **B**) We show previously published (gray) and newly reported ancient data: magenta = Brazil; brown = Belize; green = Peru/northern Chile; blue = Southern Cone. The numbers give sample size in each grouping.

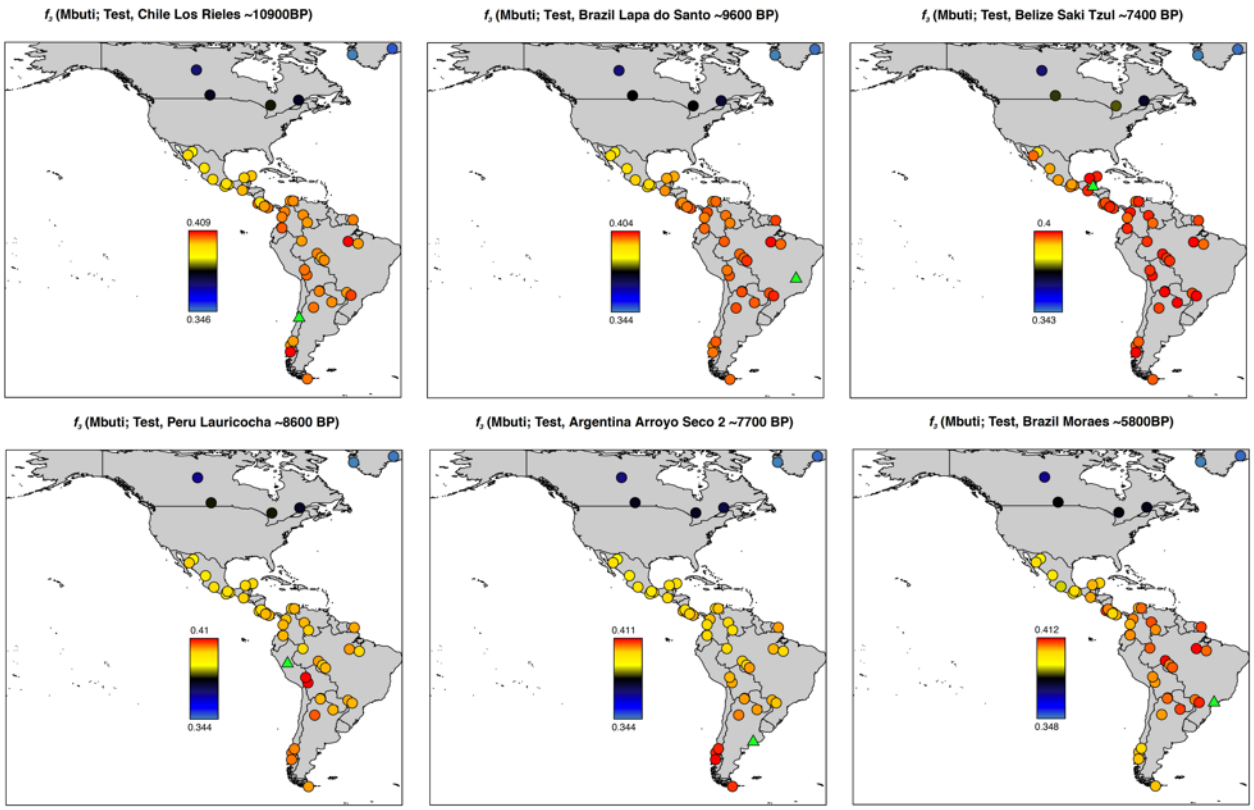


**Figure 2. Relatedness of ancient to present-day people.** Allele sharing statistics of the form $f_3$(*Mbuti; Test, Ancient*), where the "Ancient" individuals represented by a green triangle are *Chile_LosRieles_10900BP*, *Brazil_LapaDoSanto_9600BP*, *Belize_SakiTzul_7400BP*, *Peru_Lauricocha_8600BP*, *Argentina_ArroyoSeco2_7700BP*, and *Brazil_Moraes_5800BP*.

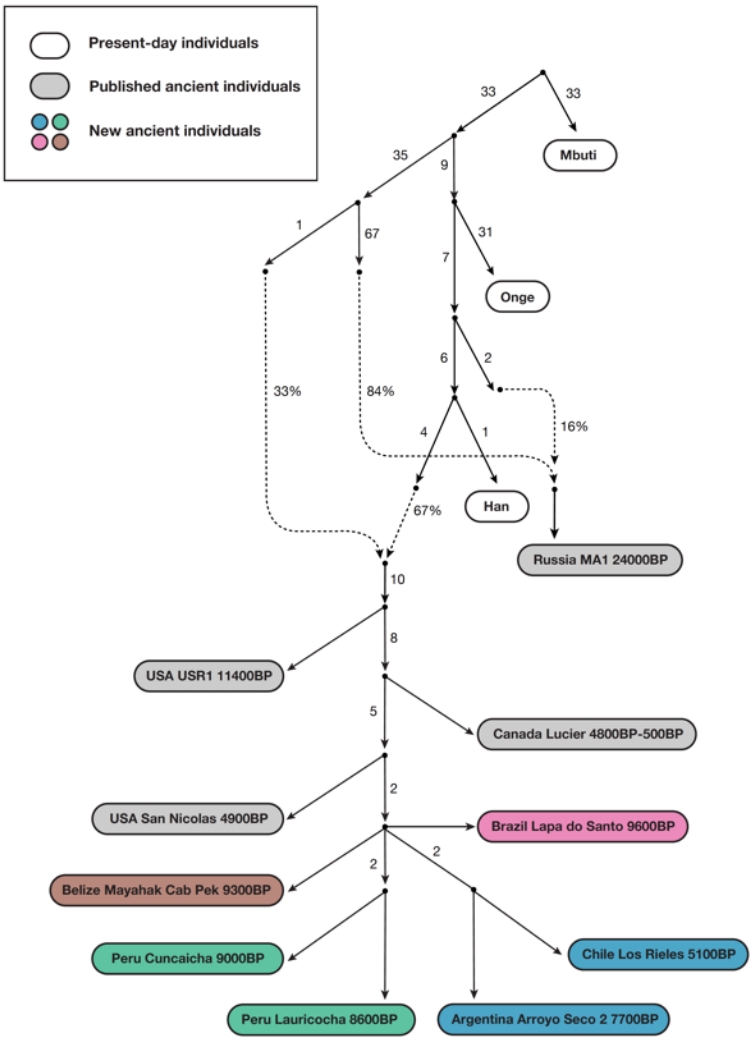


**Figure 3. Skeleton model that fits the data with minimal admixture.** This graph models nine of the ancient North, Central and South American groups without admixture (branch lengths are in units of $F_{ST}$ x 1000). The maximum deviation between observed and expected *f*-statistics is Z=2.9 (Z=3.1 when restricting to transversions). Drift lengths in the terminal edges are unlabeled as randomly sampling an allele to represent each individual makes them artifactually long.

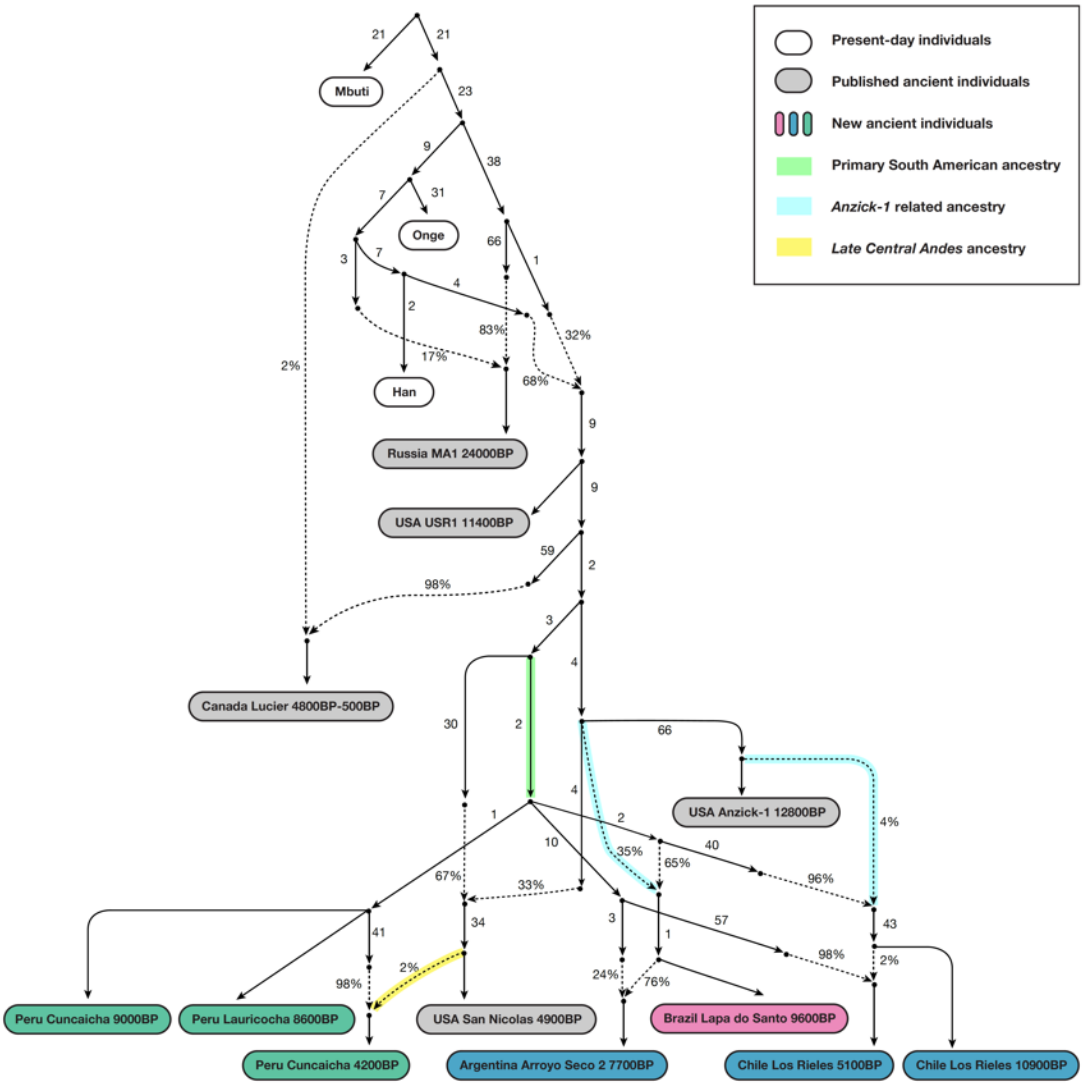


**Figure 4. Adding in the ~12800 BP Anzick-1 and ~10900 BP Los Rieles.** We used Figure 3 that models all analyzed Native Americans as unadmixed as a framework graph (excluding *Belize_MayahakCabPek_9300BP* because of relatively low coverage). We then added in *Anzick-1 and Chile_LosRieles_10900BP*. This model specifies three sources of North American related ancestry in South America, indicated by color-coding (*Population Y* ancestry is not included but Figures S5B-I show related fits some of which do include it). The maximum deviation between observed and expected *f*-statistics is Z=3.4 (Z=3.0 when restricting to transversions). The inferred 2% West Eurasian admixture into *Canada Lucier 4800-500BP* is most likely explained by contamination in these samples by people of European ancestry.

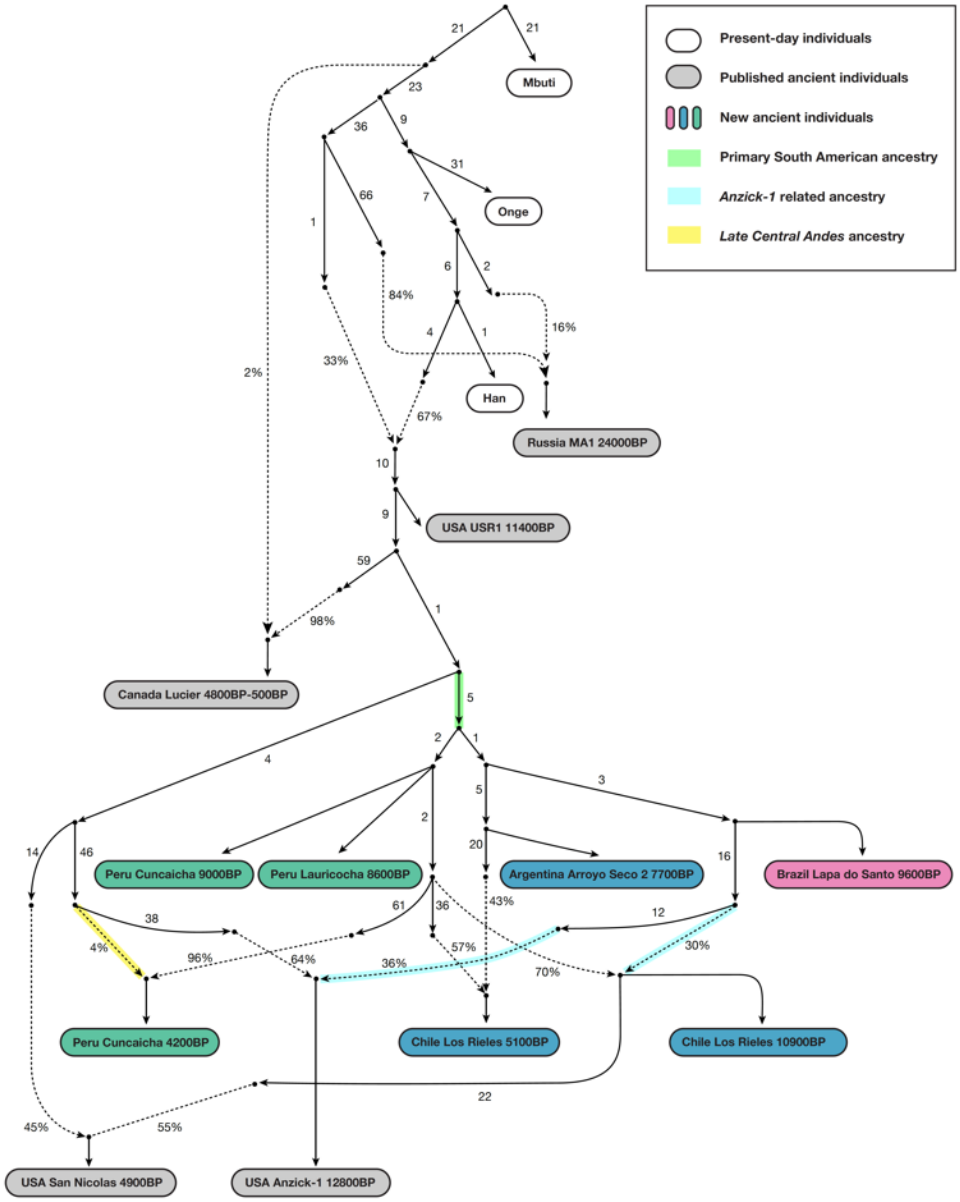


**Figure 5. An alternative fitting admixture graph obtained by a semi-automated method.** We also applied a semi-automated approach that aims to fit population relationships while minimizing the number of admixture events (Methods) [26]. This is less plausible than Figure 4 on archaeological grounds, but it has a lower maximum Z-score for the same number of admixture edges (Z=2.9 for all sites, Z=2.9 when restricting to transversions). Like Figure 4, this model specifies a minimum of three genetic exchanges between North and South America, indicated here by color-coding.

# Materials and Methods:

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David Reich (reich@genetics.med.harvard.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

**Archaeological site information**

We generated new genome-wide data from skeletal remains of 49 ancient individuals: 15 from Peru, 3 from Belize, 5 from Chile, 11 from Argentina, and 15 from Brazil.

Arroyo Seco 2, Argentina (n=8)
Laguna Chica, Argentina (n=3)
Mayahak Cab Pek, Belize (n=1)
Saki Tzul, Belize (n=2)
Jabuticabeira 2, Brazil (n=5)
Lapa do Santo, Brazil (n=7)
Laranjal, Brazil (n=2)
Moraes, Brazil (n=1)
Los Rieles, Central Chile (n=2)
Conchali, Santiago, Central Chile (n=2)
Pica Ocho, Northern Chile (n=1)
Cuncaicha, Highlands, Peru (n=3)
La Galgada, Highlands, Peru (n=1)
Laramate, Highlands, Peru (n=6)
Lauricocha, Highlands, Peru (n=5)

Details of the ancient individuals and their archaeological contexts can be found in Posth*, Nakatsuka* *et al*., 2018 [37].

**Method Details:**

**Direct AMS $^{14}$C bone dates:**

We report 31 new direct AMS $^{14}$C bone dates from eleven radiocarbon laboratories (Arizona [AA] – 1; Mannheim [MAMS] – 18; Poznan [Poz] – 1; Pennsylvania State University [PSUAMS] – 6; UC Irvine [UCIAMS] - 4; University of Georgia [UGAMS] – 1) and recalibrate 23 previously published radiocarbon dates (Online Table 1 of Posth*, Nakatsuka* *et al*., 2018 [37]). Bone preparation and quality control methods for most of these samples are described elsewhere and the details can be found on laboratory-specific websites. Detailed methods are provided below for PSUAMS, UCIAMS and MAMS.

**PSUAMS and UCIAMS Protocols and Pretreatment**

*Ultrafiltration*: At PSUAMS and UCIAMS, bone collagen for $^{14}C$ and stable isotope analyses was extracted and purified using a modified Longin method with ultrafiltration [38]. Bones were initially cleaned of adhering sediment and the exposed surfaces were removed with an X-acto blade. Samples (200–400 mg) were demineralized for 24–36 h in 0.5N HCl at 5 °C followed by a brief (<1 h) alkali bath in 0.1N NaOH at room temperature to remove humates. The residue was rinsed to neutrality in multiple changes of Nanopure $H_2O$, and then gelatinized for 12 h at 60 °C in 0.01N HCl. The resulting gelatin was lyophilized and weighed to determine percent yield as a first evaluation of the degree of bone collagen preservation. Rehydrated gelatin solution was pipetted into pre-cleaned Centriprep ultrafilters (retaining 430 kDa molecular weight gelatin) and centrifuged 3 times for 20 min, diluted with Nanopure $H_2O$, and centrifuged 3 more times for 20 min to desalt the solution.

*XAD Amino Acids*: In some instances, collagen samples were too poorly preserved and were pre-treated at Penn State using a modified XAD process [39]. Samples were physically cleaned using hand tools and sectioned with disposable Dremel cut-off wheels and then demineralized in 0.5 N HCl for 2-3 days at 5°C. The demineralized collagen pseudomorph was gelatinized at 60°C in 1-2 mL 0.01 N HCl for eight to ten hours. Sample gelatin was pipetted into a pre-cleaned 10 mL disposable syringe with an attached 0.45 mm Millex Durapore PVDF filter (precleaned with methanol and Nanopure H2O) and driven into a thick-walled culture tube. The filtered solution was lyophilized and percent gelatinization and yield determined by weight. The sample gelatin was then hydrolyzed in 2 mL 6 N HCl for 22 h at 110°C. Supelco ENVI-Chrom® SPE (Solid Phase Extraction; Sigma-Aldrich) columns were prepped with 2 washes of HCl (2 mL) and rinsed with 10 mL DI H2O. With a 0.45 mm Millex Durapore filter attached, the SPE Column was equilibrated with 50 mL 6 N HCl and the washings discarded. 2 mL collagen hydrolyzate as HCl was pipetted onto the SPE column and driven with an additional 10 mL 6 N HCl dropwise with the syringe into a 20 mm culture tube. The hydrolyzate was finally dried into a viscous syrup by passing UHP $N_2$ gas over the sample heated at 50°C for ~12 h.

**PSUAMS and UCIAMS Quality Control and Measurement**

Carbon and nitrogen concentrations and stable isotope ratios of the XAD amino acid samples were measured at the Yale Analytical and Stable Isotope Center with a Costech elemental analyzer (ECS 4010) and Thermo DeltaPlus analyzer [38]. Sample quality was evaluated by % crude gelatin yield, %C, %N and C/N ratios before AMS $^{14}C$ dating. C/N ratios for all samples fell between 2.9 and 3.6, indicating good collagen preservation. Samples (~2.1 mg) were then combusted for 3 h at 900°C in vacuum-sealed quartz tubes with CuO and Ag wires. Sample $CO_2$ was reduced to graphite at 550°C using $H_2$ and a Fe catalyst, with reaction water drawn off with $Mg(ClO_4)_2$. Graphite samples were pressed into targets in Al boats and loaded on a target wheel with OX-1 (oxalic acid) standards, known-age bone secondaries, and a $^{14}C$-free Pleistocene whale blank. $^{14}C$ measurements were at UCIAMS on a modified National Electronics Corporation compact spectrometer with a 0.5 MV accelerator (NEC 1.5SDH-1). The $^{14}C$ ages were corrected for mass-dependent fractionation with measured $\delta^{13}C$ values and calibrated with samples of Pleistocene whale bone (backgrounds, 48000 $^{14}C$ BP), late Holocene bison bone (~1,850 $^{14}C$ BP),

late AD 1800s cow bone and OX-2 oxalic acid standards.

**PSUAMS Acid Etch/Hydrolysis**

Enamel bioapatite splits from two samples from Saki Tzul, Belize (I5456 and I5457) were processed at the Pennsylvania State University for AMS $^{14}C$ dating by acid hydrolysis. Samples were cleaned with dental tools to remove adhering residues, and then acid-etched to removed secondary carbonate prior to hydrolysis. After rinsing in Nanopure $H_2O$ and drying at 50°C, samples were evaluated for the integrity of their enamel using FTIR analysis. Samples and standards were placed then in BD Vacutainer septum-stopper vials, and digested with 85% orthophosphoric acid. The evolved $CO_2$ was graphitized as above and the $^{14}C$ measurement made on a modified National Electronics Corporation compact spectrometer with a 0.5 MV accelerator (NEC 1.5SDH-1) at Penn State University.

**MAMS Protocols and Pretreatment**

The samples from Lapa do Santo MAMS-28703, -28706, and -17190 were pretreated at the Department of Human Evolution at the Max Planck Institute for Evolutionary Anthropology (MPI-EVA), Leipzig, Germany, using the method described in [40]. The outer surface of the bone samples was first cleaned by a shot blaster and then 500 mg of the whole bone was taken. The samples were then decalcified in 0.5M HCl at room temperature until no $CO_2$ effervescence was observed, usually for about 4 hours. 0.1M NaOH was added for 30 minutes to remove humic acids. The NaOH step was followed by a final 0.5M HCl step for 15 minutes. The resulting solid was gelatinized at pH3 in a heater block at 75°C for 20h. The gelatine was then filtered in an Eeze-Filter™ (Elkay Laboratory Products (UK) Ltd.) to remove small (>80μm) particles. The gelatine was then ultrafiltered with Sartorius "VivaspinTurbo" 30 KDa ultrafilters. Prior to use, the filter was used to remove carbon containing humectants. The samples were lyophilized for 48 hours. All dates were corrected for a residual preparation background estimated from $^{14}C$ free bone samples. These bones were kindly provided by the Mannheim laboratory and pretreated in the same way as the archaeological samples [41]. To assess the preservation of the collagen, C:N ratios together with isotopic values need to be evaluated. The C:N ratio should be between 2.9 and 3.6 and the collagen yield not less than 1% of the weight. Stable isotopic analysis is evaluated at MPI-EVA, Leipzig (Lab Code R-EVA), using a ThermoFinnigan Flash EA coupled to a Delta V isotope ratio mass spectrometer. All the samples fall within the acceptable range of the evaluation criteria mentioned above.

**Calibration of radiocarbon dates**

All calibrated $^{14}C$ ages were calculated using OxCal version 4.3 [42]. The IntCal13 northern hemisphere curve [43] was used for four samples from Belize, while the remainder were calibrated using the SHCal13 curve [44]. Dates from two coastal sites—Los Rieles in Chile and Jabuticabeira II in Brazil—were calibrated according to a mixture with the Marine13 curve [43] based on an estimate of a 40% marine dietary component. For each site, ΔR values were calculated based on the most proximate sample locations in the 14CHRONO Marine Reservoir Database [45] (see Online Table 1 for details).

To define genetic group labels we in general used the following nomenclature: "*Country_SiteName_AgeBP*" [20]. "*AgeBP*" of a genetic group comprised of more than one individual is calculated by averaging the mean calibrated date in years before present (BP) of the directly dated samples that provided nuclear DNA data. For samples that were not directly dated we considered the averaged value of the corresponding genetic group.

**Ancient DNA sample processing:**
We screened skeletal samples for DNA preservation in dedicated clean rooms at Harvard Medical School in Boston (USA), UCSC Paleogenomics in Santa Cruz (USA), the Max Plank Institute for Science of Human History in Jena (Germany), University of Tübingen in Tübingen (Germany) and the Australian Centre for Ancient DNA in Adelaide (Australia) (Online Table 1). Powder was prepared from the skeletal samples and DNA extraction [11] and library preparation [13] were performed using previously established protocols. Except for samples processed with the single-stranded library protocol (*Brazil_Jabuticabeira2_2000BP*) [12], all other samples were treated with uracil-DNA glycosylase (UDG) to greatly reduce the presence of errors characteristic of ancient DNA at all sites except for the terminal nucleotides [13], or including at the terminal nucleotides (UDGplus) [46]. We enriched for sequences overlapping 1,233,013 SNPs ('1240k SNP capture') [14] and sequenced the DNA on Illumina NextSeq500 or HiSeq 4000 instruments. We removed adapters from the sequences using *SeqPrep* [47] or *AdapterRemoval v2* [48], mapped the data to hg19 with BWA [49], removed duplicates with bamrmdup [50] or Dedup [51] and merged data from different libraries of the same individual using *SeqPrep*. The damage patterns were quantified using mapDamage2.0 [52]. We extracted genotypes from the ancient genomes by drawing a random sequence at each position, ignoring the first and last 2 bp of every read as well as any read containing insertions or deletions in their alignment to the human reference genome. For samples not treated with UDG (UDGminus) (e.g. *USA_Anzick1_11400BP* and *Brazil_Jabuticabeira2_2000BP*), we also clipped 10bp and created a second dataset to represent the sample following this processing. If the randomly drawn haploid genotype of an ancient individual did not match either of the alleles of the biallelic SNP in the reference panel, we set the genotype of the ancient individual as missing. For the great majority of analyses we analyzed all autosomal SNPs from the 1.24 million SNP enrichment reagents. For a subset of analyses we restricted to transversion SNPs which are unaffected by the characteristic ancient DNA errors that occur at transition SNPs.

QUANTIFICATION AND STATISTICAL ANALYSIS

**Contamination estimation in mitochondrial DNA, the X chromosome, and the autosomes:**
We tested for contamination in mtDNA using *schmutzi* (parameters: --notusepredC --uselength) [18], which iteratively determines the endogenous mitochondrial genome while also estimating human mitochondrial contamination given a database of potential contaminant mitochondrial genomes. For males we estimated contamination on the X-chromosome with ANGSD [19], which creates an estimate based on the rate of heterozygosity observed on the X-chromosome. We used the parameters minimum base quality=20, minimum mapping quality=30, bases to clip for damage=2, and set all other parameters to the default. Finally, we measured autosomal contamination using a recently developed tool based on breakdown of linkage disequilibrium

that works for both males and females (Nakatsuka, Harney *et al*., in preparation). We report but do not include in our main analyses samples with evidence of contamination greater than 5% by any of the contamination estimation methods (only sample CP26 was excluded). Due to high contamination levels (the non-damage restricted samples skewed towards West Eurasians on global PCA (not shown), sequences of all *Brazil_Jabuticabeira2_2000BP* samples were filtered with PMDtools [53] to retain only fragments with a typical ancient DNA signature and then trimmed 10bp on either end before analysis. All contamination estimates are reported in Online Table 1.

**Present-day human data:**

We used present-day human data from the Simons Genome Diversity Project [54], which included 26 Native American individuals from 13 groups with high coverage full genome sequencing. We also included data from 48 Native American individuals from 9 different populations genotyped on the Affymetrix Human Origins array [9, 55] as well as 493 Native American individuals genotyped on Illumina arrays either unmasked or masked to remove segments of possible European and African ancestry [7].

**Y chromosome and mitochondrial DNA analyses:**

For Y chromosome haplogroup calling, we used the original BAM files and performed an independent processing procedure. We filtered reads with mapping quality < 30 and bases with base quality < 30, and for UDGhalf treated libraries we trimmed the first and last 2-3bp of each sequence to remove potential damage induced mutations. We determined the most derived mutation for each sample using the tree of the International Society of Genetic Genealogy (ISOGG) and confirmed the presence of upstream mutations consistent with the assigned Y chromosome haplogroup using Yfitter [56]. For mtDNA haplogroup assignment, we used *Haplofind* [57] on the consensus sequences reconstructed with *schmutzi* (parameters: --notusepredC --uselength) [18] after applying a quality filter of ≥ 10 (or ≥ 11 for LapaDoSanto_Burial28, LapaDoSanto_Burial17 and ArroyoSeco2_AS6) for a total of 48 newly reported sequences, including samples for which no nuclear data was obtained (Online Table 1). We produced a multiple genome alignment of our newly reconstructed sequences (excluding *LagunaChica_SC50_L763* because of low coverage) along with 17 previously published ancient sequences older than ~4000 BP (Online Table 1) and 230 present-day sequences [22] using MUSCLE (parameter: -maxiters 2) [58]. We thus analyzed a total of 295 mtDNAs and used an African sequence as outgroup. We used the program MEAGA6 [59] to build a Maximum Parsimony tree with 98% partial deletion (16518 positions) and 500 bootstrap iterations, and visualized it in FigTree (http://tree.bio.ed.ac.uk/software/) (Figure S7).

We used the newly reconstructed mtDNA combined with previously published present-day and ancient sequences (Online Table 1) to generate a maximum parsimony tree (Figures S7A and S7B). This tree recapitulates the star-like phylogeny of the founding Southern Native American mtDNA haplogroups A2, B2, C1b, C1c, C1d, D1 and D4h3a reported previously [60]. We report five new Central and South American individuals belonging to the rare haplogroup D4h3a (3 Brazil, 1 Chile, 1 Belize), which among ancient individuals has been identified so far only in two individuals from the North American Northwest Coast [61] and in the *Anzick-1* individual [6] but not in Southern Ontario, ancient Californians (Scheib et al. 2018), or Western South America [22]

where it has the highest frequency today [31]. Previously this haplogroup was hypothesized to be a possible marker of human dispersal along the Pacific coast, but its presence in early individuals from Belize and Brazil (as well as in the inland *Anzick-1* genome from Montana in the U.S.A.) suggests an ancient spread toward the Atlantic coast as well with its lower frequency there today being due to population replacement or to genetic drift.

The maximum parsimony tree is also striking in showing that the lineage leading to haplogroup D4h3a has a much longer branch than all other Native American-specific mtDNA haplogroups. The diversification of haplogroup D4h3a dates to ~16000 BP which temporally overlaps with the coalescence time of A2, B2, C1, and D1 haplogroups [22]. This suggests that a rate acceleration took place on the lineage leading to the radiation of D4h3a, similar to what has been observed among African L2 lineages [62].

**Principal component analysis:**

We used *smartpca* from EIGENSOFT and default settings [63] to compute principal components using present-day populations. We projected ancient individuals with at least ~10,000 overlapping SNPs using the option *lsqproject:* YES, on eigenvectors computed using the present-day populations genotyped on the Illumina array (we restricted our analysis to the subset of Native Americans without evidence of post-colonial mixture [7]).

**Symmetry statistics and admixture tests (*f*-statistics):**

We computed *D*-statistics, $f_4$-statistics and $f_3$-statistics with ADMIXTOOLS [21] using the programs qp3Pop and qpDstat with default parameters and "f4mode: YES". We computed standard errors with a weighted block jackknife over 5-Mb blocks. For $f_3$-statistics we set the "inbreed: YES" parameter to account for the fact that we are representing the ancient samples by a randomly chosen allele at each position rather than using their full diploid genotype which we do not have enough data to discern. The details of the inbreeding correction, which computes the expected value of statistics taking into account this random sampling, are presented in the section 1.1 of the Appendix of [64]. We computed "outgroup" $f_3$-statistics of the form $f_3(Mbuti; Pop_1, Pop_2)$, which measures the shared genetic drift between population 1 and population 2. Where relevant we plot the statistics on a heatmap using R[65][64][63][79] (https://github.com/pontussk/point_heatmap/blob/master/heatmap_Pontus_colors.R). We also created a matrix of the outgroup $f_3$ values between all pairs of populations. We converted these values to proxies for distances by subtracting the values from 1 and generating a multi dimensional scaling (MDS) plot with a custom-made R script. We converted the original values to distances by taking the inverse of the values and generating a Neighbor joining tree using PHYLIP version 3.696's [66] "neighbor" function and setting USA_USR1_11400BP as the outgroup (default settings were used for the rest of the analysis). We displayed the tree using Itol [67].

***qpWave* analyses:**

To determine the minimum number of streams of ancestry contributing to Central and South American populations, we used the software *qpWave* [7] which assesses whether the set of $f_4$-statistics of the form $f_4$*(A=South American 1, B=South American 2; X=outgroup 1, Y=outgroup 2)*, which is proportional to the product of allele frequencies summed over all SNPs $(p_A-p_B)(p_X-p_Y)$, forms a matrix that is consistent with different ranks (rank 0 would mean consistency with a

single stream of ancestry relative to the outgroups; rank 1 would mean 2 streams of ancestry, and so on). The significance of the statistic is assessed using a Hotelling $T^2$ test that appropriately corrects for the correlation structure of $f_4$-statistics (and thus multiple hypothesis testing). For most analyses, we used ancient California individuals from [8] (*USA_MainlandChumash_1400BP, USA_SanFranciscoBay_300BP, USA_SanNicolas_4900BP, and USA_SanClemente-SantaCatalina_800BP*), *Chipewyan, Russia_MA1_24000BP (MA1), Anzick-1, Han, Papuan, Karelia Hunter Gatherer*, and modern Mexican groups (*Zapotec, Mixtec, Mixe,* and *Mayan*) as outgroups. We also performed the analyses with different outgroups to determine the effect of outgroups on the results (for a detailed list, see Online Table 3). We used all possible pairs, triplets, and quadruplets of South American groups as test populations. We also tried different combinations of South American groups—up to 15 different groups together—as test populations. For *qpWave* analyses we used the default settings except for the change that we set allsnps: YES.

**Admixture graph modeling:**

We used *qpGraph* [21] to model the relationships between diverse samples. This software assesses the fit of admixture graph models to allele frequency correlation patterns as measured by $f_2$, $f_3$, and $f_4$-statistics. We started with a skeleton phylogenetic tree consisting of *Mbuti, Russia_MA1_24000BP (MA1*), *Onge*, and *Han* from prior publications [9, 68]. We added the ancient South American populations in different combinations and retained only the graph solutions that provided no individual $f_4$-statistics with $|Z|>3.5$ between empirical and predicted statistics (except for the case of adding *Surui* due to the difficulties of modeling in the *Population Y* signal). We created the graphs with all overlapping SNPs among the included groups. We used the default settings of *qpGraph* for all runs exept for the options “outpop: NULL” instead of setting an outgroup population and “allsnps: YES” to compute each *f*-statistic on the common SNPs present in the populations involved in the statistic, rather than the intersection of all SNPs present in the dataset. To reduce the impact of damage-induced substitutions in UDGminus data of the *Anzick-1* individual we restricted the analysis to a version of this sample where sequences were 10bp trimmed on both sides before genotyping. In addition, we performed all analyses with the transitions at CpG sites removed, and we also report the maximum Z-scores of many of the analyses with all transition sites removed. Lastly, for the graphs in Figures S6A-D we computed standard errors for the lengths of different graph edges by performing a block jackknife by dropping each of 100 contiguous blocks (with an equal number of SNPs) in turn [69].

Scheib et al. analyzed data from diverse Native American populations—ancient and modern—and proposed that in Central and South Americans today there is a history of widespread admixture between the two deepest branches of Native American genetic variation (*ANC-A* and *ANC-B*), with a minimum of ~30% of each branch admixed into all populations [8]. They write "The summary of evidence presented here allows us to reject models of a panmictic “first wave” population from which the ASO [the Ancient South Ontario population] diverged after the population of South America or in which solely the ANC-A population contributed to modern southern branch populations.”

The evidence for the claim that Central and South Americans do not have entirely *ANC-A* ancestry is based on fitting the admixture graph model of Figure 2A in Scheib et al. 2018, which the authors show is a fit to the data jointly for *Han, Anzick-1, USA_SanNicolas_4900BP (ESN), USA_SanNicolas_1400BP (LSN), Pima, Surui,* and *Canada_Lucier_4800BP-500BP (ASO)*. They then

added a diverse set of other Native American populations into the graph as mixtures of the same two lineages, and report the mixture proportions in Table S8 of their study.

We began by replicating the finding of Scheib et al. 2018 that their proposed admixture graph was a fit to the data (maximum mismatch between observed and expected *f*-statistics of |Z|=1.1) (Figure S6A). However, when we added to the admixture graph additional non-American populations whose phylogenetic relationship to Native American populations has been well worked out (*Russia_MA1_24000BP*, *Onge*, and *Mbuti*), the model is a poor fit (maximum mismatch of observed and expected *f*-statistics of |Z|=4.8) (Figure S6B). This implies that the model of Scheib et al. 2018 does not capture some important features of the history relating these populations, and suggests that we may not be able to rely on the inferred proportions of ancestry.

If Scheib et al. 2018 were correct that there was widespread *ANC-B* ancestry in Central and South America, then *Canada_Lucier_4800BP-500BP* would not be an outgroup to *Anzick-1* and all Central and South Americans; that is, statistics of the form $f_4$*(USR1, Canada_Lucier_4800BP-500BP; Anzick-1, Test Central or South America)* would often be positive. In fact, *Canada_Lucier_4800BP-500BP* is consistent with being an outgroup to all Central and South America in our analysis, as statistics of the form $f_4$*(USR1, Canada_Lucier_4800BP-500BP; Anzick-1, Test Central or South America)* are all consistent with zero except for the special *Late Central Andes* individuals (as we describe elsewhere, this signal could be explained either by less than 2% *Canada_Lucier_4800BP-500BP* admixture into *Late Central Andes* groups, or alternatively *USA_SanNicolas_1400BP*-related admixture into *Canada_Lucier_4800BP-500BP*) (modern South Americans such as *Piapoco* and *Quechua* had statistics consistent with zero as well) (Online Table 2). This is in line with Figure S13 of Scheib et al. 2018, where *Canada_Lucier_4800BP-500BP* is also fit as an outgroup to Central and South Americans; the fit of Figure S13 of their study is reasonable, with the maximum mismatch between observed and expected *f*-statistics being |Z|=2.0, which is not surprising after correcting for the number of hypotheses tested.

To obtain some insight into why models such as Figure 2A of [8] could fit the data even while statistics like $f_4$*(USR1, Canada_Lucier_4800BP-500BP; Anzick-1, Test Central or South America)* are for the most part consistent with being zero, we estimated the genetic drift along the edge leading to *Canada_Lucier_4800BP-500BP* that mixed into South Americans in Figure 2A. We found that it is not significantly different from zero in any of the graphs that we analyzed (Figures S6A-D, Methods), meaning the ancestry on the *Canada_Lucier_4800BP-500BP* branch that mixes into the South American groups does not share a significant amount of genetic drift with *Canada_Lucier_4800BP-500BP* and there is no need to propose widespread mixing between *ANC-A* and *ANC-B*.

A supporting piece of evidence cited by Scheib et al. 2018 in favor of mixture between *ANC-A* and *ANC-B* lineages in Central and South Americans is that they identify present-day *Pima* and *Surui* haplotypes that match *Anzick-1* haplotypes (as a representative of *ANC-A*) more closely than *CK-13* (as a representative of *Canada_Lucier_4800BP-500BP*), and vice versa. However, Native American populations (like all human populations) have a large proportion of shared ancestral haplotypes, and incomplete lineage sorting means that even if two populations are not most closely related, in some sections of the genome they will be most closely related on a

haplotypic basis. Thus, it is not clear to us that this analysis demonstrates that *Pima* and *Surui* derive from *ANC-A/ANC-B* mixtures

In conclusion, given that *Canada_Lucier_4800BP-500BP* is consistent with being an outgroup to nearly all Central and South Americans based on *f*-statistic analysis (with the exception of the special *Late Central Andes* populations), and that there is no compelling haplotype-based evidence for *ANC-A* and *ANC-B* admixture in the history of Central and South Americans, the genetic data are in fact consistent with the scenario in which an *ANC-A* population was the sole contributor to southern branch (Central and South American populations). Thus, our results are consistent with the originally suggested null hypothesis of entirely *ANC-A* ancestry leading to Central and South Americans [5-7].

To build the admixture graph shown in Figure 3, we used a skeleton graph from previous publications [9, 68]. We added in groups based on previous findings (e.g. the Ancient Beringian *USR1* as an outgroup [4] and the split between *ANC-A* and *ANC-B* [7, 8]). We then added additional groups new to this study using guidance from other results such as the outgroup-$f_3$ matrix-based neighbor-joining tree. We stopped building the admixture graph once we had fit as many representative ancient individuals as possible that could fit without strong evidence of mixture (worst Z-score outlier $f_4$ (*Han*, *USA_SanNicolas_4900BP; Argentina_ArroyoSeco2_7700BP, Canada_Lucier_4800BP-500BP)* Z=2.9).

To build the complex admixture graphs shown in Figures 4 and 5, we used two approaches. For Figure 4, we started with the admixture graph of Figure 3, and then grafted onto it admixture events motivated by our *qpWave* results, namely mixture from an *Anzick-1*-related lineage into the earliest Chilean individual and some of the Brazil and Argentina groups, and mixture of *USA_SanNicolas_4900BP*-related ancestry into *Late Central Andes* groups. We compared models with and without admixture edges and used the model with an extra admixture edge if it decreased the maximum Z score by over 0.3.

For Figure 5 we carried out a semi-automated search in which we began with a skeleton model including all non-Native Americans and *USA_USR1_11400BP*, and then iteratively added as many other populations as we could in a greedy approach, first as simple clades in order to minimize graph complexity, and then as 2-way mixtures if the sample clade approach did not fit. Thus, for *N* populations, we first fit graphs of *m* populations and then considered all remaining *N-m* populations as candidates to be grafted in all fitting models with *m* populations. Each grafted population was either placed anywhere on the graph (or its two components in case of mixture were placed anywhere on the graph). This approach is described in more detail in [26].

The two admixture graphs shown in Figures 4 and 5 have many qualitative points of agreement including: i) *USA_USR1_11400BP* as an outgroup to all other Native Americans, ii) a split of *ANC-A* and *ANC-B* such that *ANC-B* had minimal genetic influence on all South Americans, iii) A rapid radiation of the earliest South Americans, with the earliest South Americans having very little drift on the lineages separating them, iv) distinctive shared ancestry between *Brazil_LapaDoSanto_9600BP* and *Chile_LosRieles_10900BP* on the one hand and *USA_Anzick-1_12800BP* on the other, v) distinctive shared ancestry between *USA_Anzick-1_12800BP* and *USA_SanNicolas_4900BP*, and vi) mixture of a source of ancestry with distinct relatedness to North Americans into *Late Central Andes* groups.

The primary disagreement between the admixture graphs concerns the question of whether or not *USA_Anzick_12800BP* is admixed.

In Figure 4 *USA_Anzick_12800BP* is modeled as unadmixed, and ancestry related to this group mixes into some of the Brazil, Chile, and Argentina groups as well as into *USA_SanNicolas_4900BP*. The ancestry sources can be interpreted as resulting from North to South America migrations in successive streams. There are an initial two streams from an *Anzick-1*-related group retained in *Chile_LosRieles_10900BP*, *Brazil_LapaDoSanto_9600BP*, and *Argentina_ArroyoSeco2_7700BP* and another ancestry stream that is pervasive throughout ancient South America (we cannot resolve the order of these two streams). There is a third ancestry source contributing to *Late Central Andes* groups, and a fourth ancestry source that corresponds to the *Population Y* signal in *Karitiana* and *Surui* but that we do not specifically model in the graph.

In Figure 5, most South Americans can be modeled as a mixture of a lineage that split into regional branches in Peru (*Lauricocha_8600BP* and *Cuncaicha_9000BP*), the Southern Cone (*Argentina_ArroyoSeco2_7700BP* and *Chile_LosRieles_5100BP*), and *Brazil_LapaDoSanto_9600BP*, with the lineage more closely related to *Brazil_LapaDoSanto_9600BP* then mixing into the shared ancestors of *USA_Anzick_12800BP* and *USA_SanNicolas_4900BP* (possibly reflecting a back-flow from South to North America, although, alternatively, all the splits could have occurred in North America). The model also specifies more recent admixture into *Late Central Andes* population of a lineage with a distinctive relatedness to North Americans (this model also included West Eurasian related admixture in *Canada_Lucier_4800BP-500BP* that likely reflects a low level of contamination in these samples).

Both models shown in Figures 4 and 5 are reasonable statistical fits (maximum Z-scores of 3.4 and 2.9 with only transitions in CpG sites removed, and 3.0 and 2.9 when all transitions are removed), and we were unable to resolve which was better. Additional sampling of early North and South Americans could help to resolve the true model.

In Figure S5, we present various modifications of these models, including some that add *Surui* which has evidence of a fourth source of "*Population Y*" ancestry that bears a different relationship to Asians.

**Analyses of phenotypically relevant SNPs:**

We analyzed sequences at SNPs previously known to be relevant to interesting phenotypic traits (Supplemental Note 1). We used *samtools* version 1.3.1 [70, 71] *mpileup* with the settings -d 8000 -B -q30 -Q30 to obtain information about each read from the bam files of our samples. We used the fasta file from human genome GRCh37 (hg19) for the pileup. We counted the number of derived and ancestral variants at each analyzed position using a custom Python script.

Besides *EDAR* we analyzed several other phenotypically relevant variants including one variant *TBX15* which affects body fat distribution [72], a variant in *SLC16A11*, which predisposes individuals to diabetes [73], 2 variants in *NOS3* and *EGLN1* believed to facilitate life at high altitudes [74], top 10 variants from a recent study on natural selection in Andeans [75], and a variant at the fatty acid desaturase gene *FADS2* with evidence of natural selection [34].

We observed that the *SLC16A11* variant rs13342232 is homozygous ancestral (A) in *USR1* and *Anzick-1*, but the derived allele is present in 17/32 of the individuals that had coverage at the variant, which is approximately the frequency observed in present day Native Americans (we observe no significant correlation with time or location).

The *TBX15* allele is heterozygous in *USR1* but homozygous derived (A) in *Anzick-1* and all other individuals with at least one sequence covering the SNP (39) except 2 *Belize_SakiTzul_7400BP* and a present-day *Mayan* individual that were heterozygous, and 1 *Argentina-Chile_FuegoPatagonian_100BP* that had a single sequence supporting the ancestral position. This reflects the approximate allele frequency present in present day Native Americans [72], meaning that selection, if it occurred, likely did so prior to the diversification of Native Americans.

The *NOS3* and *EGLN1* derived and ancestral alleles were present in appreciable frequencies in all time periods and locations, and we lacked enough samples to assess whether a prior report of selection on these variants [74] was consistent with our data.

As expected, all individuals were homozygous for the ancestral allele at *SLC45A2* (C) and *LCT* (G), and *SLC24A5* (G), indicating darker skin color and lack of lactase persistence.

In the Greenland Inuit *FADS2* has been shown to have experienced selection related to cold adaptation and to a diet rich of proteins [76]. Selection scans in non-Arctic Native American groups who share a substantial proportion of their ancestry with the Inuit also identified the *FADS2* locus as being under positive selection, and it has been proposed that the adaptation took place in a common ancestral group before their entrance in the Americas [34]. All our ancient individuals harbor the derived variant of a *FADS2* SNP (chr11:pos61597212, rs174570) supporting the view that the selection could have taken place in the ancestral population of Native Americans (Online Table 4).

**Insights into more recent history of Brazil based on Jabuticabeira 2 individuals:**

Maritime societies are documented on the coast of southern and southeastern Brazil since about 8000 BP. Even without agriculture and pottery, these groups achieved impressive demographic densities. Their most outstanding cultural practice was the building of hundreds of shell-mounds, some of monumental magnitude, that were used as a funerary ground (up to one thousand skeletons are estimated to be included in mounds that could be 50 meters high and 200 meters in diameter). Analysis of sex-biased morphological variation suggest these groups were matrilocal [25]. It is still not clear whether shell-mound builders constituted a pan-regional society with a single origin sharing ancestry and language, or if they were groups of independent origins who adopted a similar subsistence strategy focused on a maritime economy and the construction of shell mounds. Nevertheless, around 2000-1000 BP there is a clear decline and eventually cessation of shell-mound construction. A highly debated question in South American pre-colonial history concerns the nature of the disappearance of the Sambaqui archaeological tradition and the role in this process of two main population movements in the region – arrival of Ge speakers and arrival of Tupi speakers, hypothesized to have occurred at ~2000 BP and ~1000 BP, respectively.

While it has been proposed that Sambaqui shell mound builders and proto-Ge speakers probably interacted, the arrival of Tupi-Guarani speakers (such as present-day *Parakana* and *Guarani*) is hypothesized to have involved a rather abrupt population replacement leading to the complete disappearance of the Sambaqui culture – represented in the current study by Jabuticabeira 2 [24, 25].

Our analyses of the genetic affinities of *the Brazil_Jabuticabeira2_2000BP* individuals to modern groups provides the first genetic evidence to test this model. In Figure S3 it appears that

the *Kaingang* (a Ge speaking group from southern Brazil) and the *Arara* (a Carib speaking group part of the Ge-Pano-Carib family from northern Brazil) have a greater genetic affinity with *Brazil_Jabuticabeira2_2000BP* than do groups that speak Tupi-Guarani languages. This pattern is confirmed with statistics of the form $f_4$*(Mbuti, Brazil_Jabuticabeira2_2000BP; Guarani, Kaingang)* and $f_4$*(Mbuti, Brazil_Jabuticabeira2_2000BP; Parakana, Arara)* that are significantly positive (Z scores of 2.7 and 3.2, respectively, despite fewer than 50,000 SNPs being available for analysis) (Table S1), suggesting that present-day Ge speakers across a wide geographic region harbor specific affinities to Sambaqui shell mound builders, consistent with some elements of shared ancestry in Ge speakers. This pattern holds even when we remove the most recent Jabuticabeira 2 individual dated to ~1200BP (Table S1).

Within Jabuticabeira 2, there are two distinct periods of occupation. The earliest is dated to 2500-2200 BP and can be considered a late expression of the classic Sambaqui phenomenon that first appears in the region around 8000 BP. The later occupation event is dated to around 1500 BP and archaeologically is expressed by the formation of layers of dark earth on top of the shell-mounds. This transformation is not restricted to Jabuticabeira 2 as it has been documented at several other locations along the Atlantic coast. Possible factors to explain this transformation range from environmental changes (e.g. total depletion of mollusk banks) to the arrival of new people or simply the implementation of new practices. In the present study the only individual of Jabuticabeira 2 coming from the second event of occupation is Burial 102 - all others are from the classic Sambaqui phase of occupation. Our genome-wide data is thin for this individual and lack the resolution to test for differences in ancestry between the groups. However, mtDNA of all seven individuals from the classic Sambaqui horizon share the same haplogroup C1c indicating low mtDNA diversity, while Burial 102 carries mtDNA haplogroup B2. More individuals from the later period of occupation of the site should be able to reveal if Burial 102 is representative of a population shift.

**Data Availability**

Raw sequences (bam files) from the 49 newly reported ancient individual with genome-wide data and 48 newly reported individuals with mtDNA data are available from the European Nucleotide Archive. The accession number for the sequence data reported in this paper is ENA: (PRJEB28961).

# Chapter 5.2: A Paleogenomic Reconstruction of the Deep Population History of the Andes

Nathan Nakatsuka[1,2], Iosif Lazaridis[1], Chiara Barbieri[3,4], Pontus Skoglund[5], Nadin Rohland[1], Swapan Mallick[1,6,7], Cosimo Posth[3], Kelly Harkins-Kinkaid[8], Matthew Ferry[1,6], Éadaoin Harney[1,6], Megan Michel[1,6], Kristin Stewardson[1,6], Jannine Novak-Forst[8], José M. Capriles[9], Marta Alfonso Durruty[10], Karina Aranda Álvarez[11] , David Beresford-Jones[12], Richard Burger[13], Lauren Cadwallader[14], Ricardo Fujita[15], Johny Isla[16], George Lau[17], Carlos Lémuz Aguirre[18], Steven LeBlanc[19], Sergio Calla Maldonado[18], Frank Meddens[20], Pablo G. Messineo[21], Brendan J. Culleton[22], Thomas K. Harper[23], Jeffrey Quilter[19], Gustavo Politis[21], Kurt Rademaker[24], Markus Reindel[25], Mario Rivera[26,27] Lucy Salazar[12], José R. Sandoval[15], Calogero M. Santoro[28], Nahuel Scheifler[21], Vivien Standen[29], Maria Ines Barreto[30], Isabel Flores Espinoza[30], Elsa Tomasto-Cagigao[31], Guido Valverde[32], Douglas J. Kennett[22,23,33], Alan Cooper[32], Johannes Krause[3], Wolfgang Haak[3], Bastien Llamas[32], David Reich[1,6,7,34,36], Lars Fehren-Schmitz[8,35,36]

[1] Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
[2] Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02115, USA.
[3] Max Planck Institute for the Science of Human History, Jena 07745, Germany.
[4] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich 8057, Switzerland.
[5] Francis Crick Institute, London NW1 1AT, UK.
[6] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02446, USA.
[7] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA.
[8] UCSC Paleogenomics, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.
[9] Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA.
[10] Department of Sociology, Anthropology and Social Work, Kansas State University, Manhattan, KS 66506, USA.
[11] Sociedad de Arqueología de La Paz, 5294 La Paz, Bolivia.
[12] McDonald Institute for Archaeological Research, University of Cambridge, Downing St., Cambridge, CB2 3ER, UK.
[13] Department of Anthropology, Yale University, New Haven, CT 06511, USA.
[14] Office of Scholarly Communication, Cambridge University Library, Cambridge CB3 9DR, UK.
[15] Centro de Genética y Biología Molecular, Facultdad de Medicina, Universidad de San Martín de Porres, Lima 15011, Peru.
[16] Peruvian Ministry of Culture, DDC Ica, Directos of the Nasca-Palpa Management Plan, Calle Juan Matta 880, Nasca 11401, Peru
[17] Sainsbury Research Unit, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK.
[18] Carrera de Arqueología, Universidad Mayor de San Andrés, Edificio Facultad de Ciencias Sociales 3er Piso, La Paz 1995, Bolivia.
[19] Harvard Peabody Museum, Harvard University, Cambridge, MA 02138, USA.
[20] School of Archaeology, Geography and Environmental Sciences, University of Reading, Reading, Berkshire, RG6 6AH, UK.
[21] INCUAPA-CONICET, Facultad de Ciencias Sociales, Universidad Nacional del Centro de la Provincia de Buenos Aires, Olavarría 7400, Argentina.
[22] Institutes for Energy and the Environment, The Pennsylvania State University, University Park, PA 16802, USA.
[23] Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA.
[24] Department of Anthropology, Michigan State University, East Lansing, MI 48824, USA
[25] Commission for Archaeology of Non-European Cultures, German Archaeological Institute, Berlin 14195, Germany.

[26] Universidad de Magallanes, Punta Arenas 6210427, Chile.
[27] Field Museum Natural History 1400 S Lake Shore Dr., Chicago, IL 60605, USA.
[28] Instituto de Alta Investigation, Universidad de Tarapaca, Antafogasta 1520, Arica, 1000000, Chile.
[29] Departamento de Antropología, Universidad de Tarapacá, Antafogasta 1520, Arica, 1000000, Chile.
[30] Museo de Sitio Huaca Pucllana, Calle General Borgoño, Cuadra 8, Miraflores, Lima 18, Peru.
[31] Department of Humanities, Pontifical Catholic University of Peru, San Miguel 15088, Peru.
[32] Australian Centre for Ancient DNA, School of Biological Sciences and The Environment Institute, Adelaide University, Adelaide, SA 5005, Australia.
[33] Current address: Department of Anthropology, University of California, Santa Barbara, Santa Barbara, CA 93106, USA.
[34] Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[35] UCSC Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA 95064, USA.
[36] Senior author

**Keywords:**
Andes, population genetics, archaeology, anthropology, ancient DNA

**Correspondence:**
N.N. (nathan_nakatsuka@hms.harvard.edu); D.R. (reich@genetics.med.harvard.edu); L.F-S. (lfehrens@ucsc.edu)

**Lead Contact:** L.F-S. (lfehrens@ucsc.edu)

**Supplementary Material:** All supplementary material can be found in the supplement of Nakatsuka, *et al*., 2020 *Cell*.

# Abstract

There are many unanswered questions about the population history of the Central and South Central Andes, particularly regarding the impact of large-scale societies, such as the Moche, Wari, Tiwanaku, and Inca. We assembled genome-wide data on 89 individuals dating from ~9000-500 years ago (BP), with a particular focus on the period of the rise and fall of state societies. Today's genetic structure began to develop by 5800 BP, followed by bi-directional gene flow between the North and South Highlands, and between the Highlands and Coast. We detect minimal admixture among neighboring groups between ~2000-500 BP, although we do detect cosmopolitanism (people of diverse ancestries living side-by-side) in the heartlands of the Tiwanaku and Inca polities. We also reveal cases of long-range mobility connecting the Andes to Argentina, and the Northwest Andes to the Amazon Basin.


# Introduction

The South American Andean regions have a long and dynamic history beginning with the arrival of the first hunter-gatherers at least ~14500 BP. In the Central and South-Central Andean regions (present-day Peru, Bolivia, and North Chile), early settlements in both the Coast and the Highlands [1-5] were followed by the development of sedentary lifestyles, complex societies, and eventually archaeological cultures with wide spheres of influence, such as the Wari (~1400-950 BP), Tiwanaku (~1400-950 BP), and Inca (~510-420 BP) (BP: before present, defined as years before 1950 CE; in what follows, all radiocarbon dates are corrected with appropriate calibration curves as justified in STAR Methods and summarized by the midpoint of their estimated date ranges rounded to the closest century; Table S1).

Archaeological research in the Central Andes is extraordinarily rich [6], but ancient DNA studies to date have been limited, so there has been little information about demographic change over time. Studies of uniparental DNA indicated evidence for a degree of genetic homogeneity of the Central and Southern Highlands, especially for the Y chromosome [7-11], while studies with aDNA suggested substantial continuity as well as gene flow between the Coast and the Highlands [12-17]. High coverage genome-wide ancient DNA data from South America from the time before European contact began to be published in 2018, with most data from mid- to early-Holocene hunter-gatherers (in the Central and South-Central Andes 23 individuals were reported) [18-20]. Although these studies had large geographic and temporal gaps, they were critical in showing that individuals from the Central and South-Central Andes up to at least ~9000 BP are more closely related to modern Andean highland, rather than coastal or Amazonian populations [18, 19]. An additional lineage was found to have begun spreading in this region by at least ~4200 BP [18] and had a significant genetic affinity (excess allele sharing) with groups in Mexico and the California Channel Islands.

We assembled genome-wide data from 89 individuals from the Central and South-Central Andes over the past ~9000 years, including 65 newly reported individuals, and added data from a ~1600 BP individual from the Argentine Pampas region (Figure 1A-B, Figure S1, and Table S1). We also report 39 direct radiocarbon dates (Table S1). The dataset includes individuals associated with a wide range of

archaeological cultures from the Highlands and Coast of three geographic regions within present-day Peru: a northern zone we call “North Peru” (including sites in the Departments La Libertad and Ancash), a central zone we call “Central Peru" (Department of Lima), and a southern zone we call "South Peru” (including sites in the Departments of Ica, Ayacucho, Arequipa, Apurimac) spanning thousands of years through each of the regions. We also assembled data from Cusco in Peru, the South-Central Andean "Titicaca Basin" Highlands (spanning an ecologically and culturally unique region of southernmost Peru as well as western Bolivia), and "North Chile" (see Figure 1). Here we use the term “archaeological cultures” as a proxy for the particular material cultures and site contexts from which our ancient individuals are derived, acknowledging that the actual human societies that produced the artifacts representing these material cultures often had substantially different social organizations, and our data is not sufficient to capture the full breadth and internal dynamics of each of them.

We combined the new data with previously published ancient DNA data from [18-26] and compared it with the genetic diversity of different present-day peoples [27-29]. We determined when the genetic structure observed today in the Central Andes first began to develop and assessed the degree to which gene flows over time have modulated this structure. Further, we investigated how changes in the population structure might correlate to archaeologically documented episodes of cultural, political, and socioeconomic change (summary of findings in Data S1).

## Ethics and Community Engagement

We acknowledge the Indigenous Peoples of Peru, Bolivia, Chile, and Argentina who supported this study as well as the ancient and present-day individuals whose samples we analyzed. The analysis of DNA from ancient individuals can have significant implications for present-day communities both because the studies can reveal how ancient people relate to present-day groups and also because the physical handling of the skeletal materials might be sensitive to the groups involved. Thus, it is important to engage with local communities and with scholars who work closely with these communities to incorporate these perspectives [30, 31], and to do so in a way appropriate for the particular Indigenous communities and political and social history in each region.

This study is the result of an international and inter-institutional collaborative effort that includes scientists from the countries where the ancient individuals originated. In all cases the interest in genetic investigation of human remains centrally involved local co-authors, in most cases the archaeologists that excavated the sites. In many of these countries, archaeological investigations, as well as the permission to conduct biomolecular research on archaeological skeletal remains, is governed by national regulations. In Peru, for example, this is addressed in Ley General del Patrimonio Cultural de la Nación (Law No. 28296) (see also [32, 33]). Our primary approach was thereby by necessity to consult with the provincial and state-based offices of the responsible institutions to obtain permission for analysis. In addition to this, however, we engaged with local communities throughout the study as detailed below.

All but one of the sample sets presented here were exported from their country of origin for this analysis and studied with direct permission of the local government. For example, the great majority of the samples newly reported in this study come from Peru, where this study was approved by the Ministry of Culture of Peru, which was originally created to revalue indigenous culture, past and present, to promote interculturality, and to fight against racism. The only exception is the San Sebastian samples (Cusco, Peru) that were part of a US collection and were studied there as part of a repatriation effort with permission of Peruvian institutions, and are now curated in Cusco. Some of the samples, especially from coastal Peru, come from looted cemetery contexts, and the genomic data and direct radiocarbon dates generated here help to confirm their assignation to cultural epochs. Thus, this work helps to re-contextualize the individuals and has the potential to provide local communities with new ways to engage the past at disturbed sites.

For the individual from Argentina, in addition to obtaining permits from the provincial heritage institutions, the Indigenous community living near the site (Comunidad Indígena Mapuche-Tehuelche Cacique Pincen) approved the study after consultation and participation in the rescue excavation (the skeletal remains will be re-buried). The results of this and prior studies and their implications have been discussed with the community, and they have indicated support for this research in discussion with co-authors of the study. The regulations in Bolivia require archaeologists to consult with local communities before field research and turn in their research field reports to these communities. For the individuals from Chile, we obtained permits from the local heritage institutions, but no local Indigenous community lived near the site or indicated a connection to the analyzed skeletons.

Both before and during this study, there was substantial engagement with local communities by co-authors JS, RF, CB, and GP, who have a long-term presence in each region and years of experience collecting data and returning results to the communities. Several of the co-authors presented the outcomes of this study and related archaeological and paleogenetic studies in the form of publicly accessible talks.

Data S1 is a translation of the Summary and Key Findings sections into Spanish to increase accessibility for non-English speakers, following the precedent established and used by the journal *Latin American Antiquity*.

# Results and Discussion

***Authenticity of Ancient DNA and Single Locus Patterns***

We evaluated the authenticity of the data based on: 1) characteristic cytosine-to-thymine substitution rate at the ends of the sequenced fragments over 3% [34]; 2) point estimates of contamination in mitochondrial DNA (mtDNA) below 5% [35, 36]; 3) point estimates of X chromosome contamination below 3% (only possible in males) [37]; and 4) point estimates of genome-wide contamination below 5% based on a method that leverages breakdown in linkage disequilibrium due to contamination [38]. Individuals I1400, I01, and MIS6 were removed based on these analyses (full metrics are in Table S1). All common South American mtDNA haplogroups A2, B2a, B2b, C1b, C1c, D1, and D4h3a were represented (Figure S2), likely reflecting persistently large population sizes in the Central Andes [9, 17].

***Population Structure Has Early Holocene Roots***

Restricting to autosomal data, we performed a qualitative assessment of the population structure using unsupervised ADMIXTURE (Figure S3) and PCA (Figure 2 and Figure S4). We also generated a neighbor-joining tree and multi-dimensional scaling (MDS) plots of the matrix of "outgroup-$f_3$" statistics of the form $f_3$*(Mbuti; Pop1, Pop2)*, which measure shared genetic drift between population pairs (Figure 3, Figure S5). Genetic structure strongly correlates with geography since at least ~2000 BP, with the first eigenvector in PCA corresponding to a north-to-south cline and the second separating Northwest Amazon groups from Central and South-Central Andean groups (Figure 2). The genetic structure is consistent with patterns expected from isolation-by-distance or gene flow among neighbors, with geographically closer individuals sharing more alleles than people separated by long distances (Figure 3). The oldest individuals (*Peru_NorthHighlands_Lauricocha_8600BP* and *Peru_SouthHighlands_Cuncaicha_9000BP*) did not plot in a position corresponding to their location, as expected since these individuals were not affected by the shared drift and gene flow among geographic neighbors that shaped the population structure of much more recent individuals (Figure S4).

We examined how genetic structure evolved over time using statistics of the form $f_4$*(Mbuti, Test; Pop1, Pop2)* where *Pop1* and *Pop2* were groups of similar time period, iterating over all other populations in our dataset as *Test*. We created analysis clusters for the regions labeled in Figure 1 (*NorthPeruHighlands*, *NorthPeruCoast*, *CentralPeruCoast*, *SouthPeruHighlands*, *SouthPeruCoast*, and *NorthChile)* where all ancient individuals younger than ~2000 BP were grouped (the Cusco individuals were excluded from *SouthPeruHighlands* for reasons discussed below) based on our empirical finding of a high degree of genetic homogeneity in each region since ~2000BP (see below).

We first computed statistics of the form $f_4$*(Mbuti, Test; Coast, Highlands),* declaring significance if the statistics were more than 3 standard errors from zero*. Test* individuals that share significantly more alleles with either Coast or Highlands groups must have lived when population structure existed that distinguished the Highlands regions of the Central and South-Central Andes from other parts of South America, and thus provide a minimum on the date of the structure.

The North Peru Highlands reveal substantial continuity over seven millennia as shown by excess allele sharing of *Peru_NorthHighlands_Lauricocha_8600BP* to *NorthPeruHighlands* relative to *NorthPeruCoast* (Table S2A). Similarly, in the South Peru Highlands, *Peru_SouthHighlands_Cuncaicha_4200BP* shares more alleles with *SouthPeruHighlands* than

with *SouthPeruCoast*. Thus, the oldest Highlands individuals were from populations that contributed more to later Highlands than to Coast groups, suggesting that the distinctive ancestry of late Highlands groups was already beginning to be established in the Highlands many thousands of years before. The long-standing genetic distinctiveness of the Highlands and Coast peoples is consistent with archaeological evidence that inhabitants of the Coast and the Highlands often relied on different subsistence strategies and had very different mobility patterns for millennia [39-41]. A better understanding of the distinctive ancient lineages that we detect in the Coast groups will require older genomes from the Coast.

A minimum date of development of North-South substructure can be inferred from the age of the earliest pair of North vs. South groups that show asymmetric relationships with later individuals from the North and South. The earliest Peruvian Highlands individuals were symmetrically related to post-2000 BP individuals (except for a degree of local continuity over ~5,000 years in the Lauricocha site), but structure is evident by ~4200 BP, because *NorthPeruCoast*, *CentralPeruCoast*, and *NorthPeruHighlands* had significant affinity for *Peru_NorthHighlands_Lauricocha_5800BP* or *Peru_NorthHighlands_LaGalgada_4100BP* relative to *Peru_SouthHighlands_Cuncaicha_4200BP*, which instead had significant affinity for *NorthChile* and Titicaca Basin groups (Table S2B). The northern and southern lineages must have split at least ~5800 BP (the date of *Peru_NorthHighlands_Lauricocha_5800BP*), although we can only be confident that a north/south structure that correlates with the post-2000 BP structure was established by the date of *Peru_SouthHighlands_Cuncaicha_4200BP*. This roughly corresponds to the onset of the Late Preceramic Period (~5000 BP), when increasing economic, political, and religious differentiation between Central Andean regions becomes evident archaeologically, and when levels of mobility decreased at the Coast and slightly later in the Highlands and Altiplano [40, 42, 43]. This occurred in tandem with increasing reliance on plant cultivation [42, 44-48], which has been hypothesized to have contributed to rapid population growth in some regions [8, 49, 50]. A greater reliance on plant cultivation documented in the archaeological record from this period could plausibly contribute to increased sedentism and reduced gene flow, potentially contributing to the North-South substructure we observe beginning to develop by this period. However, it is important to note that demographic changes in the Andes most likely had various tempos and sequences in different regions; thus, data from larger samples sizes from this region around this time is necessary to gain greater clarity on this development of substructure.

***Gene Flow After the Establishment of Population Structure***

We document gene flow between the North and South Peru Highlands after the establishment of initial population structure through significantly more allele sharing of *SouthPeruHighlands* with *NorthPeruHighlands* than with *Peru_NorthHighlands_Lauricocha_5800BP* (Table S3A). We fit an admixture graph [51] (Figure S6A) by systematically searching through all graphs with three or fewer admixture events among ancient Native Americans. *SouthPeruHighlands* could only fit as a mixture between groups related to *Peru_SouthHighlands_Cuncaicha_4200BP* and *Peru_NorthHighlands_Lauricocha_5800BP*. This could reflect gene flow between the regions and/or a mixture from a third unsampled population that affected both regions. We could not determine the directionality of gene flow due to a lack of very ancient South Peru Highlands individuals (Cuncaicha is further south than our later South Peru Highlands series and has ancestry more consistent with later Titicaca Basin individuals). A speculative possibility is that this admixture relates to the archaeologically documented Chavin sphere of influence [52] that

involved cultural interaction between the North Peru Highlands (Ancash) to at least the Ayacucho region ("*SouthPeruHighlands*" in this study) ~2900-2350 BP as reflected in the exchange of goods like cinnabar and obsidian, and by a widespread shared material culture style manifest across the Central Andes between Jaen in the north and Ayacucho in the south and along the north-central Pacific coast [52-54]. This scenario does not imply that the gene flow must have originated from Chavin, but that increased cultural and material exchange between the regions was accompanied by gene flow in one or both directions, though future work is necessary to test this hypothesis.

We also document gene flow between the Highlands and the Coast in North Peru based on significantly more allele sharing of *NorthPeruHighlands* with *NorthPeruCoast* than to *Peru_NorthHighlands_Lauricocha_5800BP* and of *CentralPeruCoast* with *Peru_NorthHighlands_Lauricocha_5800BP* relative to *Peru_NorthHighlands_Lauricocha_8600BP* [16] (Table S4, Table S3A). We detect gene flow connecting the Titicaca Basin to the South Peru Highlands and North Chile prior to ~2000 BP through significant allele sharing of *SouthPeruHighlands* and *NorthChile* with *Peru_TiticacaBasin_RioUncallane_1700BP* relative to *Peru_TiticacaBasin_SoroMikayaPatjxa_6800BP* and *Peru_TiticacaBasin_RioUncallane_1700BP* and *NorthChile* with *Peru_Cuncaicha_4200BP* relative to *Peru_Cuncaicha_9000BP*. This accords with archaeological evidence of cultural exchange prior to ~2000 BP between these regions [55, 56] as well as observations of gene flow between the regions based on mtDNA, though our date estimates precede the estimated dates from the mtDNA studies by ~1000 years [57-60].

***Continuity in Most Regions After ~2000 BP***

After ~2000 BP, we observe genetic homogeneity within most regions to the limits of our statistical resolution. This is evident when we group individuals by geography, time period, and archaeological cultural context and compute statistics of the form $f_4$*(Mbuti, Test; Pop1, Pop2)*, where Pop1 and Pop2 are two groups within the same geographical/temporal/archaeological category, and *Test* is a range of other groups outside the region. Statistics in almost all regions were consistent with zero (Table S3B), indicating that the *Test* population shares alleles at about an equal rate with *Pop1* or *Pop2*. We also used *qpWave* [29] to agglomerate the $f_4$-statistics for each (Pop1, Pop2) pair, computing a single p-value that takes into account the correlation in ancestry among the *Test* populations used as outgroups (Table S5). The only exceptions to the evidence of homogeneity are in Cusco and the Titicaca Basin (see below); hence, we split post ~2000 BP individuals in these two regions into homogeneous analysis subgroups.

The persistent regional substructure we detect over the last two millennia is notable given the dynamic changes of archaeological cultures, territorial expansions, and ever-changing intercultural interactions. Within the span of the *NorthPeruCoast* time series at the site of El Brujo from ~1750-560 BP, the Moche (~1850-1250 BP) developed and were succeeded by the Lambayeque (~1250-575 BP) [61], yet we detect no significant difference in ancestry relative to individuals from outside the region. In the *NorthPeruHighlands* we find continuity in the Ancash region at Chinchawas and LaGalgada (~1200-550 BP). In the *CentralPeruCoast* time series we find continuity in the Lima region from ~1850-480 BP through the period of cultural influence of the Highland Wari polity (~1350-950 BP) [62]. In the *SouthPeruCoast* we find continuity at Ica and Palpa from (~1480-515 BP), spanning the demise of the Nasca culture (~2050-1200 BP). In the *SouthPeruHighlands* time series we find continuity in the region including the Laramate Highlands, Mesayocpata, Charangochayoc, and Campanayuq (~1150-

390 BP) despite Wari influence. Thus, the peoples of each region in each period are consistent with having become the primary demographic substrate for those in the next, suggesting that cultural changes were largely driven by political/territorial restructuring with little evidence of large-scale mass migrations such as those that have been documented in some other regions of the world through ancient DNA [15]. This of course does not exclude the possibility of smaller scale movements of people, Elite-Dominance scenarios, or other dynamic demographic processes that left genetic signatures not detectable by our analysis; it is possible that phenomena of this type could be detected with large sample sizes which could reveal outlier individuals from some regions with different ancestries. However, our results add to the body of evidence consistent with conflict not having had a strong influence on demography over this period, most notably showing that the cultural impact of the Wari on coastal regions previously dominated by the Moche and Nasca was not mediated by large-scale population replacement or admixture [63].

The genetic structure established in each region from ~2000-500 BP is strongly echoed in the genetic structure of present-day Indigenous peoples. This is evident in the outgroup-$f_3$ based tree and MDS plots where ancient individuals cluster with modern individuals from the same region (Figure 3, Figure S5). In addition, when we computed statistics of the form $f_3$*(Mbuti; Ancient Andean, Present-Day South American)* [27], we observe qualitatively that the present-day individuals are most closely related to the ancient individuals from their region (Figure S7), a finding that is significant as measured by $f_4$-statistics (Table S6). For example, we observe excess allele sharing of the *NorthPeruCoast* individuals with Sechura (a present-day North Peru Coast group) compared to Puno (a present-day Titicaca Basin group). Thus, the forced migrations imposed by the Inca and Spanish in these regions did not completely disrupt the genetic population structure that existed prior to these events [64]. Another example is significantly more allele sharing of *CentralPeruCoast* with Quechua speakers relative to Aymara speakers from the same region [29], and, conversely, significantly more allele sharing of *NorthChile* and Titicaca Basin groups with Aymara speakers relative to Quechua speakers. This correlates to the geographic range of speakers of these two languages found today, with Aymara more circumscribed to the shores of the Titicaca Basin and southern territories, and Quechua in the north [65] (Table S6). We emphasize that this is a statement about genetic continuity, not a connection to speakers of specific languages: due to our lack of data from ancient individuals known to speak Quechua or Aymara, we cannot determine whether the language distribution followed the same pattern of geographic continuity, especially because the modern distribution of Quechua and Aymara is strongly influenced by Spanish colonial politics, as well as post-colonial state marginalization of those languages.

***Cosmopolitanism During the Tiwanaku and Inca Periods***

We document long-range mobility and genetic heterogeneity at the sites of Tiwanaku in the Titicaca Basin and Cusco associated with the administrative centers of the Tiwanaku polity (1400-950 BP) and Inca Empire (~550-420 BP), respectively and successively. At Tiwanaku, this is evident in significantly more allele sharing of *SouthPeruHighlands* with *Bolivia_Tiwanaku_1000BP* (individuals from Tiwanaku's administrative center) than with all other Titicaca Basin groups from this period in the Tiwanaku sphere of influence (spanning North Chile, Western Bolivia, and South Peru) (Table S4, Table S7A). This could potentially be explained by the pull-factor that a major administrative, religious and urban center like Tiwanaku [62] had on individuals from surrounding groups. While what we call *SouthPeruHighlands* broadly falls into the sphere of influence of the Wari polity at the time of

Tiwanaku, this does not seem to restrict such movement, which could be a sign of the limited impact the Wari polity had on some regions in their sphere of influence as suggested by some scholars (e.g. [66]). After the Tiwanaku disintegration but before the expansion of the Inca Empire, we observe two ~700 BP individuals from close to the border of present-day Chile, Peru, and Bolivia that shared more alleles with *SouthPeruHighlands* than with Titicaca Basin groups, including even Tiwanaku (Figures 2-3, Table S4, Table S7A). These individuals were from a cemetery of herders and their ancestry could be reflective of migrants from the South Peru Highlands. The archaeological record indicates that the end of both Wari and Tiwanku led to a spread of camelid pastoralism, which involved increased regional mobility and could have led to the observed migration [67, 68].

During the Inca Empire we detect significant heterogeneity in individuals within the Cusco region (San Sebastian) and the Sacred Valley (Torontoy). This is seen in Figures 3, S3 and S4 where the individuals cluster with *NorthPeruCoast*, *SouthPeruHighlands*, and Titicaca Basin groups (Table S4, Figure 4 and Table S7B). The pre-Hispanic Cusco samples are less related to present-day Cusco individuals [27] than to groups outside the region (Table S4 and Table S7B). Specifically, we find that relative to the ancient Cusco individuals from the San Sebastian or Torontoy sites, *SouthPeruHighlands* always shows significantly more allele sharing with present-day Cusco, with the signal maximized by the *Peru_Chanka_Charangochayoc_700BP* group, which originates from the site of the same name in the Lucanas province, Ayacucho Region, about 300 km to the west of Cusco. The process that led to present-day peoples of Cusco harboring ancestry distinctive from the ancient Cusco individuals is an important topic for future research. Possible scenarios include policies by the Inca or Spanish to move groups into or out of that region (*mitma* forced relocation) or recent economic diasporas into the region [69] (a large-scale rural exodus into urban areas was documented in the 19$^{th}$ and 20$^{th}$ century). These patterns could also be explained if the ancient Cusco individuals were immigrants or recent descendants of immigrants, as has been shown for burials at Machu Picchu employing morphological and isotopic data [70-72].

The dataset also highlights a case of extreme mobility during the Inca period. Published data from an Inca culture-associated boy found in the Southern Andes [20] (*Argentina_Aconcagua_500BP*) is most closely related to *NorthPeruCoast* (Figure 3, Table S3C), reflecting long-distance movement of the child for his sacrifice [17, 73, 74], likely from the same region as the two Inca period *NorthPeruCoast*-related *Cusco_Torontoy* individuals, as they form a clade with each other in *qpWave* analyses (Table S7B). This suggests that a particular site in the North Peru coastal region was likely important for the Inca (differing from prior reports that suggested the Inca sacrifice was from the Central Coast [17, 73, 74]).

***Genetic Exchange between the Northwest Amazon and North Peru***

We tested for gene flows between the Central Andes and other regions. We observe excess allele sharing between the Northwest Amazon and North Peru as shown by significantly more affinity of Indigenous peoples from the western Amazon (*SanMartin*, *Ticuna*, *Wayku*, and *Surui*) to *NorthPeruCoast*, *NorthPeruHighlands*, or *CentralPeruCoast* than to *SouthPeruCoast* or the earlier *Lauricocha_5800BP* individual (Table S7C). We used *qpAdm* to model *Peru_Amazon_SanMartin_Modern* as a mixture of 29% ancient North Highlands ancestry (related to *Peru_NorthHighlands_LaGalgada_4100BP*) and 71% Amazonian ancestry (related to *Brazil_Amazon_Karitiana_Modern*) (±11%, quoting one standard error), suggesting that at least some modern Northwest Amazonian groups harbor Andean-related ancestry. When we modeled Peruvian groups as a mixture of *Peru_WestAmazon_SanMartin_Modern* and

*Peru_NorthHighlands_LaGalgada_4100BP*, we also identified coastal individuals with significantly non-zero Amazonian-related ancestry (e.g. 39±14% in *CentralPeruCoast*) (Table S7C and Figure S6B), suggesting bi-directional gene flow with Amazonian-related ancestry affecting the Peruvian North and Central Coast more than the Highlands [7, 9, 29, 75-77]. We detect no gene flow between the North Coast and Amazonian groups to the North, as we found no significant affinity of any modern Amazonian groups from Ecuador or Colombia to *NorthPeruCoast* relative to *Lauricocha_5800BP*.

The stronger signal of Amazonian-related ancestry in the North and Central Coast relative to the North Highlands suggests that gene flow could have occurred over the low mountain ranges of North Peru (Huancabamba deflection), rather than across the high-altitude mountain ranges that dominate the Andes further south and potentially are a larger barrier to gene flow, or if Highlands groups maintained high social barriers to admixture from the Amazonian groups. We used the software *DATES* [78], which models allele covariance over genetic distance to measure admixture dates, and found that the admixtures occurred ~1478±252 years ago in *NorthPeruCoast* and ~1153±90 years ago in *CentralPeruCoast* (Table S7C), consistent with the hypothesis of a southward migration pattern.

We do not observe tropical lowlands-derived gene-flow into the Titicaca Basin or Northern Chile as reported in studies based on mitochondrial DNA [77, 79]. There is strong archaeological evidence for the exchange of food crops and other goods between the lowlands east of the Andes and the Chilean North coast (e.g. [79, 80]), but it is possible this did not lead to gene flow detectable in the *NorthChile* individuals tested here, which post-date the postulated exchange by 2000-3000 years [79]. Since we do not have any DNA from ancient Amazonians, we cannot exclude gene flow from past groups exhibiting so far undetected linages.

***Gene Flow Between the Argentine Pampas and South-Central Andes***

We detect significantly more allele sharing of *SouthPeruHighlands*, *SouthPeruCoast*, *CentralPeruCoast*, and Titicaca Basin groups to *Argentina_LagunaChica_1600BP* relative to *Argentina_LagunaChica_6800BP*. This likely reflects gene flow between the Pampas and the Central Andes, consistent with previous claims [8, 81]. Using *qpAdm,* we fit *Argentina_LagunaChica_1600BP* as a mixture of 80±12% ancestry related to *Argentina_LagunaChica_6800BP* and 20±12% ancestry related to a representative Andes group giving the lowest standard error (*CentralPeruCoast*). We also fit *CentralPeruCoast* as 77±17% related to *Peru_Cuncaicha_4200BP* and 23±17% related to *Argentina_LagunaChica_1600BP* (Figure S6C, Table S7D). Pottery and metal objects of South Andean origin are found in the Araucania region in the western Pampas dating to at least ~1000 BP [82], and skeletons from Chenque 1 in the Pampas have been suggested to have South Andean isotopic signatures [83]. Taken together, there is thus compelling evidence for human movements as well as cultural interactions between these regions at least ~1600 BP.

***Distinctive Ancestry Profile that Arrived by ~4200 BP Fully Integrated by ~2000 BP***

A previous study [18] detected a signal of differential North American-relatedness in groups from Southern Peru and North Chile after ~4200 BP relative to earlier groups. We used our data to explore the timing and geographic extent of the spread of this ancestry, using the same approach as the previous work on this topic [18]. All of the groups after ~4200 BP except for the Lauricocha individuals, *Chile_CaletaHuelen_1100BP*, and *Bolivia_Iroco_1050BP* were significant for two sources of ancestry ($p<0.05$) (Table S8 and Figure S8), suggesting that the

California Channel Island-related ancestry spread throughout all of the Andes by at least ~2000 BP. With the software DATES we measured the admixture time to be ~5000±1500 years ago (Table S8).

***Summary Model and Conclusion***

We used a semi-automated procedure to build an admixture graph to model representative ancient Central and South-Central Andeans [51] (Figures 4-5). Our best fit recapitulates key findings from this study. The earliest Peruvians do not share genetic drift with the later groups in our dataset, except for local continuity at the Lauricocha site. The differentiation between North and South Peru Highlands correlating to later structure is only evident by 5800-4100 BP. Post ~2000 BP South Peru Highlands individuals are modeled as a mixture of earlier South Highlands and North Highlands-related ancestry. Deep ancestry is inferred in Coast individuals, while North Chile individuals can only be fit with ancestry from a different basal lineage. Post ~2000 BP individuals from the socio-political center of Tiwanaku exhibit mixtures of ancestry related to contemporary people from the Central Peru Coast and South Peru Highlands. An important direction for future work is to obtain ancient DNA from the Coast prior to ~1600 BP, as well as equally rich ancient DNA data from regions to the north, west, and south of the Central Andes, which will provide further important insights.

# Acknowledgements

We thank the local Peruvian cultural heritage institutions, the Peruvian Ministry of Culture, the National Museum of Archaeology, Anthropology and History of Peru (MNAAHP), and the Universidad Nacional San Antonio Abad del Cusco for the permission to sample and do research on the archaeological skeletal remains. Permits for this work were granted by the Ministry of Culture of the government of Peru (the former National Institute of Cultural Heritage–INC) Resoluciónes Viceministeriales and Resoluciónes Directoral Nacional RDN-419-96/INC, RDN-1346, 017-2010, 120-2010, 0028-2010, 545-2011, 369-2011, 019-2010, 092-2016, 026-2018 - VMPCIC-MC, Credencial No 0/0-83-DCIRBM, and Acuerdo No 043-CRTA-INC-80. We thank the Bolivian Ministerio de Culturas y Turismo, the Viceministerio de Interculturalidad, and the Unidad de Arqueología y Museos as well as the Gobiernos Autónomos Municipales de Oruro, Tihuanacu and La Paz for granting research and export permits including Autorización UNAR N° 093/2007, UDAM–Autorización N° 015/2012, and MDCyT-UDAM N° 101/2017. We thank the Chilean Government, the Consejo de Monumentos Nacionales, Chile, the Museo San Miguel de Azapa, the Instituto de Alta Investigación, Dr. Bernardo Arriaza, and the Universidad de Tarapacá, for granting permission, and facilitating the access to the individuals from Northern Chile. The Consejo de Monumentos Nacionales, Chile, granted Order CMN No. 3904-18 for the excavation of the Pukara-6 site. We thank Mark Lipson, Vagheesh Narasimhan, Iñigo Olalde, and Nick Patterson for critical comments and discussions. **Funding:** N.N. is supported by an NIGMS (GM007753) fellowship. L.F.S. was supported by a U.S. National Science Foundation (NSF) grant (1515138), a UC-MRPI-Catalyst grant (UC-17-445724) and the Wenner-Gren Foundation (SC-14-62). P.S. is supported by the Francis Crick Institute (FC001595), which receives its core funding from Cancer Research UK, the UK Medical Research Council, and the Wellcome Trust. D.R. was supported by the U.S. National Science Foundation HOMINID grant BCS-1032255, the U.S. National Institutes of Health grant GM100233, by an Allen Discovery Center grant, by grant 61220 from the John Templeton Foundation, and is an investigator of the Howard Hughes Medical Institute. C.B. is supported by the University Research Priority

Program of Evolution in Action of the University of Zurich. M.A-D. was supported by the National Geographic project in the pilot program “Ancient DNA: Peopling of the Americas, 2018.” P.G.M. was supported by the National Geographic Society (NGS-50543R-18) and CONICET (PIP Nº 0414).

## Author contributions

WH, BL, DR, and LFS initiated the study. It was further developed working with NN, CB, PS, CP, AC, and JK. JF, MAD, KA, DBJ, RB, LC, JMC, JI, GL, CL, SLB, SCM, FM, PGM, BC, JQ, GP, MR, MRA, LS, CMS, NS, VS, MIB, IE, ETC, GV, WH, BL, and LFS excavated archaeological sites, provided or acquired samples, and contextualized the archaeological findings. RF, JSL, and CB acquired the modern genomic reference data and supervised the contextualization. BJC and TKH analyzed radiocarbon data, supervised by DJK. NR, SM, KHK, MFY, EH, MM, KS, JF, GV performed ancient DNA laboratory and data processing work, supervised by NR, BL, DR, and LFS. NN and LFS led the genetic data analysis supported by IL, PS, SM, CP and supervised by DR. NN, DR, and LFS wrote the manuscript with input from CB, PS, RB, JMC, JQ, GP, ETC, AC, WH, and BL. All authors discussed the results and contributed to the final manuscript.

## Declaration of Interests

The authors declare no competing interests.

## References

1. Rademaker K, Hodgins G, Moore K, Zarrillo S, Miller C, Bromley GR, Leach P, Reid DA, Álvarez WY, Sandweiss DH: **Paleoindian settlement of the high-altitude Peruvian Andes.** *Science* 2014, **346:**466-469.
2. Dillehay TD: *Where the land meets the sea: fourteen millennia of human history at Huaca Prieta, Peru.* University of Texas Press; 2017.
3. Chala-Aldana D, Bocherens H, Miller C, Moore K, Hodgins G, Rademaker K: **Investigating mobility and highland occupation strategies during the Early Holocene at the Cuncaicha rock shelter through strontium and oxygen isotopes.** *Journal of Archaeological Science: Reports* 2018, **19:**811-827.
4. Santoro CM, Gayo EM, Capriles JM, Rivadeneira MM, Herrera KA, Mandakovic V, Rallo M, Rech JA, Cases B, Briones L: **From the Pacific to the Tropical Forests: Networks of Social Interaction in the Atacama Desert, Late in the Pleistocene 1.** *Chungara* 2019, **51:**5-25.
5. Capriles JM, Albarracin-Jordan J, Lombardo U, Osorio D, Maley B, Goldstein ST, Herrera KA, Glascock MD, Domic AI, Veit H: **High-altitude adaptation and late Pleistocene foraging in the Bolivian Andes.** *Journal of Archaeological Science: Reports* 2016, **6:**463-474.
6. Silverman H, Isbell W: *Handbook of South American Archaeology.* Springer Science & Business Media; 2008.
7. Barbieri C, Heggarty P, Yang Yao D, Ferri G, De Fanti S, Sarno S, Ciani G, Boattini A, Luiselli D, Pettener D: **Between Andes and Amazon: The genetic profile of the Arawak-speaking Yanesha.** *American Journal of Physical Anthropology* 2014, **155:**600-609.

8. Gómez-Carballa A, Pardo-Seco J, Brandini S, Achilli A, Perego UA, Coble MD, Diegoli TM, Álvarez-Iglesias V, Martinón-Torres F, Olivieri A: **The peopling of South America and the trans-Andean gene flow of the first settlers.** *Genome research* 2018.
9. Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C: **Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire.** *Proceedings of the National Academy of Sciences* 2018**:**201720798.
10. Sandoval JR, Lacerda DR, Acosta O, Jota MS, Robles-Ruiz P, Salazar-Granara A, Vieira PPR, Paz-y-Miño C, Fujita R, Santos FR: **The Genetic History of Peruvian Quechua-Lamistas and Chankas: Uniparental DNA Patterns among Autochthonous Amazonian and Andean Populations.** *Annals of human genetics* 2016, **80:**88-101.
11. Sandoval JR, Salazar-Granara A, Acosta O, Castillo-Herrera W, Fujita R, Pena SD, Santos FR: **Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors.** *Journal of human genetics* 2013, **58:**627.
12. Baca M, Doan K, Sobczyk M, Stankovic A, Węgleński P: **Ancient DNA reveals kinship burial patterns of a pre-Columbian Andean community.** *BMC genetics* 2012, **13:**30.
13. Russo MG, Mendisco F, Avena SA, Crespo CM, Arencibia V, Dejean CB, Seldes V: **Ancient DNA reveals temporal population structure within the South-Central Andes area.** *American journal of physical anthropology* 2018.
14. Fehren-Schmitz L, Harkins KM, Llamas B: **A paleogenetic perspective on the early population history of the high altitude Andes.** *Quaternary International* 2017, **461:**25-33.
15. Valverde G, Romero MIB, Espinoza IF, Cooper A, Fehren-Schmitz L, Llamas B, Haak W: **Ancient DNA analysis suggests negligible impact of the Wari empire expansion in Peru's central coast during the Middle Horizon.** *PloS one* 2016, **11:**e0155508.
16. Fehren-Schmitz L, Haak W, Machtle B, Masch F, Llamas B, Cagigao ET, Sossna V, Schittek K, Isla Cuadrado J, Eitel B, Reindel M: **Climate change underlies global demographic, genetic, and cultural transitions in pre-Columbian southern Peru.** *Proc Natl Acad Sci U S A* 2014, **111:**9443-9448.
17. Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A, et al: **Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas.** *Sci Adv* 2016, **2:**e1501385.
18. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E: **Reconstructing the Deep Population History of Central and South America.** *Cell* 2018.
19. Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Llave CV, Witonsky D, Beall C: **The genetic prehistory of the Andean highlands 7000 years BP though European contact.** *Science Advances* 2018, **4:**eaau4921.
20. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T: **Early human dispersals within the Americas.** *Science* 2018**:**eaav2621.
21. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas AS, et al: **POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans.** *Science* 2015, **349:**aab3884.
22. Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, de Leon MP, Allentoft ME, Moltke I, et al: **The ancestry and affiliations of Kennewick Man.** *Nature* 2015, **523:**455-458.

23. Moreno-Mayar JV, Potter BA, Vinner L, Steinrucken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas AS, Sikora M, et al: **Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans.** *Nature* 2018, **553:**203-207.
24. Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Morseburg A, Johnson JR, Potter A, et al: **Ancient human parallel lineages within North America contributed to a coastal expansion.** *Science* 2018, **360:**1024-1027.
25. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Jr., Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al: **The genome of a Late Pleistocene human from a Clovis burial site in western Montana.** *Nature* 2014, **506:**225-229.
26. Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Delser PM, Velasco MS, Schraiber JG, Rasmussen S, Homburger JR, Ávila-Arcos MC: **Origins and genetic legacies of the Caribbean Taino.** *Proceedings of the National Academy of Sciences* 2018, **115:**2341-2346.
27. Barbieri C, Barquera Lozano RJ, Arias L, Sandoval JR, Acosta O, Zurita C, Aguilar-Campos A, Tito-Álvarez AM, Serrano-Osuna R, Gray RD: **The current genomic landscape of western South America: Andes, Amazonia and Pacific Coast.** *Molecular biology and evolution* 2019.
28. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al: **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.** *Nature* 2016, **538:**201-206.
29. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al: **Reconstructing Native American population history.** *Nature* 2012, **488:**370-374.
30. Bardill J, Bader AC, Garrison NA, Bolnick DA, Raff JA, Walker A, Malhi RS, Summer internship for IpiGC: **Advancing the ethics of paleogenomics.** *Science* 2018, **360:**384-385.
31. Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Nanibaa'A G: **A framework for enhancing ethical genomic research with Indigenous communities.** *Nature communications* 2018, **9:**2957.
32. Silverman H: **Cultural Resource Management and Heritage Stewardship in Peru.** *CRM: Journal of Heritage Stewardship* 2006, **3:**57-72.
33. Herrera A: *Indigenous Archaeology…in Peru?* New York: Routledge; 2011.
34. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D: **Partial uracil-DNA-glycosylase treatment for screening of ancient DNA.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20130624.
35. Renaud G, Slon V, Duggan AT, Kelso J: **Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA.** *Genome Biol* 2015, **16:**224.
36. Furtwängler A, Reiter E, Neumann GU, Siebke I, Steuri N, Hafner A, Lösch S, Anthes N, Schuenemann VJ, Krause J: **Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis.** *Scientific reports* 2018, **8:**14075.
37. Korneliussen TS, Albrechtsen A, Nielsen R: **ANGSD: Analysis of Next Generation Sequencing Data.** *BMC Bioinformatics* 2014, **15:**356.
38. Nakatsuka NJ, Harney E, Mallick S, Mah M, Patterson N, Reich DE: **ContamLD: Estimation of Ancient Nuclear DNA Contamination Using Breakdown of Linkage Disequilibrium.** *bioRxiv* 2020.
39. Silverman H: *Andean archaeology.* Blackwell; 2004.

40. Aldenfelder MS: **High elevation foraging societies.** In *The handbook of South American archaeology.* Springer; 2008: 131-143
41. Capriles JM, Santoro CM, Dillehay TD: **Harsh environments and the terminal Pleistocene peopling of the Andean highlands.** *Current Anthropology* 2016, **57:**99-100.
42. Quilter J: *The Ancient Central Andes.* Routledge; 2013.
43. Pozorski S, Pozorski T: **Early cultural complexity on the coast of Peru.** In *The Handbook of South American Archaeology.* Springer; 2008: 607-631
44. Rick JW: **The character and context of highland preceramic society.** *peruvian prehistory* 1988**:**3-40.
45. Dillehay TD, Rossen J, Andres TC, Williams DE: **Preceramic adoption of peanut, squash, and cotton in northern Peru.** *Science* 2007, **316:**1890-1893.
46. Hastorf CA: **The formative period in the Titicaca Basin.** In *The handbook of South American archaeology.* Springer; 2008: 545-561
47. Arriaza BT, Standen VG, Cassman V, Santoro CM: **Chinchorro culture: pioneers of the coast of the Atacama Desert.** In *The handbook of South American archaeology.* Springer; 2008: 45-58
48. Rivera MA: **The preceramic Chinchorro mummy complex of northern Chile: context, style, and purpose.** *Tombs for the living: Andean mortuary practices* 1995**:**43-78.
49. Goldberg A, Mychajliw AM, Hadly EA: **Post-invasion demography of prehistoric humans in South America.** *Nature* 2016, **532:**232-235.
50. Gayo EM, Latorre C, Santoro CM: **Timing of occupation and regional settlement patterns revealed by time-series analyses of an archaeological radiocarbon database for the South-Central Andes (16–25 S).** *Quaternary International* 2015, **356:**4-14.
51. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
52. Burger RL: **Understanding the Socioeconomic Trajectory of Chavín de Huántar: A New Radiocarbon Sequence and Its Wider Implications.** *Latin American Antiquity* 2019**:**1-20.
53. Burger RL: **Chavin de Huantar and its sphere of influence.** In *The handbook of South American archaeology.* Springer; 2008: 681-703
54. Matsumoto Y, Nesbitt J, Glascock MD, Palomino YIC, Burger RL: **Interregional Obsidian Exchange During the Late Initial Period and Early Horizon: New Perspectives from Campanayuq Rumi, Peru.** *Latin American Antiquity* 2018, **29:**44-63.
55. Santoro CM, Capriles JM, Gayo EM, de Porras ME, Maldonado A, Standen VG, Latorre C, Castro V, Angelo D, McRostie V: **Continuities and discontinuities in the socio-environmental systems of the Atacama Desert during the last 13,000 years.** *Journal of Anthropological Archaeology* 2017, **46:**28-39.
56. Olson E, Dodd J, Rivera M: **Prosopis sp. tree-ring oxygen and carbon isotope record of regional-scale hydroclimate variability during the last 9500 years in the Atacama Desert.** *Palaeogeography, Palaeoclimatology, Palaeoecology* 2020, **538:**109408.
57. Moraga M, Santoro CM, Standen VG, Carvallo P, Rothhammer F: **Microevolution in prehistoric Andean populations: chronologic mtDNA variation in the desert valleys of northern Chile.** *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists* 2005, **127:**170-181.
58. Rothhammer F, Moraga M, Rivera MA, Santoro C, Standen V, Carvallo P: **Contratación de las principales hipótesis sobre el origen de los constructores de**

**Tiwanaku recurriendo al análisis de ADNmt de restos esqueletales exhumados en el sitio arqueológico homónimo.** *Tiwanaku, Aproximaciones a Sus Contextos Históricos y Sociales* 2004**:**151-163.
59. Rivera MA: **The prehistory of northern Chile: A synthesis.** *Journal of World Prehistory* 1991, **5:**1-47.
60. Aufderheide AC, Kelley MA, Rivera M, Gray L, Tieszen LL, Iversen E, Krouse HR, Carevic A: **Contributions of chemical dietary reconstruction to the assessment of adaptation by ancient highland immigrants (Alto Ramirez) to coastal conditions at Pisagua, North Chile.** *Journal of Archaeological Science* 1994, **21:**515-524.
61. Castillo Butters LJ, Uceda S: **The Mochicas.** In *The Handbook of South American Archaeology.* Springer; 2008: 707-729
62. Isbell WH: **Wari and Tiwanaku: international identities in the central Andean Middle Horizon.** In *The handbook of South American archaeology.* Springer; 2008: 731-759
63. Castillo LJ: **Los últimos mochicas en Jequetepeque.** *Moche: hacia el final del milenio* 2003, **2:**65-123.
64. Iannacone G, Parra R, Bermejo M, Rojas Y, Valencia C, Portugues L, Medina M, Vallejo A, Prochanow A: **Peruvian genetic structure and their impact in the identification of Andean missing persons: A perspective from Ayacucho.** *Forensic Science International: Genetics Supplement Series* 2011, **3:**e291-e292.
65. Heggarty P: **Linguistics for archaeologists: a case-study in the Andes.** *Cambridge Archaeological Journal* 2008, **18:**35-56.
66. Jennings J: *Beyond Wari walls: regional perspectives on middle horizon Peru.* University of New Mexico Press; 2010.
67. Stanish C: *Ancient Titicaca: The evolution of complex society in southern Peru and northern Bolivia.* Univ of California Press; 2003.
68. Covey RA: **Multiregional perspectives on the archaeology of the Andes during the Late Intermediate Period (c. AD 1000–1400).** *Journal of Archaeological Research* 2008, **16:**287-338.
69. Alconini S, Covey RA: *The Oxford handbook of the Incas.* Oxford University Press; 2018.
70. Burger RL, Lee-Thorp J, Van der Merwe N: *The 1912 Yale Peruvian scientific expedition collections from Machu Picchu: human and animal remains.* Peabody Museum of Natural History: New Haven: Department of Anthropology, Yale University Division of Anthropology.; 2003.
71. Turner BL, Kamenov GD, Kingston JD, Armelagos GJ: **Insights into immigration and social class at Machu Picchu, Peru based on oxygen, strontium, and lead isotopic analysis.** *Journal of archaeological science* 2009, **36:**317-332.
72. Verano JW: *Human skeletal remains from Machu Picchu: a reexamination of the Yale Peabody Museum's collections.* New Haven, CT: Dept. of Anthropology, Yale University Division of Anthropology, Peabody Museum of Natural History; 1912.
73. Salas A, Catelli L, Pardo-Seco J, Gómez-Carballa A, Martinón-Torres F, Roberto-Barcena J, Vullo C: **Y-chromosome Peruvian origin of the 500-year-old Inca child mummy sacrificed in Cerro Aconcagua (Argentina).** *Science Bulletin* 2018.
74. Gómez-Carballa A, Catelli L, Pardo-Seco J, Martinón-Torres F, Roewer L, Vullo C, Salas A: **The complete mitogenome of a 500-year-old Inca child mummy.** *Scientific reports* 2015, **5:**16462.
75. Di Corcia T, Sanchez Mellado C, Davila Francia T, Ferri G, Sarno S, Luiselli D, Rickards O: **East of the Andes: The genetic profile of the Peruvian Amazon populations.** *American journal of physical anthropology* 2017, **163:**328-338.

76. Gnecchi-Ruscone GA, Sarno S, De Fanti S, Gianvincenzo L, Giuliani C, Boattini A, Bortolini E, Di Corcia T, Sanchez Mellado C, Francia D: **Dissecting the Pre-Columbian genomic ancestry of Native Americans along the Andes-Amazonia divide.** *Molecular biology and evolution* 2019.
77. Rothhammer F, Fehren-Schmitz L, Puddu G, Capriles J: **Mitochondrial DNA haplogroup variation of contemporary mixed South Americans reveals prehistoric displacements linked to archaeologically-derived culture history.** *American Journal of Human Biology* 2017, **29:**e23029.
78. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M: **The formation of human populations in South and Central Asia.** *Science* 2019, **365:**eaat7487.
79. Rothhammer F, Dillehay TD: **The late Pleistocene colonization of South America: an interdisciplinary perspective.** *Annals of human genetics* 2009, **73:**540-549.
80. Santoro C: **Fase Azapa transición del arcaico al desarrollo agrario inicial en los valles bajos de Arica.** *Chungara* 1980, **6:**46-56.
81. Muzzio M, Motti JM, Sepulveda PBP, Yee M-c, Cooke T, Santos MR, Ramallo V, Alfaro EL, Dipierri JE, Bailliet G: **Population structure in Argentina.** *PloS one* 2018, **13:**e0196325.
82. Berón M: **Circulación de bienes como indicador de interacción entre las poblaciones de la pampa occidental y sus vecinos.** *Arqueología en las Pampas* 2007, **1:**345-364.
83. Berón M, Luna L, Barberena R: **Isótopos de oxígeno en restos humanos del sitio Chenque I: primeros resultados sobre procedencia geográfica de individuos.** *Tendencias Teórico-metodológicas y Casos de Estudio en la Arqueología de Patagonia* 2013**:**27-38.
84. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
85. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987-2993.
86. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
87. Schubert M, Lindgreen S, Orlando L: **AdapterRemoval v2: rapid adapter trimming, identification, and read merging.** *BMC research notes* 2016, **9:**88.
88. Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K: **EAGER: efficient ancient genome reconstruction.** *Genome biology* 2016, **17:**60.
89. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
90. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
91. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M: **Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal.** *Proceedings of the National Academy of Sciences* 2014, **111:**2229-2234.
92. Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C: **HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment.** *Hum Mutat* 2013, **34:**1189-1194.
93. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F: **HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups.** *Human mutation* 2011, **32:**25-32.

94. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, Kronenberg F, Salas A, Schönherr S: **HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing.** *Nucleic acids research* 2016, **44:**W58-W63.
95. Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J: **A revised timescale for human evolution based on ancient mitochondrial genomes.** *Current biology* 2013, **23:**553-559.
96. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: **MEGA6: molecular evolutionary genetics analysis version 6.0.** *Molecular biology and evolution* 2013, **30:**2725-2729.
97. Jónsson H, Ginolhac A, Schubert M, Johnson PL, Orlando L: **mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters.** *Bioinformatics* 2013, **29:**1682-1684.
98. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic acids research* 2004, **32:**1792-1797.
99. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.
100. Díaz FP, Latorre C, Carrasco G, Wood JR, Wilmshurst JM, Soto DC, Cole TL, Gutierrez RA: **Multiscale climate change impacts on plant diversity in the Atacama Desert.** *Global Change Biology* 2019, **25:**1733-1745.
101. Hyslop J: **Chulpas of the Lupaca zone of the peruvian high plateau.** *Journal of Field Archaeology* 1977, **4:**149-170.
102. Castro V, Berenguer J, Gallardo F, Llagostera A, Salazar D: **Vertiente Occidental Circumpuneña. Desde las sociedades posarcaicas hasta las preincas (ca. 1500 años aC a 1470 dC).** *Prehistoria de Chile: desde sus Primeros Habitantes hasta los Incas* 2016**:**239-279.
103. Santoro C, Standen V: **proyecto: catastro y evaluación del patrimonio cultural arqueológico de la provincia de parinacota.** II Informe SNASPE-CONAF, Arica; 1999.
104. Capriles JM: *The Economic Organization of Early Camelid Pastoralism in the Andean Highlands of Bolivia.* Archaeopress; 2014.
105. Capriles J: **Arqueología del pastoralismo temprano de camélidos en el Altiplano central de Bolivia.** *Instituto Francés de Estudios Andinos, Plural Editores, La Paz* 2017.
106. C. Lémuz KA, and E. Arratia: *Arqueología de PutuPutu.* La Paz; 2019.
107. Reindel M, Isla J: **Los Molinos und La Muña. Zwei Siedlungszentren der Nasca-Kultur in Palpa, Südperu/Los Molinos y La Muña. Dos centros administrativos de la cultura Nasca en Palpa, costa sur del Perú.** *Beiträge zur Allgemeinen und Vergleichenden Archäologie* 2001, **21:**241-319.
108. Isla J, Reindel M (Eds.): **Palpa and Lucanas: Cultural Development Under Changing Climatic Conditions on the Western Slope of the Andes in Southern Peru**. Nova Science; 2017.
109. Meddens FM, Cook AG: **La administración Wari y el culto a los muertos: Yako, los edificios en forma "D" en la sierra sur-central del Peru.** *Wari: arte precolombino peruano* 2001**:**213-228.
110. Isbell WH: *Mummies and mortuary monuments: a postprocessual prehistory of Central Andean social organization.* University of Texas Press; 1997.
111. Cadwallader L: **Investigating 1500 Years of Dietary Change in the Lower Ica Valley, Peru Using an Isotopic Approach.** University of Cambridge, 2013.

112. Cadwallader L, Torres SA, O'Connell TC, Pullen AG, Beresford-Jones DG: **Dating the dead: new radiocarbon dates from the Lower Ica Valley, south coast Peru.** *Radiocarbon* 2015, **57:**765-773.
113. Cadwallader L, Beresford-Jones DG, Sturt FC, Pullen AG, Arce Torres S: **Doubts about How the Middle Horizon Collapsed (ca. AD 1000) and Other Insights from the Looted Cemeteries of the Lower Ica Valley, South Coast of Peru.** *Journal of Field Archaeology* 2018, **43:**316-331.
114. Matsumoto Y, Cavero Palomino Y, Gutierrez Silva R: **The Domestic Occupation of Campanayuq Rumi: Implications for Understanding the Initial Period and Early Horizon of the South-Central Andes of Peru.** *Andean Past* 2013, **11:**15.
115. Nesbitt J, Matsumoto Y, Palomino YC: **Campanayuq Rumi and Arpiri: Two Civic-Ceremonial Centers on the Southern Periphery of the Chavín Interaction Sphere.** *Ñawpa Pacha* 2019, **39:**57-75.
116. Lau GF: *Ancient Community and Economy at Chinchawas (Ancash, Peru)*. New Haven: Peabody Museum of Natural History & Yale University Publications in Anthropology (Vol. 90); 2010.
117. Lau GF: **Core-periphery relations in the Recuay hinterlands: economic interaction at Chinchawas, Peru.** *Antiquity* 2005, **79:**78-99.
118. Dillehay TD, Bonavia D, Goodbred SL, Pino M, Vásquez V, Tham TR: **A late Pleistocene human presence at Huaca Prieta, Peru, and early Pacific Coastal adaptations.** *Quaternary Research* 2012, **77:**418-423.
119. Cesareo R, Bustamante A, Jordán RF, Fernandez A, Azeredo S, Lopes RT, Alva W, Chero LZ, Brunetti A, Gigante GE: **Gold and Silver joining technologies in the Moche Tombs "Señor de Sipán" and "Señora de Cao jewelery.** *ACTA IMEKO* 2018, **7:**3-7.
120. Pablo G. Messineo NAS, Mariela E. González, Alfonsina Tripaldi, Ivana L. Ozán Jazmín Paonessa: **Estado actual de las investigaciones en la localidad arqueológica Laguna Chica (Sistema Lagunar Hinojo-Las Tunas, Trenque Lauquen).** In *XX Congreso Nacional de Arqueología Argentina*. Córdoba, Universidad Nacional de Córdoba; 2019.
121. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M: **The genomic history of the Iberian Peninsula over the past 8000 years.** *Science* 2019, **363:**1230-1234.
122. Beverly RK, Beaumont W, Tauz D, Ormsby KM, von Reden KF, Santos GM, Southon JR: **The keck carbon cycle AMS laboratory, University of California, Irvine: status report.** *Radiocarbon* 2010, **52:**301-309.
123. Ramsey CB, Lee S: **Recent and planned developments of the program OxCal.** *Radiocarbon* 2013, **55:**720-730.
124. Marsh EJ, Bruno MC, Fritz SC, Baker P, Capriles JM, Hastorf CA: **IntCal, SHCal, or a Mixed Curve? Choosing a 14 C Calibration Curve for Archaeological and Paleoenvironmental Records from Tropical South America.** *Radiocarbon* 2018, **60:**925-940.
125. Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Ramsey CB, Buck CE, Cheng H, Edwards RL, Friedrich M: **IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1869-1887.
126. Hogg AG, Hua Q, Blackwell PG, Niu M, Buck CE, Guilderson TP, Heaton TJ, Palmer JG, Reimer PJ, Reimer RW: **SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1889-1903.
127. Reimer PJ, Reimer RW: **A marine reservoir correction database and on-line interface.** *Radiocarbon* 2001, **43:**461-463.
128. Eisenmann S, Bánffy E, van Dommelen P, Hofmann KP, Maran J, Lazaridis I, Mittnik A, McCormick M, Krause J, Reich D: **Reconciling material cultures in archaeology**

**with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis.** *Scientific reports* 2018, **8:**13003.
129. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W: **From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era.** *STAR: Science & Technology of Archaeological Research* 2017, **3:**1-14.
130. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.
131. Korlevic P, Gerber T, Gansauge MT, Hajdinjak M, Nagel S, Aximu-Petri A, Meyer M: **Reducing microbial and human contamination in DNA extractions from ancient bones and teeth.** *Biotechniques* 2015, **59:**87-93.
132. Troll CJ, Kapp J, Rao V, Harkins KM, Cole C, Naughton C, Morgan JM, Shapiro B, Green RE: **A ligation-based single-stranded library preparation method to analyze cell-free DNA and synthetic oligos.** *BMC genomics* 2019, **20:**1-14.
133. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S: **Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA.** *Nucleic acids research* 2009, **38:**e87-e87.
134. Maricic T, Whitten M, Paabo S: **Multiplexed DNA sequence capture of mitochondrial genomes using PCR products.** *PLoS One* 2010, **5:**e14004.
135. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al: **An early modern human from Romania with a recent Neanderthal ancestor.** *Nature* 2015, **524:**216-219.
136. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al: **Genome-wide patterns of selection in 230 ancient Eurasians.** *Nature* 2015, **528:**499-503.
137. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al: **Massive migration from the steppe was a source for Indo-European languages in Europe.** *Nature* 2015, **522:**207-211.
138. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589-595.
139. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, Rohland N, Mallick S, Stewardson K, Kistler L, et al: **Archaeogenomic evidence reveals prehistoric matrilineal dynasty.** *Nat Commun* 2017, **8:**14115.
140. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D: **Genetic evidence for two founding populations of the Americas.** *Nature* 2015, **525:**104-108.
141. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al: **Ancient human genomes suggest three ancestral populations for present-day Europeans.** *Nature* 2014, **513:**409-413.
142. Van Oven M: **PhyloTree Build 17: Growing the human mitochondrial DNA tree.** *Forensic Science International: Genetics Supplement Series* 2015, **5:**e392-e394.
143. Van Oven M, Kayser M: **Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.** *Human mutation* 2009, **30:**E386-E394.
144. Fehren-Schmitz L, Llamas B, Lindauer S, Tomasto-Cagigao E, Kuzminsky S, Rohland N, Santos FR, Kaulicke P, Valverde G, Richards SM: **A re-appraisal of the early Andean human remains from Lauricocha in Peru.** *PloS one* 2015, **10:**e0127141.

145. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, et al: **Beringian standstill and spread of Native American founders.** *PLoS One* 2007, **2:**e829.
146. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
147. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39:**W475-478.
148. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
149. Lipson M, Reich D: **A working model of the deep relationships of diverse modern human genetic lineages outside of Africa.** *Molecular biology and evolution* 2017, **34:**889-902.
150. Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B: **Inferring admixture histories of human populations using linkage disequilibrium.** *Genetics* 2013, **193:**1233-1254.

# Figures:

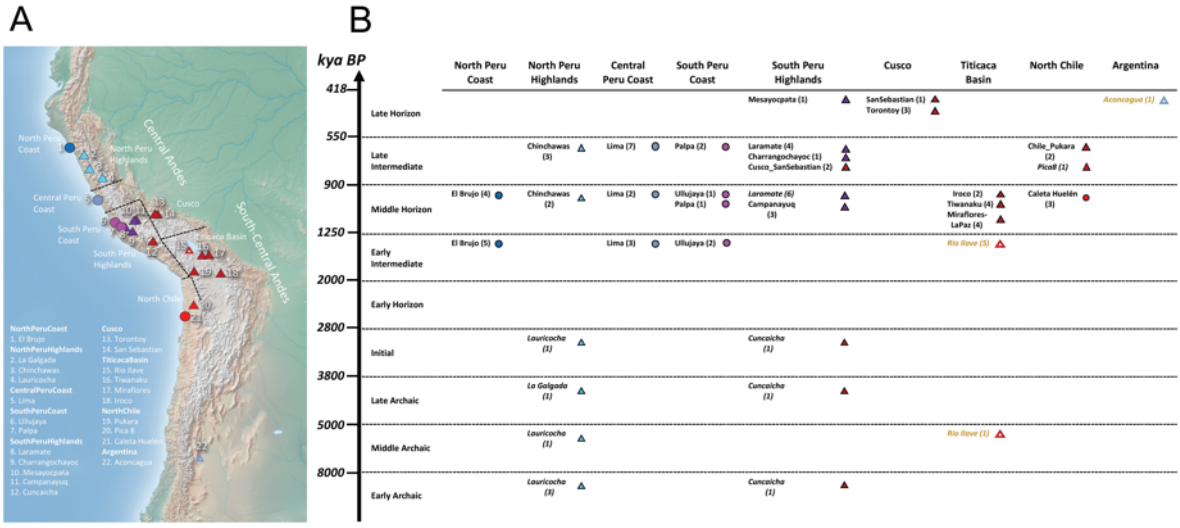


**Figure 1. Distribution of Pre-Hispanic Individuals Over Space and Time. (A)** Map with the locations of 86 ancient individuals (3 from our study are not included here due to very low coverage). Dotted lines represent regions defined for this study. Highland individuals are triangles and Coast individuals are circles. Coloring corresponds to genetic profiles, which in most cases match the geographic regions. **(B)** Groupings of ancient individuals based on geography and archaeological period (Table S1). Italics indicate previously published individuals, and sample sizes are in parentheses (yellow indicates shotgun sequences). Map was made with Natural Earth.

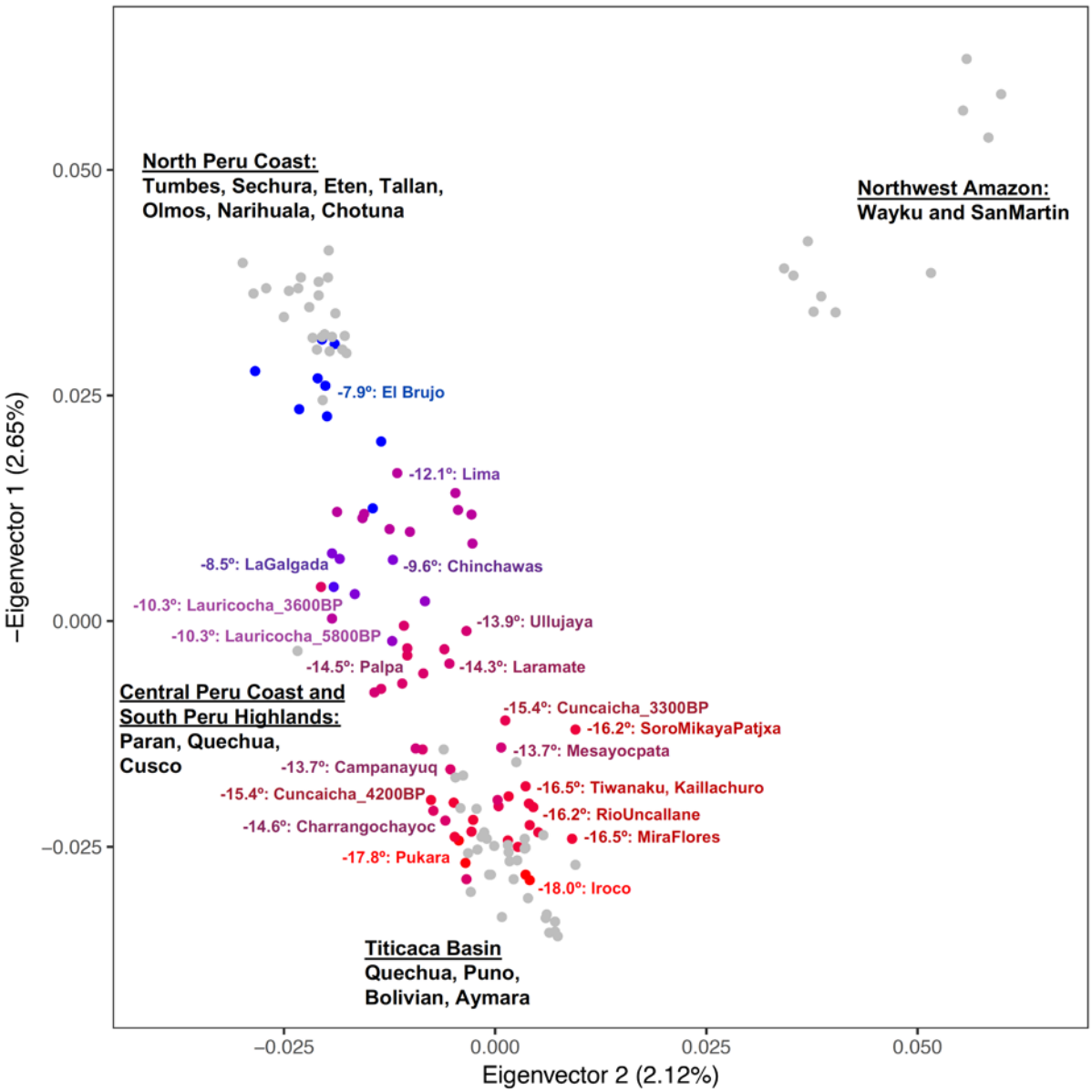


**Figure 2. Principal Components Analysis (PCA) of ancient individuals projected onto modern variation from labeled groups.** Modern individuals are in gray, and ancient individuals form a gradient that correlates to latitude (coloring is directly based on latitude with blue most north and red most south; numbers are latitude degrees). We removed 16 outliers from North Chile, Cusco, and Argentina that have evidence of ancestry from gene flows outside each region, and *Peru_Lauricocha_8600BP* and *Peru_Cuncaicha_9000BP*, which were too old to share the latitudinal cline (Figure S4 includes them). The percentage of total variation explained by each PC is shown in parentheses on each axis.

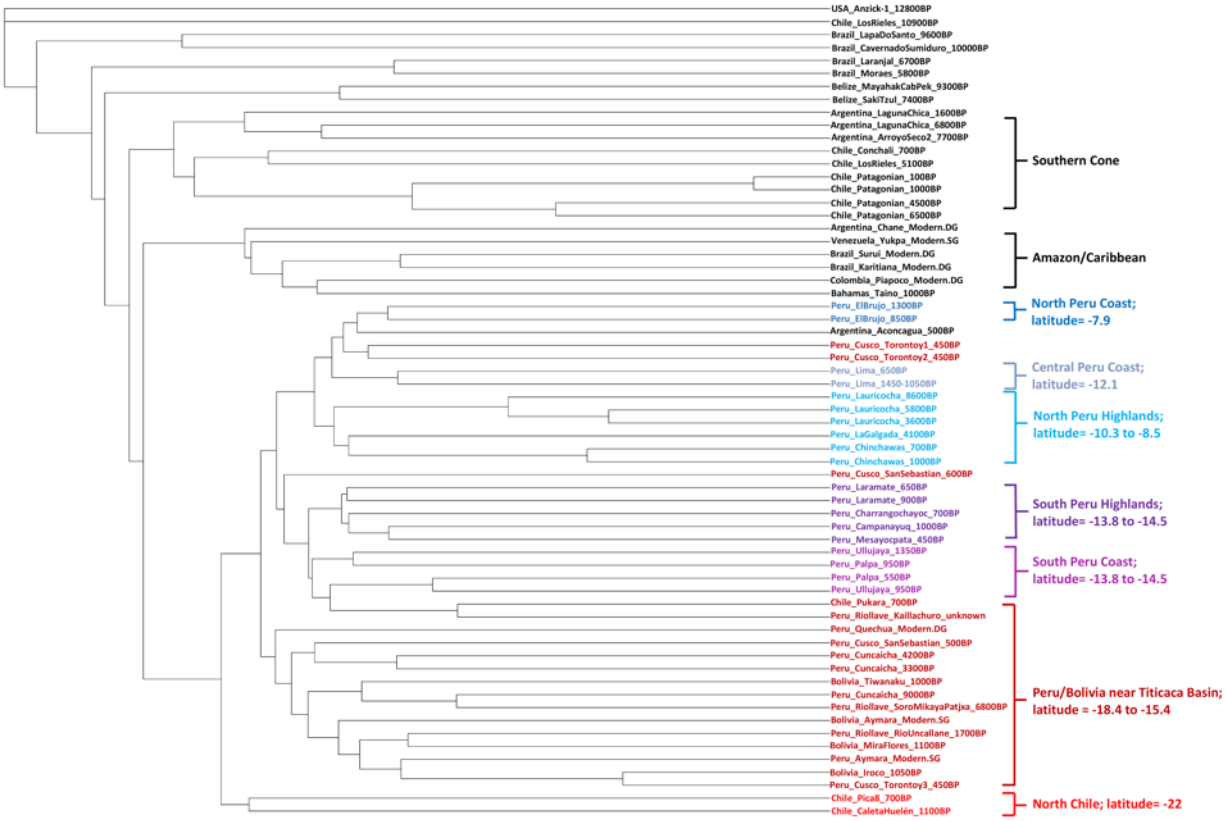


**Figure 3. Neighbor-joining tree based on inverted outgroup-$f_3$ statistics (1/$f_3$*(Mbuti; Group1, Group2)*).** Only individuals with >40,000 SNPs are included.

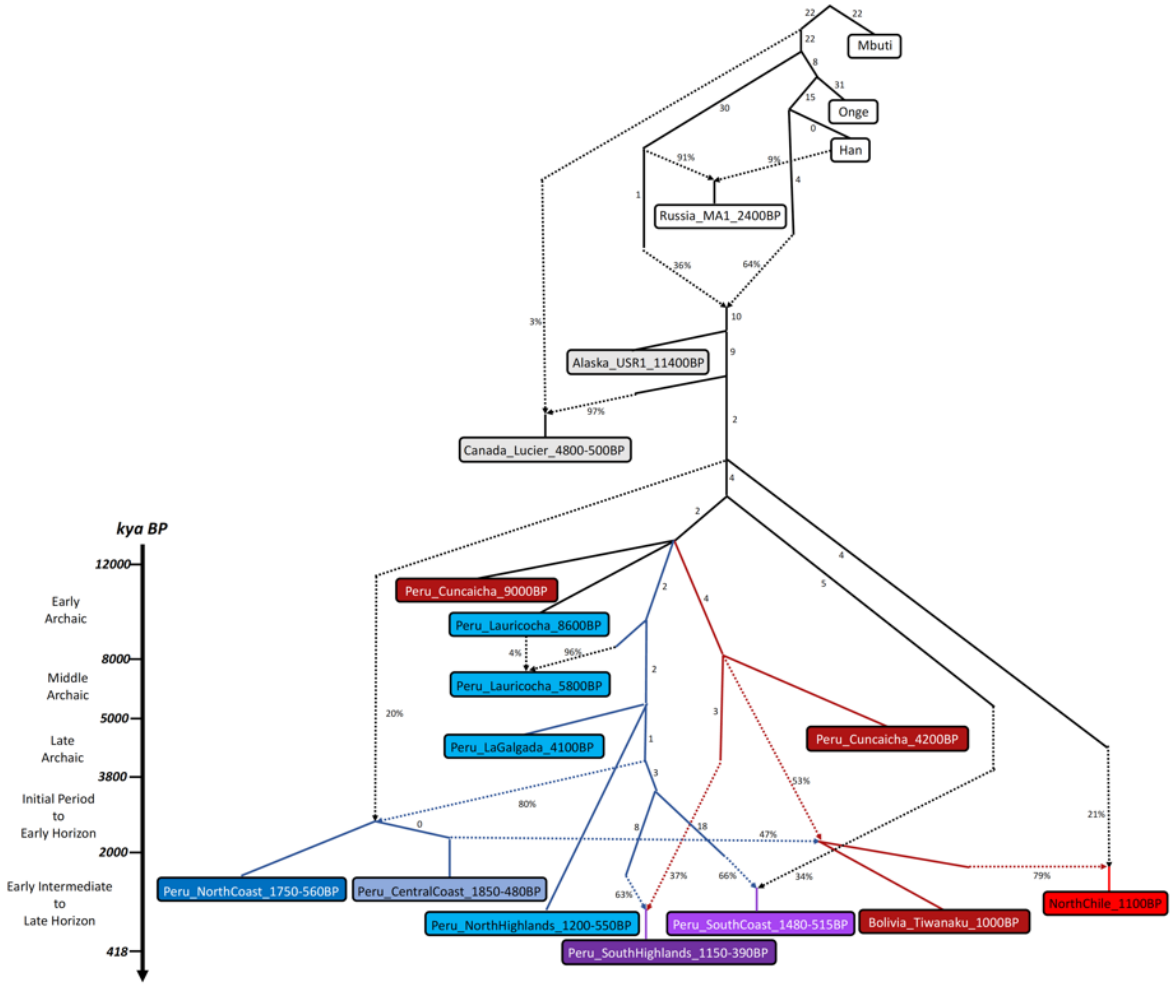


**Figure 4. Overview of findings.** Admixture graph fit (maximum |Z|-score between observed and expected *f*-statistics is 3.1). *Chile_Pica8_700BP* was removed from *NorthChile* due to low coverage.

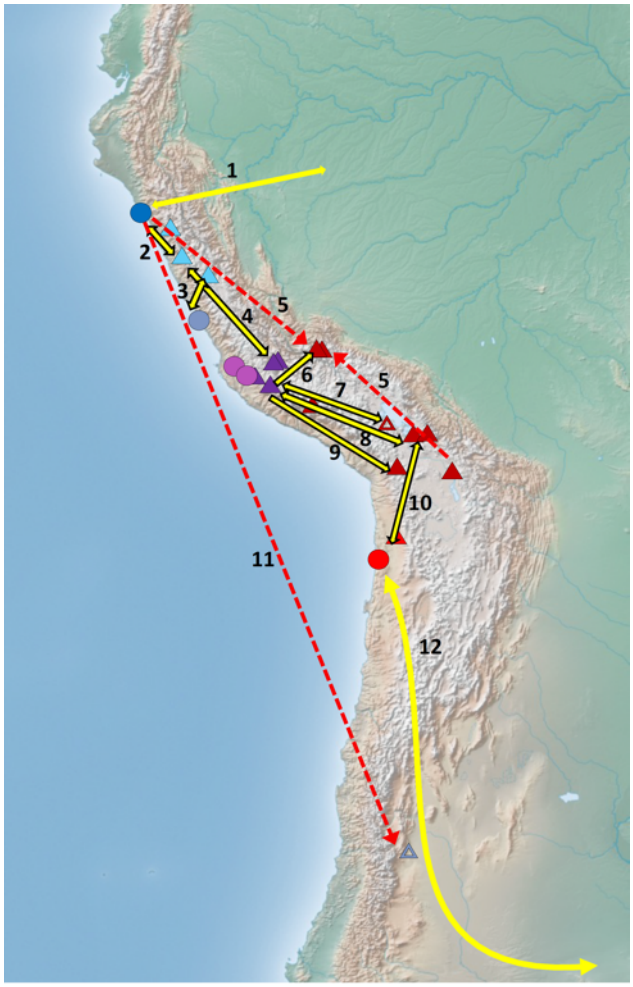


**Figure 5. Map summarizing genetic exchanges in the Central Andes. (1)** Bi-directional mixture between the North and Central Coasts and the Northwest Amazon. **(2)** Genetic exchange between *NorthPeruCoast* and *NorthPeruHighlands*. **(3)** Genetic interaction between *CentralPeruCoast* and *NorthPeruHighlands_Lauricocha* before ~5800 BP. **(4)** Genetic exchange between *NorthPeruHighlands* and *SouthPeruHighlands*. **(5)** Individuals of *NorthPeruCoast* and Titicaca Basin-related ancestry found in Cusco (Torontoy) during the Inca Empire (~450 BP). **(6)** Spread of *SouthPeruHighlands*-related ancestry into the Cusco region 450 BP-present. **(7)** Genetic exchange between *SouthPeruHighlands* and Titicaca Basin before 1700 BP. **(8)** Greater allele sharing between Tiwanaku and *SouthPeruHighlands* relative to other individuals in Titicaca Basin during the Tiwanaku period (~1000 BP). **(9)** *SouthPeruHighlands*-related ancestry found in Pukara in Northern Chile ~700 BP. **(10)** Genetic exchange between *NorthChile* and Titicaca Basin before ~1700 BP. **(11)** *NorthPeruCoast*-related ancestry found in an Inca sacrifice victim in Argentina. **(12)** Gene flow between *NorthChile* or *SouthPeruHighlands* and the Pampas region of Argentina.

# Materials and Methods:

## LEAD CONTACT AND MATERIALS AVAILABILITY

### Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Lars Fehren-Schmitz (lfehrens@ucsc.edu).

### Materials Availability Statement

This study did not generate new unique reagents.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Archaeological site information:

We generated new genome-wide data from skeletal remains of 66 ancient individuals:

| | |
|---|---|
| Caleta Huelen 12, Chile: | 3 |
| Pukara, Chile: | 2 |
| Iroco, Oruro, Bolivia: | 2 |
| Miraflores, La Paz, Bolivia: | 4 |
| Tiwanaku, La Paz, Bolivia: | 4 |
| Monte Grande, Peru: | 1 |
| Los Molinos, Palpa, Peru: | 2 |
| Laramate, Peru: | 4 |
| Charangochayoc, Peru: | 1 |
| Ullujaya, lower Ica Valley, Peru: | 3 |
| Mesayocpata, Peru: | 1 |
| Torontoy, Cusco, Peru: | 3 |
| Campanayuq Rumi, Peru: | 3 |
| San Sebastian, Cusco, Peru: | 3 |
| Huaca Pucllana, Lima: | 12 |
| Chinchawas, Peru: | 5 |
| El Brujo, Peru: | 9 |
| La Galgada, Peru: | 1 |
| Pampas, Laguna Chica, Argentina: | 1 |
| Paracas, Peru | 1 |
| Huaca Prieta | 1 |

Details of the ancient individuals and their archaeological contexts can be found in Nakatsuka *et al.*, 2020 *Cell*.

**Method Details:**

**Direct AMS $^{14}$C bone dates:** We report 39 new direct AMS $^{14}$C bone dates from 5 radiocarbon laboratories (Arizona [AA] – 2; Mannheim [MAMS] – 5; Oxford Radiocarbon Accelerator Unit [ORAU] – 11; Pennsylvania State University [PSUAMS] – 17; UC Irvine [UCIAMS] - 4) (Table S1). Bone preparation and quality control methods for most of these samples are described elsewhere. Methods for each lab are described in the following: Arizona: [104]; Mannheim: [18]; Oxford: [15, 112]; PSUAMS: [121]; UCIAMS: [122].

**Calibration of radiocarbon dates:** All calibrated $^{14}$C ages were calculated using OxCal version 4.3 [123]. Northern or southern hemisphere calibration curves used were based generally on the position of the summer Intertropical Convergence Zone (ITCZ) rather than the geographic hemisphere following recent archaeological studies in this region [124]. The IntCal13 northern hemisphere curve [125] was used for samples within the Amazon Basin and Altiplano, while the remainder were calibrated using the SHCal13 curve [126]. Dates from coastal sites were calibrated using a mixture of SHCal13 with the Marine13 curve [125] based on an estimate of a 40% marine dietary component. For each site, ΔR values were calculated based on the most proximate sample locations in the 14CHRONO Marine Reservoir Database [127] (see Table S1 for details). We recognize that there might still exist potential uncorrected biases due to the uncertainty in past carbon 14 variation in this region. However, in studies by others [124] and our own tests with different calibrations, maximum differences between different calibration choices amount to four decades at 3400BP or a maximum of a couple of decades during the brief Inca Late Horizon. All details and error ranges for the dates and calibrations are found in Table S1.

**Grouping of Individuals:** To define genetic group labels we generally used the following nomenclature: "*Country_SiteName_AgeBP*" [128]. "*AgeBP*" of a genetic group comprised of more than one individual is calculated by averaging the mean calibrated date in years before present (BP) of the directly dated samples that provided nuclear DNA data. For samples that were not directly dated we considered the averaged value of the corresponding genetic group.

**Ancient DNA Laboratory Work:** All samples in this study were processed in the dedicated clean rooms at UCSC Paleogenomics in Santa Cruz (USA), Harvard Medical School in Boston (USA), or the Australian Centre for Ancient DNA in Adelaide in Australia (ACAD), following strict procedures to minimize contamination [129]. In all three labs, DNA was extracted from bone or tooth powder using a method that is optimized to retain small DNA fragments [130, 131]. Double-stranded sequencing libraries were prepared for most samples using previously established protocols (UCSC & Harvard [34]; ACAD [17]). The sequencing libraries for Iroco, Miraflores and Torontoy were built using a single-stranded library preparation method described by Troll et al. [132]. All samples were treated with uracil-DNA glycosylase (UDG) to greatly reduce the presence of errors characteristic of ancient DNA at all sites except for the terminal nucleotides [34], or including at the terminal nucleotides (UDGplus) [133].

We enriched the libraries both for sequences overlapping mitochondrial DNA [134], and for sequences overlapping about 1.24 million nuclear targets after two rounds of enrichment [135-137]. We sequenced the enriched products on an Illumina NextSeq500 using v.2 150 cycle kits for 2×76 cycles and 2×7 cycles, and sequenced up to the point so that the expected number of new SNPs covered per 100 additional read pairs sequenced was approximately less than 1.

Enrichment was performed either at Harvard Medical School (USA), or the Max Plank Institute for Science of Human History in Jena (Germany).

To analyze the data computationally, we merged paired reads that overlapped by at least 15 nucleotides using SeqPrep (https://github.com/jstjohn/SeqPrep) taking the highest quality base to represent each nucleotide, and then mapped the sequences to the human genome reference sequence (GRCh37 from the 1000 Genomes project) using the *samse* command of the Burrows-Wheeler Aligner (*BWA*) (version 0.6.1) [138]. We trimmed two nucleotides from the end of each sequence, and then randomly selected a single sequence at each site covered by at least one sequence in each individual to represent their genotype at that position ("pseudo-haploid" genotyping).

We assessed evidence for ancient DNA authenticity by measuring the rate of damage in the first nucleotide (flagging individuals as potentially contaminated if they had a less than 3% cytosine-to-thymine substitution rate in the first nucleotide for a UDG-treated library and less than 10% substitution rate for a non-UDG-treated library). To determine kinship we computed pairwise mismatch rates between the different individuals following the same approached used in [139].

## QUANTIFICATION AND STATISTICAL ANALYSIS

**Contamination estimation in mitochondrial DNA, the X chromosome, and the autosomes:** We estimated mtDNA contamination using *contamMix* version 1.0-12 [95], which creates a Bayesian estimate of a consensus sequence composed of the true ancient DNA, error and contamination, which could come from any of a set of current human full-length mitochondrial genomes that span all plausible contaminating sequences. The software was ran with down-sampling to 50x for samples above that coverage, --trimBases X (2 bases for UDG-half samples and 10 bases for UDG-minus samples), 8 threads, 4 chains, and 2 copies, taking the first one that finished. For males we estimated X-chromosome contamination with ANGSD [37], which is based on the rate of heterozygosity observed on the X-chromosome. We used the parameters minimum base quality=20, minimum mapping quality=30, bases to clip for damage=2, and set all other parameters to the default. Lastly, we measured contamination in the autosomes using *ContamLD*, a tool based on breakdown of linkage disequilibrium that works for both males and females [38]. We report but do not include in our main analyses samples with evidence of contamination greater than 5% by any of the contamination estimation methods (samples I1400, I01, and MIS6 were excluded). All contamination estimates are reported in Table S1.

**Present-day human data:** We used present-day human data from the Simons Genome Diversity Project [28], which included 26 Native American individuals from 13 groups with high coverage full genome sequencing. We also included data from 224 Native American individuals from 34 different populations genotyped on the Affymetrix Human Origins array [27, 140, 141] as well as 493 Native American individuals genotyped on Illumina arrays either unmasked or masked to remove segments of possible European and African ancestry [29].

**Y chromosome and mitochondrial DNA analyses:** For Y chromosome haplogroup calling, we used the original BAM files and performed an independent processing procedure. We filtered out reads with mapping quality <30 and bases with base quality <30, and for UDGhalf treated libraries we trimmed the first and last 2 bp of each sequence to remove potential damage induced substitutions.

We determined the most derived mutation for each sample using the tree of the International Society of Genetic Genealogy (ISOGG) version 11.110 (accessed 21 April 2016) and confirmed the presence of upstream mutations consistent with the assigned Y chromosome haplogroup, manually checking each of the haplogroups.

To identify the mitochondrial haplotypes of the individuals, we manually analyzed each variant as described in [17] rather than relying on automated procedures. All mitochondrial reads mapped to the rCRS or RSRS using BWA were visualized in Geneious v7.1.3 (Biomatters; available from https://www.geneious.com/) for each sample. Initially, SNPs were called in Geneious for all polymorphisms with minimum coverage 5 and a minimum variant frequency 0.8. The assembly and the resulting list of SNPs were verified manually and compared to SNPs reported at phylotree.org (mtDNA tree Build 17 [18 Feb 2016]) [142]. Following recommendations in van Oven and Kayser 2009 [143], we excluded common indels and mutation hotspots at nucleotide positions 309.1C(C), 315.1C,AC indels at 515–522, 16182C, 16183C, 16193.1C(C), and C16519T. We embedded the consensus mitochondrial genomes in the existing mitochondrial tree (mtDNA tree Build 17 [18 Feb 2016]) using the online tool HaploGrep2 [94] to determine the haplotypes.

We generated a multiple genome alignment of 45 newly reported mtDNA sequences (excluding sample IO1 because of low coverage) together with 91 previously published ancient mtDNAs from western South America [17, 18, 144] and 196 modern-day sequences [17] using MUSCLE (parameter: -maxiters 2) [98]. The complete aligment cosists of 333 mtDNA sequences belonging to haplogroups A, B, C and D, plus an haplogroup L3 sequence as outgroup. The program MEGA6 [96] was used to construct a Maximum Parsimony tree with 99% partial deletion (16543 positions) and 500 bootstrap iterations that was visualized with FigTree (http://tree.bio.ed.ac.uk/software/) (Figure S2). This tree recapitulates the star-like phylogeny of the founding Southern Native American mtDNA haplogroups reported previously [145].

**ADMIXTURE clustering analysis:** Using PLINK2 [99], we first pruned our dataset using the --geno 0.7 option to ensure that we only performed our analysis on sites that had at least 70% of samples with a called genotype. We then ran ADMIXTURE [90] with 100 replicates for each K value, reporting the replicate with the highest likelihood. We show results for K=2 to 18 in Figure S3. Replications and automated filtering were performed using the UCSC-PL wrapper script adpipe.py (https://github.com/mjobin/UPA/blob/master/adpipe.py).

**Principal Components Analysis:** We performed principal components analysis (PCA) using the *smartpca* version 16680 in EIGENSOFT [89]. We used the default parameters and the lsqproject: YES, and newshrink: YES options and performed PCA on the Human Origins dataset of present-day un-admixed Andean individuals [27]. We projected the ancient individuals onto the principal components determined from the present-day individuals. When plotting the principal components we reversed the eigenvector 1 values so that the strong correspondence to the geography of Peru would be more apparent.

**Symmetry statistics and admixture tests (*f*-statistics):** We used the *qp3pop* and *qpDstat* packages in ADMIXTOOLS [51] to compute $f_3$-statistics and $f_4$-statistics (using the f4Mode: YES parameter in *qpDstat*) with standard errors computed with a weighted block jackknife over 5-Mb blocks. We used the inbreed: YES parameter to compute $f_3$-statistics to account for our random allele choice at each position (due to having too little data to determine the full diploid genotype). We computed "outgroup $f_3$"-statistics of the form $f_3(Mbuti; Pop1, Pop2)$, which measure the shared genetic drift between population 1 and population 2. We created a matrix of the outgroup-$f_3$

values between all pairs of populations. We converted these values to distances by subtracting the values from 1 and generating a multi-dimensional scaling (MDS) plot with a custom R script. We converted the original values to distances by taking the inverse of the values and generating a neighbor joining tree using PHYLIP version 3.696's [146] neighbor function and setting *USA-MT_Anzick1_12800BP* as the outgroup (default settings were used for the rest of the analysis). We displayed the tree using Itol and set all of the tree lengths to "ignore" [147]. In some of our analyses we plot the *f*-statistics on a heatmap using R (https://github.com/pontussk/point_heatmap/blob/master/heatmap_Pontus_colors.R).

**Grouping ancient samples into analysis clusters:** The ancient individuals were first grouped by archaeological site and time period based on archaeological designations (EIP=Early Intermediate Period, MH=Middle Horizon, LIP=Late Intermediate Period, LH=Late Holocene) (Figure 1B). We ran *qpWave* and computed statistics of the form $f_4$*(Mbuti, Test, Individual 1, Individual 2)* iterating over all possible pairs of individuals in each group. For the $f_4$-statistics we looked for asymmetries with any external group as *Test*. For *qpWave* analyses, we tested for evidence of two sources of ancestry relative to outgroups, which were one randomly chosen group from each geographic region outside of the region where the group was from (Table S5). We did not find evidence for more than one source of ancestry in any of the pairs except in Cusco unless a group from that geographic region (Figure 1B) was added to the outgroup set. We then performed the same procedure for pairs of groups in the same geographic region and found that we could not detect significant heterogeneity within a geographic region except in Cusco and the Titicaca Basin. Beyond the regions in Figure 1B, we could not cluster the groups further, because the *qpWave* analyses showed evidence for heterogeneity when comparing groups from different regions. For Cusco (Torontoy and SanSebastian), we kept the individuals as separate except where indicated in the text.

***qpWave* analyses:** To determine the minimum number of sources of ancestry contributing to Central Andes groups, we used *qpWave* [29], which assesses whether the set of $f_4$-statistics of the form $f_4$*(A=South American 1, B=South American 2; X=outgroup 1, Y=outgroup 2)*, which is proportional to the product of allele frequencies summed over all SNPs $(p_A-p_B)(p_X-p_Y)$, forms a matrix that is consistent with different ranks (rank 0 would mean consistency with a single stream of ancestry relative to the outgroups; rank 1 would mean 2 streams of ancestry, etc.). The significance of the statistic is assessed using a Hotelling $T^2$ test that corrects for the correlation structure of $f_4$-statistics (and thus multiple hypothesis testing). For all *qpWave* analyses, we used the default settings except for the change that we set allsnps: YES. For analyses to determine the number of waves of ancestry from North America, we used ancient California individuals from [24] (*USA_MainlandChumash_1400BP and USA_SanClemente-SantaCatalina_800BP*), *Russia_MA1_24000BP (MA1), USA-MT_Anzick1_12800BP, Papuan, Karelia Hunter Gatherer*, and modern Mexican groups (*Zapotec, Mixtec,* and *Mayan*) as outgroups.

**Admixture Graph analyses:** We used *qpGraph* [148] in ADMIXTOOLS to model the relationships between the different groups. For all analyses we removed transition SNPs at CpG sites and used default settings with outpop: Mbuti.DG and useallsnps: YES. We removed individuals I0044 and I0042 from these analyses due to their different processing in the laboratory (shotgun sequencing), which created artificial evidence of shared ancestry with similarly processed shotgun-sequenced outgroup populations. We used a previously published skeleton graph for Native Americans [140, 149] and successively added in additional populations in all combinations, allowing up to one admixture from the existing groups. We took the graph with the lowest

maximum Z-score and then repeated the process, adding another population until all populations of interest were added. For the main graph (Figure 4) we used the 1240K SNP set and first started with the two oldest individuals (*Peru_Lauricocha_8600BP* and *Peru_Cuncaicha_9000BP*) and then added on the individuals *Peru_Lauricocha_5800BP*, *Peru_Cuncaicha_4200BP*, and *Peru_LaGalgada_4100BP*. We then added the groups in order: *NorthPeruHighlands*, *SouthPeruHighlands*, *SouthPeruCoast*, *CentralPeruCoast*, *NorthPeruCoast*, *Bolivia_Tiwanaku_1000BP*, *NorthChile*. For the local graph (Figure S6A) to test the interactions between *NorthPeruHighlands* and *SouthPeruHighlands,* we first started with the individuals *Peru_Lauricocha_5800BP* and *Peru_Cuncaicha_4200BP*. We then added in *Peru_LaGalgada_4100BP*, then *NorthPeruHighlands* and *SouthPeruHighlands* in either order.

For the graph co-analyzing the Amazonians (Figure S6B), we used the Human Origins SNP set and first started with the structure from all individuals over 4,000 years old. We then added on *Peru_SanMartin_modern*, then *SouthPeruCoast*, and *NorthPeruCoast*. For the Argentina graph (Figure S6C) we began with *Argentina_LagunaChica_6800BP* and *NorthPeruHighlands*. We then added on *Bolivia_Tiwanaku_1000BP* and *NorthChile*, so that we had a mix of groups with differential affinity to *Argentina_LagunaChica_1600BP*. Lastly, we added in *Argentina_LagunaChica_1600BP*.

**Formal modeling of admixture history:** We used *qpAdm* [137] in the ADMIXTOOLS package to estimate the proportions of ancestry in a *Test* population deriving from a mixture of *N* 'reference' populations by taking advantage of the fact that they have shared genetic drift with a set of 'Outgroup' populations. We set the details: YES parameter, which reports a normally distributed Z-score for the fit (estimated with a block jackknife).

To model the genetic admixture between people related to those of the Amazon and Northwest Peru, we first modeled each of the Amazonian groups as a mixture of groups related to *Peru_LaGalgada_4100BP* and *Brazil_Karitiana_modern* with the following outgroups: *Argentina_ArroyoSeco2_7700BP*, *USA-MT_Anzick1_12800BP*, *Peru_Cuncaicha_4200BP*, *Chile_Conchali_700BP*, *Mexico_Mixe_modern*, *USA-CA_SanNicolas_4000BP*, *Chile_CaletaHuelen_1100BP*, and *Brazil_Moraes_5800BP*.

We then modeled each of the Andean groups as a mixture of *Peru_LaGalgada_4100BP* and *Peru_SanMartin_modern* with the following outgroups: *Argentina_ArroyoSeco2_7700BP*, *USA-MT_Anzick1_12800BP*, *Peru_Cuncaicha_4200BP*, *Chile_Conchali_700BP*, *Mexico_Mixe_modern*, *USA-CA_SanNicolas_4000BP*, *Chile_CaletaHuelen_1100BP*, *Brazil_Moraes_5800BP*, *Bahamas_Taino_1000BP*.

To study the admixture between groups related to those in the Argentine Pampas and the Andes, we modeled *Argentina_LagunaChica_1600BP* as a mix of *Argentina_LagunaChica_8600BP* and, in series, one of the following groups (*Peru_Laramate_900BP*, *Chile_CaletaHuelen_1100BP*, *SouthPeruHighlands*, *NorthChile*, or *CentralPeruCoast*) with the following outgroups (*Peru_Lauricocha_5800BP*, *USA-MT_Anzick1_12800BP*, *Mexico_Mixe_modern*, *USA-CA_SanNicolas_4000BP*, *Brazil_LapaDoSanto_9600BP*, *Bahamas_Taino_1000BP*, and *Peru_LaGalgada_4100BP*).

To study the ancestries of Aymara and Quechua, we modeled *Peru_AymaraIllumina* and *Peru_QuechuaIllumina* [29] as a mixture of *Bolivia_MiraFlores_1100BP* and *CentralPeruCoast* with the following outgroups (*Argentina_ArroyoSeco2_7700BP*, *USA-MT_Anzick1_12800BP*, *Peru_Cuncaicha_4200BP*, *Chile_Conchali_700BP*, *Mexico_Mixe_modern*, *USA-CA_SanNicolas_4000BP*, *Brazil_Moraes_5800BP*, *Bahamas_Taino_1000BP, Peru_Lauricocha_5800BP*).

**DATES (Distribution of Ancestry Tracts of Evolutionary Signals):** The *DATES* software [78] measures admixture dates in DNA samples by modeling the decrease in allele covariance over genetic distance in a group relative to the allele frequencies of the two source groups. This software does not require diploid information or high coverage data and can thus work well with ancient DNA samples (unlike ALDER [150], which measures admixture linkage disequilibrium directly and thus requires diploid information or multiple individuals, which is the equivalent). We used the default settings with jackknife: YES and used the software to analyze all potential admixtures we report in this study assuming 28.5 years per generation.

**Data and Code Availability:** All sequencing data are available from the European Nucleotide Archive, accession number: PRJEB37446. Genotype data obtained by random sampling of sequences at approximately 1.24 million analyzed positions are available at the Reich lab website: https://reich.hms.harvard.edu/datasets.

# Chapter 5.3: Ancient genomes in South Patagonia reveal population movements associated with technological shifts and geography

Nathan Nakatsuka[1,2,*], Pierre Luisi[3,*], Josefina M. B. Motti[4], Mónica Salemme[5,6], Fernando Santiago[5], Manuel D. D'Angelo del Campo[4,7], Rodrigo J. Vecchi[8], Yolanda Espinosa-Parrilla[9,10], Alfredo Prieto[11], Nicole Adamski[1], Ann Marie Lawson[1], Thomas K. Harper[12], Brendan J. Culleton[13], Douglas J. Kennett[14], Carles Lalueza-Fox[9], Swapan Mallick[1,15,16], Nadin Rohland[1], Ricardo A. Guichón[4], Graciela S. Cabana[17], Rodrigo Nores[3,18,*,+], David Reich[1,15,16,19,*,+]

[1] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[2] Harvard-MIT Division of Health Sciences and Technology, Boston, MA 02115, USA
[3] Universidad Nacional de Córdoba, Facultad de Filosofía y Humanidades, Departamento de Antropología, Córdoba 5000, Argentina
[4] NEIPHPA-CONICET, Facultad de Ciencias Sociales, Universidad Nacional del Centro de la Provincia de Buenos Aires, Quequén 7631, Argentina
[5] Centro Austral de Investigaciones Científicas (CADIC – CONICET), 9410 Ushuaia, Tierra del Fuego, Argentina
[6] Instituto de Cultura, Sociedad y Estado (ICSE), Universidad Nacional de Tierra del Fuego, Fuegia Basket 251, 9410 Ushuaia, Tierra del Fuego, Argentina
[7] Laboratorio de Poblaciones del Pasado (LAPP), Departamento de Biología, Facultad de Ciencias, Universidad Autónoma de Madrid (UAM), Calle Darwin 2, E-28049, Madrid, España
[8] CONICET - Departamento de Humanidades, Universidad Nacional del Sur, Bahía Blanca 8000, Argentina
[9] Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, España
[10] School of Medicine and Laboratory of Molecular Medicine - LMM, Center for Education, Healthcare and Investigation - CADI, Universidad de Magallanes, Punta Arenas, Chile
[11] Universidad de Magallanes, Avenida Bulnes 01855, Punta Arenas, Chile
[12] Department of Anthropology, The Pennsylvania State University, University Park, PA 16802, USA
[13] Institutes for Energy and the Environment, The Pennsylvania State University, University Park, PA 16802, USA
[14] Department of Anthropology, University of California, Santa Barbara, Santa Barbara, CA 93106, USA.
[15] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[16] Howard Hughes Medical Institute, Harvard Medical School, Boston, MA 02446, USA
[17] Molecular Anthropology Laboratories, Department of Anthropology, University of Tennessee, Knoxville, TN 37996, USA
[18] Instituto de Antropología de Córdoba (IDACOR), CONICET, Universidad Nacional de Córdoba, Córdoba 5000, Argentina
[19] Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

* Contributed equally
+ Co-directed this work

**Correspondence**
N.N. (Nathan_nakatsuka@hms.harvard.edu); P.L. (pierrespc@gmail.com), R.N. (rodrigonores@ffyh.unc.edu.ar); D.R. (reich@genetics.med.harvard.edu)

**Supplementary Material:** All supplementary material can be found in the supplement of Nakatsuka*, Luisi*, *et al*. 2020, *Nature Communications*.

# Abstract

At least since European contact, five ethnic groups have been living in southern Patagonia: Kawéskar in the Western Archipelagos and Yámana in the south with primarily marine-based economies; Aónikenk and Selk'nam in the north and east with primarily terrestrially-based economies; and Haush in the extreme east with a mosaic of cultural traits shared with maritime and terrestrial groups. We generated genome-wide data from 20 ancient individuals and document how genetic lineages already present in South Patagonia from the Middle Holocene (~6600-5800 BP individuals) contributed a large fraction of the ancestry of Late Holocene groups (<2000 BP individuals), except in the Western Archipelagos. Gene flow from the north by ~4700 BP added additional ancestry to groups practicing marine economies, and a later gene flow event had impact in all Late Holocene South Patagonians. From ~2200-1200 BP, mixture between neighbors resulted in a cline correlated to geographic ordering along the coast.


# Introduction

South Patagonia, defined here as the region south of 49$^{th}$ parallel in South America (Figure 1), has been occupied by humans since at least the time of the ~12600 BP Tres Arroyos rockshelter on Isla Grande de Tierra del Fuego (calendar years before present; all dates calibrated with marine reservoir effect correction in what follows) [1-5]. A handful of sites date to the Early (~13000-8500 BP) and Middle (~8500-3500 BP) Holocene, and site density increased considerably in the Late Holocene (<3500 BP) [1, 2, 6]. During this span, archaeological research has provided evidence of multiple material culture shifts that could potentially have been associated with movements of people [3].

The earliest shift relates to seafaring technology including adoption by at least ~6700 BP of canoes and harpoons which made possible the hunting of sea lions and other pinnipeds even in seasons when they were not available on the shore, allowing the settlement of nomadic hunter-gathering populations in the archipelagos [4, 5, 7]. The development of this technology has been hypothesized to reflect either *in situ* origin from land hunter-gatherers, or spread of techniques from the north via copying of ideas [6] or movement of people [8, 9].

The second shift occurred in the Western Archipelagos and involved changes in raw material and shape of tools, with green obsidian (probably sourced from the Otway Sound in the South of the Western Archipelago) as a characteristic marker of the first period [10] (~6700-6300 BP), and large bifacial lithic projectile points of different materials in the later period (~5500-3100 BP) [7, 11]. The disruption in green obsidian use has been hypothesized to reflect a loss of cultural knowledge about location of the source of this raw material, potentially due to arrival of new people unfamiliar with the landscape [12].

The third shift occurred in the northern part of Tierra del Fuego, involving technological innovations such as the use of *boleadoras* (stone spheres bound with rope used as throwing weapons) during the Middle Holocene, followed by their abandonment by ~1500 BP [13]. In addition, a new type of pedunculated lithic projectile point appeared by ~2000 BP and its size reduction around ~900 BP was associated with the appearance of bow and arrow technology [14, 15]. The similarity of these Late Holocene projectile points with those from historical times documents an element of cultural continuity at least from ~2000 BP [3], although this does not prove genetic continuity as techniques can be copied, and similar environments can lead to parallel innovations.

By the time Europeans arrived in the 16th century, five Native groups were recorded in South Patagonia (Figure 1) with two broad subsistence strategies optimized for different terrains: the plateaus and lowlands of the east and north vs. the irregular coast with islands and archipelagos in the west and south. Terrestrial hunter-gatherers included the Aónikenk (or Southern Tehuelche), who extended along the eastern slope of the mainland, and the Selk'nam (or Ona), who occupied the north of the island of Tierra del Fuego. These two groups relied primarily on hunting guanaco and birds, and gathering shellfish from the seashore [16]. The Yámana (or Yaghan) in the Beagle Channel region and the Kawéskar (or Kawésqar or Alacalufe) in the Western Archipelago (including the Otway Sound and Strait of Magellan shores) had a high reliance on marine resources that could easily be accessed by sea canoes. Finally, the Haush (or Mánekenk) of the southeastern tip of the island of Tierra del Fuego on the Mitre Peninsula did not have navigation technology, but archaeological evidence indicates that they hunted both terrestrial and marine prey [16-19]. The relationships between the five groups has been the subject of debate, with some arguing that mating among different groups was common in boundary areas [20], and others suggesting that such unions were rare [21] (see Supplementary Material for more details).

Genome-wide studies can provide direct information about whether or not movements of people accompanied changes evident in the archaeological record. Uniparental markers analysis showed that South Patagonians have only C and D mitochondrial haplogroups and low Y-chromosome diversity, consistent with a bottleneck in the founding groups followed by strong genetic drift and isolation [22-25]. de la Fuente *et al*. (2018) published the first genome-wide data from 4 individuals dated to around 1000 BP along with data from 61 modern Patagonians. The authors demonstrated a substantial degree of continuity from the archaeological individuals to present-day ones. The terrestrial Selk'nam shared alleles at an equal rate with the maritime Kawéskar and Yámana to the limit of the statistical resolution of that study [22]. Another study published two individuals of ~6600 and ~4700 BP, and showed they were more closely related to 1000 BP individuals and to historical groups than to any ancient or modern groups in other parts of South America, documenting more than 7000 years of detectable shared ancestry in the region [26].

We generated genome-wide data of 19 individuals in South Patagonia from ~5800-100 BP, including the first data of this kind from the Mitre Peninsula and inland at the south of the continent (Figure 1); we also report a ~2400 BP individual from the Pampas in Argentina. Compared with previous ancient DNA data from the region (6 pre-European contact [22, 26] and 11 post-contact [27]), our data fill in spatio-temporal gaps, particularly in the east and north of Tierra del Fuego. We combined this with previously reported data to address the following questions (referenced throughout the manuscript): Was there genetic continuity across time in the region or is there any detectable population change correlating with 1) marine diet specialization at ~6700 BP, 2) technological shifts such as the abandonment of green obsidian use between ~5500-3100 BP, and/or 3) the transition from the use of *boleadoras* to the use of pedunculated points ~2000 BP? 4) What was the extent of gene flow among neighboring South Patagonian groups? 5) Were the inhabitants of Mitre Peninsula genetically more similar to maritime or terrestrial groups? 6) How do the ancient groups relate to the ones after European contact?

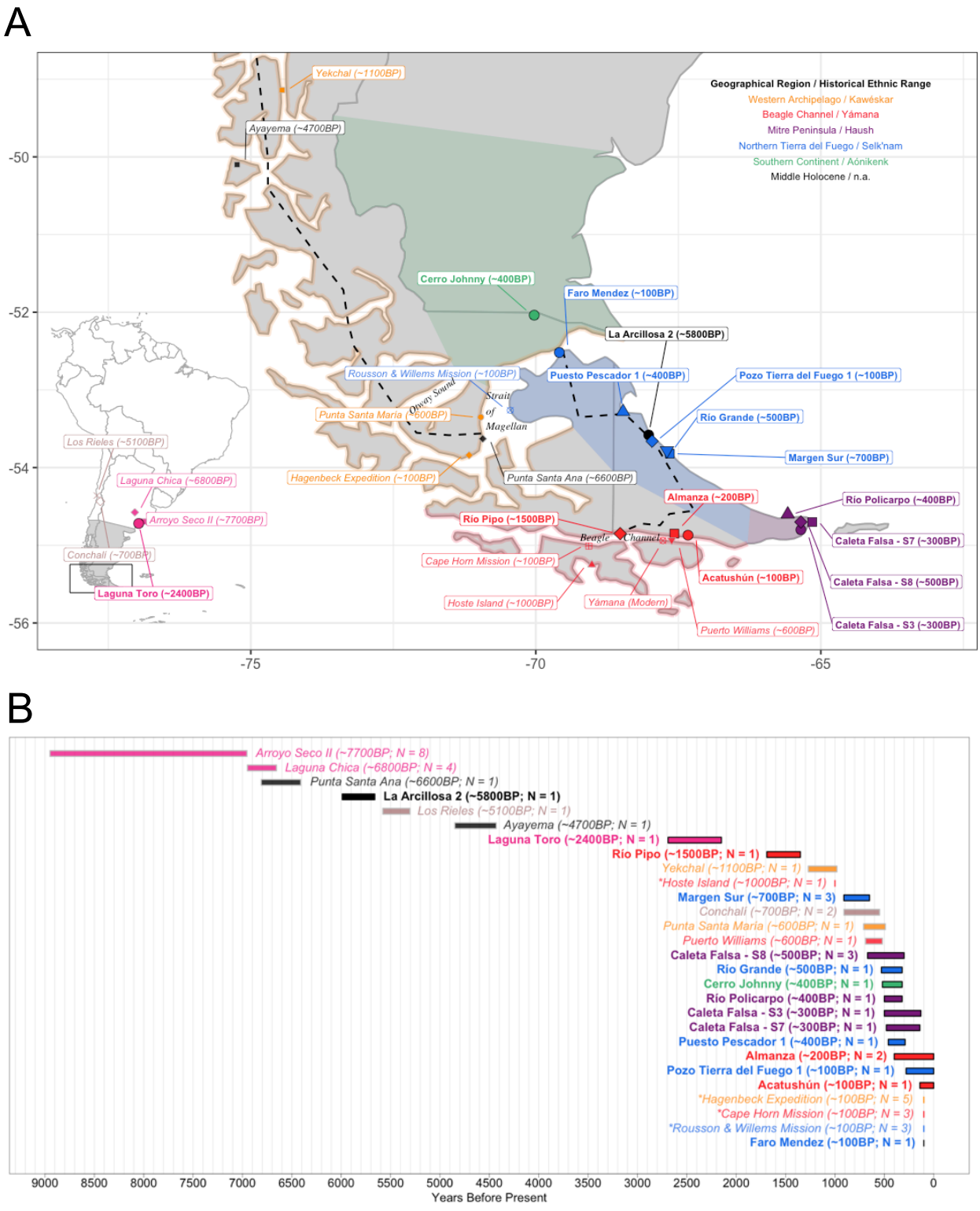


**Figure 1. Geographic and temporal distribution.** Newly reported data are in bold; color coding is in the legend. **(A) Geography.** We used site coordinates or reported location, except for Raghavan *et al*., 2015 [27] samples that were geographically reassigned according to historical evidence. The dashed lines represent routes of movement used to calculate plausible migration distances. The continuous line marks the border between Argentina in the east and Chile in the west. Inset: location of South Patagonia (rectangle) and the broader Patagonia region (following McCulloch *et al.*'s [28] definition; gray), along with the locations of ancient individuals mentioned in the main text but falling outside the range of the main map. The historical ranges of groups were adapted from Borrero *et al*., 1997 [29]. **(B) Time ranges** (number of individuals per site in parentheses).

Sites for which radiocarbon dates were not available are labeled with an asterisk. Dates were calibrated for the Southern hemisphere and corrected for maritime reservoir effect (see Methods).

### *Ethics Statement*

During more than 30 years of archaeological work in the region, one of the authors (RAG) held numerous meetings with different members of present-day Patagonian communities living in the same geographic areas where the ancient samples were located. Reaching spaces for dialogue and joint learning between members of the Native and scientific communities has been a central theme since the beginning of his research and allowed exchanging perspectives on the objectives of bioanthropological studies. In addition, on several occasions, RAG organized and carried out educational activities in schools at different educational levels in the cities of Río Grande and Ushuaia, Tierra del Fuego (Argentina). Some members of Native communities expressed interest in establishing through genetic analyses how the ancient humans found by chance and conserved in the museums relate to present-day people. This study followed that interest, and to facilitate the distribution and understanding of our findings, we translated the abstract and the main conclusions into Spanish (File S1) and shared them with members of the Native communities.

We performed this study following ethical guidelines for working with human remains, treating them with due respect as deceased humans [30]. The ancient skeletal samples we analyzed were all curated at the Museo del Fin del Mundo (Ushuaia, Argentina), the Centro Austral de Investigaciones Científicas (Ushuaia, Argentina), the Universidad Nacional del Sur (Bahía Blanca, Argentina), or the Universidad de Magallanes (Punta Arenas, Chile). Samples of the skeletal material were exported with full Argentinean and Chilean governmental permissions (see Supplement for details).

## Results and Discussion

### *Authenticity of Ancient DNA*

We verified the authenticity of the analyzed data based on all samples meeting the following criteria: 1) a rate of cytosine-to-thymine changes at the ends of the aligned fragments of >3% [31], 2) mitochondrial DNA (mtDNA) contamination point estimates below 5% [32], 3) X-chromosome contamination point estimates in males below 3% [33], and 4) genome-wide contamination [34] point estimates below 5%. No individuals were removed based on these analyses. We report but do not analyze one individual (I12365) who was found to be the brother of I12367. Supplementary Online Table 1 includes details on all the ancient individuals we analyzed.

### *Uniparental Markers, Population Size Estimates, and Variants of Phenotypic Relevance*

All mitochondrial haplotypes of the South Patagonians were C or D, reflecting the only haplogroups found in the Fuegian archipelago to date, with higher rates of D1g5 and C1c in Northern Tierra del Fuego, C1b in the Beagle Channel, and co-occurrence of C1b and D1g5 in the Mitre Peninsula (Supplementary Online Table 1). D1g5 is a widespread clade in ancient and modern samples from Argentina and Chile [35, 36], and probably differentiated in the early stages of the Southern Cone colonization, since it has geographically structured internal clades [36]. We also observed one individual who was D4h3a, which today is concentrated along the Pacific coast

of both South and North America [37]. All Y-chromosomes fall into the Q1a2a1a haplogroup except for one (Q1a2a1b), a similar skew to that seen across South America today.

We performed conditional heterozygosity analyses and found that ancient Patagonian groups had rates of variation at polymorphic sites [38] (Supplementary Figure S1) as low as the groups in the world with the lowest variation today, suggesting persistent low population size, consistent with previous inferences based on uniparental marker analyses [22-25]. We were not able to determine the date of the population bottlenecks that produced this low variation, because the three Middle Holocene Patagonians were from different sites, so there might be an upward bias when we grouped them. Moreover, higher resolution reconstruction of population size change over time requires high coverage whole genome sequencing data [39], which we do not have.

We examined the status of the analyzed individuals for several previously reported variants associated with cold tolerance [40-43]. However, the small sample sizes were insufficient to allow us to document significant allele frequency change over time and thus we were not able to carry out formal tests for natural selection (Supplementary Online Table 2).

***Correlation of Genetic Ancestry with Geography and Language***

We detect significant genetic continuity in South Patagonia since 6600 BP, as symmetry $f_4$-statistics show that the earliest Patagonians share more alleles with later Patagonians relative to Pampas, Argentina (*Argentina_ArroyoSeco2_7700BP* or *Argentina_LagunaChica_6800BP*) or Central Chile (*Chile_LosRieles_5100BP*) [44] (Supplementary Online Table 3A). We also analyzed all individuals with unsupervised ADMIXTURE (Supplementary Figure S2), Principal Components Analysis (PCA) (Supplementary Figure S3), an $F_{ST}$-based heatmap (Supplementary Figure S4), and measurements of shared genetic drift between pairs of individuals using statistics of the form $f_3$*(Mbuti; Ind1, Ind2)* (Figure 2). A Multidimensional Scaling (MDS) plot of $f_3$-statistics-based matrix shows that Middle Holocene individuals over ~5000 BP are distinct from the Late Holocene individuals (Figure 2B), with the important exception of *Chile_Ayeyama_4700BP*, which shows a slight shift toward later Western Archipelago individuals, a signal that reflects an important genetic event that we discuss in detail in what follows.

In the Late Holocene, the genetic structure of South Patagonia correlated with geography, diet/technology, and linguistic group, with largely separated clusters in the Beagle Channel region, Western Archipelago, and Southern Continent/North Tierra del Fuego. However, there are also gradients, with individuals from the Mitre Peninsula forming a cline between the North Tierra del Fuego/Southern Continent and Beagle Channel individuals; and the modern Yámana individual lying between ancient individuals from the Western Archipelago and Beagle Channel (Figure 2).

We correlated pairwise genetic drift distances to geographical, temporal, and linguistic distances (based on historically attested languages), as well as distances based on differences in subsistence resources (Table 1; Online Methods and Supplementary Online Table 4). Mantel tests were significant for distances based on all four variables (*P-values* based on 10,000 permutations <0.0002). When performing partial Mantel tests controlling for the other variables to determine if each variable had additional explanatory power beyond the others, the association remained significant for language (*P*=0.0004) and geography (*P*=0.0183*)*; we observed qualitatively similar findings when performing these analyses in other ways (Supplementary Online Table 5).

**Table 1. Correlation of Genetics with Geography, Diet/Technology, Language and Time.** *P*-values are based on 10,000 permutations. Multiple $R^2$ for Partial Mantel Test: $R^2$=0.30931

| Explanatory Variable Distance | Simple Mantel Test R | *P* | Partial Mantel Test *P* |
|---|---|---|---|
| Geography | 1.90E-01 | 2.00E-04 | 1.83E-02 |
| Diet/Technology | 6.62E-02 | 4.00E-04 | 9.75E-01 |
| Language | 2.32E-01 | <1e-4 | 4.00E-04 |
| Time | 2.31E-02 | <1e-4 | 2.58E-01 |

Based on the clear evidence for correlation of genetics with geography and post-European contact language family, we named the Late Holocene groups in each region according to the ethnic groups recognized at the time of European contact. Thus, we refer to individuals from the Western Archipelago as Kawéskar, the Beagle Channel area as Yámana, the Mitre Peninsula as Haush, the North Tierra del Fuego island as Selk'nam, and the South Continent as Aónikenk. We recognize that this is an over-simplification, because we cannot know how the individuals from ~1000-500 BP self-identified, if language differentiation reflected the languages in historical periods, if the cultures of the past were similar enough to those of historical periods to be considered continuous [45], or if there were further meaningful subdivisions beyond the groupings used at the time of European contact. Categorization into five discrete groups also masks substructure and differentiation within groups (for example, the cline in ancestry we observe in the Haush of differential relatedness to the Yámana on the one hand and the Selk'nam on the other). However, we use these names because the genetic data do not contradict the traditional terms and indeed correlate it to them strongly.

***Genetic Differentiation Between Maritime and Terrestrial Regions Apparent by ~4700 BP***

When we computed $f_4$-statistics comparing the oldest individuals, we observed that *Chile_PuntaSantaAna_6600BP* and *Argentina_LaArcillosa2_5800BP* were equally distant genetically to later groups. However, the Late Holocene Kawéskar and Yámana groups are significantly more related (|Z|>3) to *Chile_Ayayema_4700BP* than to *Chile_PuntaSantaAna_6600BP* or *Argentina_LaArcillosa2_5800BP* (Supplementary Online Table 3B), consistent with the pattern evident in Figure 2B. The fact that the ancient Selk'nam, Aónikenk, or Haush did not show significant (|Z|<1.5) affinity for *Chile_Ayayema_4700BP* suggests that the ancestry present in *Chile_Ayayema_4700BP* made a larger contribution to western groups that later relied mainly on marine resources accessible from sea canoes than to eastern groups like Selk'nam and likely Aónikenk that relied mainly on terrestrial resources. This persisted to historical times as *Chile_Ayayema_4700BP* shares more alleles (Z=3.5) with *Yamana_CapeHorneMission_Grouped_100BP* than with *Selknam_RoussonandWillemsMission_100BP* (Supplementary Online Table 3C).

Based on isotope data [46] (Supplementary Figure S5), the Western Archipelago *Chile_PuntaSantaAna_6600BP* individual is one of the first known South Patagonians with a primarily marine-based diet; while the North Tierra del Fuego *Argentina_LaArcillosa2_5800BP* individual had a primarily terrestrial diet [47, 48]. This addresses our first question: the fact that there is no significant genetic affinity of Late Holocene Western Archipelago individuals to *Chile_PuntaSantaAna_6600BP* relative to *Argentina_LaArcillosa2_5800BP* suggests that the spread of marine specialists like *Chile_PuntaSantaAna_6600BP* was not responsible for a large part of the genetic ancestry of marine specialists in the same region in the Late Holocene.

In contrast to the archaeological evidence for a shift to marine specialization, the shift away from green obsidian use in the Western Archipelagos between ~5500-3100 BP does have a

correlate in our genetic findings (our second question), as it occurred during the time of the marine-adapted *Chile_Ayayema_4700BP* individual who bears significant additional affinity to Late Holocene people from this region [26]. The fact that this individual, who is from the most northern part of South Patagonia, shows specific genetic affinity to later marine-adapted groups in South Patagonia is consistent with population change during this time frame. Specifically, it suggests gene flows connecting marine-adapted groups throughout South Patagonia (although we are not able to determine the direction of such gene flows from our genetic analysis).

All groups outside of Patagonia were symmetrically related to these earliest Patagonian groups to the limits of our resolution (Supplementary Online Table 3B), consistent with all these changes being due to local developments in the southern tip of South America, albeit with important movements within this broad region.

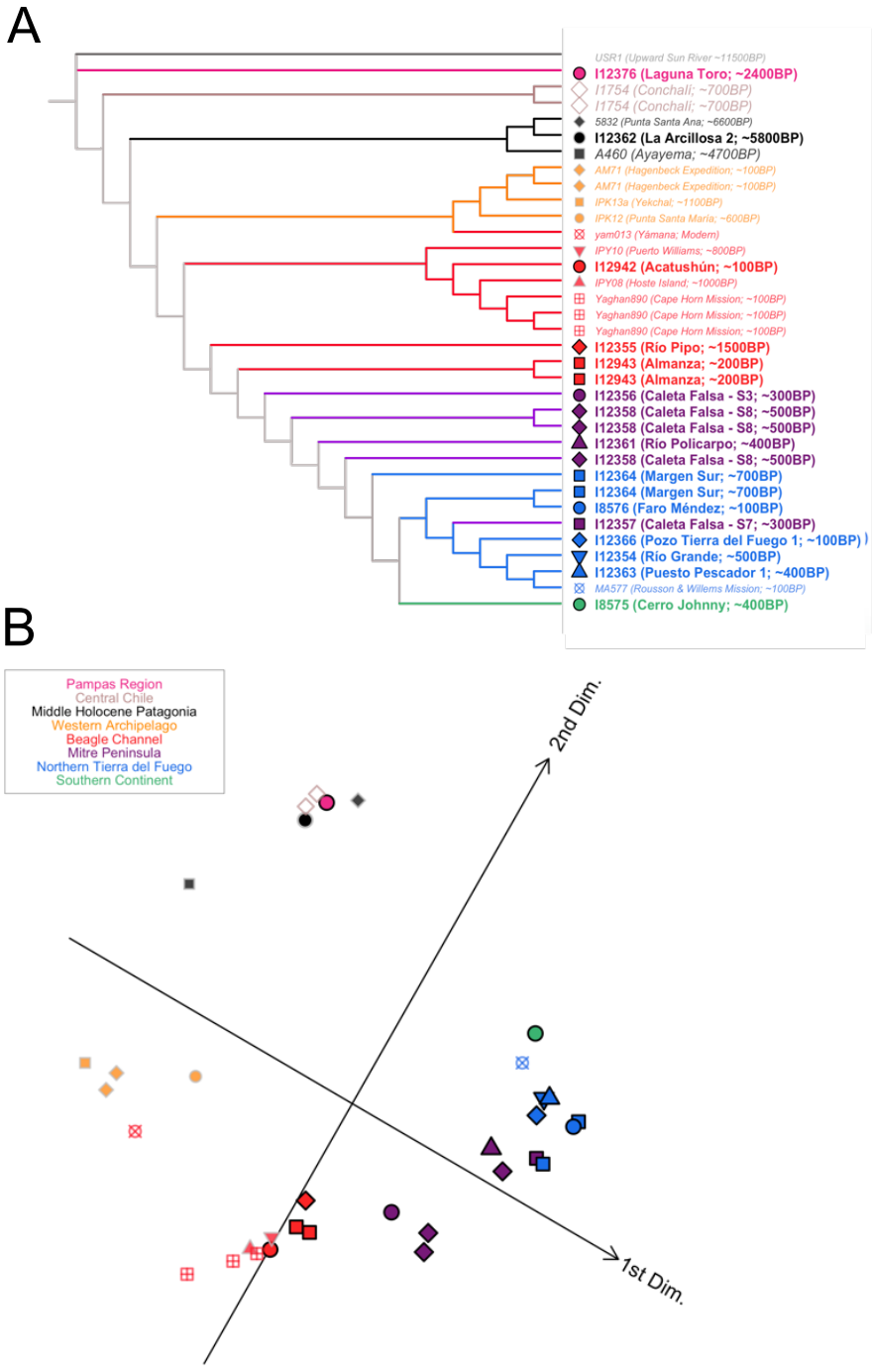

**Figure 2. Population structure in South Patagonia inferred from pairwise genetic drift. (A)** Neighbor-joining tree created using the matrix of inverted statistics 1/$f_3$*(Mbuti; Ind1, Ind2)*, with an ancient Beringian [49] as an outgroup [50, 51]; branch lengths are not meaningful but topology is. Only individuals with >100,000 SNPs were included; newly reported data are in bold; and color coding is in the legend as in Figure 1. **(B)** Multi-dimensional scaling (MDS) plot of the matrix of statistics 1-$f_3$*(Mbuti; Ind1, Ind2)*. The matrix of the first two dimensions of MDS was rotated 30 degrees to emphasize the striking geographic correlation of the genetic cline of Late Holocene samples to the coastline.

***Gene Flow from North Patagonia into South Patagonia in the Middle to Late Holocene***

To test for genetic interaction between Patagonia and other regions in America after ~4700 BP (question 3), we tested for asymmetry between Late Holocene (~1500-100 BP) vs. Middle Holocene (over ~4700 BP) Patagonian groups compared to other Native Americans, assessing if statistics of the form $f_4$(Mbuti, OtherSouthAmericans; MiddleHolocenePatagonians, LateHolocenePatagonians) were significantly different from zero. The only consistently significant signal was an excess allele sharing of *Chile_Conchali_700BP* from Central Chile far to the north with some of the later groups (Aónikenk, Haush, Yámana, and Selk'nam) relative to the Middle Holocene individuals (Supplementary Online Table 3D). There is no evidence this is due to South Patagonian gene flow into Central Chile, as statistics like $f_4$(Outgroup, SouthPatagoniaAfter4700BP; *Chile_LosRieles_5100BP_MA*, *Chile_Conchali_700BP*) were all consistent with zero (Supplementary Online Table 3E). We obtained further support for this direction of gene flow when we used *qpAdm* to attempt to model *Chile_Conchali_700BP* as a mixture of *Chile_LosRieles_5100BP* and any Late Holocene Patagonian group; in all cases, *Chile_Conchali_700BP* is modeled as consistent with having no Late Holocene Patagonian ancestry (Supplementary Online Table 6A), providing little scope for a scenario of large-scale South Patagonian ancestry moving northward into Central Chile.

Using *qpAdm* [52] to model the Late Holocene South Patagonian groups, we found that the maritime Kawéskar and Yámana could be modeled as 45-65% (±4-7%; we use 1 standard error in what follows) *Chile_Conchali_700BP*-related ancestry and the rest *Chile_Ayayema_4700BP*-related (Supplementary Online Table 6A, all models pass at p>0.02). These models work with *Chile_PuntaSantaAna_6600BP* and *Argentina_LaArcillosa2_5800BP* among the outgroups in *qpAdm*; in contrast, when *Chile_PuntaSantaAna_6600BP* or *Argentina_LaArcillosa2_5800BP* were used as the second source instead of *Chile_Ayayema_4700BP* (used as an outgroup), the models do not fit (p<0.005). These results suggest little if any direct continuity from 6600-5800 BP groups to Late Holocene maritime groups in South Patagonia, consistent with substantial re-peopling of western South Patagonia by North Patagonians (the location of *Chile_Ayayema_4700BP*) in the Middle Holocene. In contrast, the eastern Selk'nam could not be modeled with *Chile_Ayayema_4700BP*-related ancestry (p<0.005) and instead only fit (p>0.02) as a mixture of ~50-60% *Chile_Conchali_700BP*-related and the rest *Argentina_LaArcillosa2_5800BP* or *Chile_PuntaSantaAna_6600BP*-related ancestry. The Haush fit as ~50-60% *Chile_Conchali_700BP*-related and the rest as any Middle Holocene groups (we do not have resolution to resolve the source). The Aónikenk had a borderline fit (0.005<p<0.015) with ~50-60% (±6-7%) *Chile_Conchali_700BP* and either *Chile_PuntaSantaAna_6600BP* or *Chile_Ayayema_4700BP* (*Argentina_LaArcillosa2_5800BP* did not fit). Additional analyses (below) suggest Aónikenk has an ancestry most similar to that of Selk'nam, and so we favor the model of *Chile_Conchali_700BP* and *Chile_PuntaSantaAna_6600BP* which works for both groups.

In summary, all our working *qpAdm* models for the Late Holocene South Patagonians involve a mixture of about half ancestry from a group related to *Chile_Conchali_700BP,* and about half ancestry from one of the two Mid-Holocene South Patagonian lineages (*Chile_PuntaSantaAna_6600BP* or *Chile_Ayayema_4700BP*) that we have sampled in the studied dataset and that diverged from each other at least by ~6600 BP (the date of *Chile_PuntaSantaAna_6600BP*). This could be explained by north-to-south gene flow of *Chile_Conchali_700BP*-related ancestry into South Patagonia admixing into each of the divergent groups across the region, but cannot be explained by gene flow in the reverse direction, which would be expected to cause *Chile_Conchali_700BP* to be modeled as having ancestry from either the *Chile_PuntaSantaAna_6600BP*-related or *Chile_Ayayema_4700BP*-related lineage, which is not supported by $f_4$-statistics or *qpGraph* modeling (Supplementary Online Table 3E, Figure 3).

Taken together, our analyses thus suggest at least three major north-to-south gene flows affecting South Patagonia: the first bringing *Chile_PuntaSantaAna_6600BP* ancestry by at least the date of this individual, the second bringing *Chile_Ayayema_4700BP* ancestry into the western part of South Patagonia at least by ~2000 BP (the average date of formation-by-admixture of the Late Holocene ancestry cline), and the third bringing *Chile_Conchali_700BP* ancestry into all of South Patagonia again by at least ~2000 BP.

We did not detect excess allele sharing of genetic affinity of groups outside Patagonia (such as the *Argentina_LagunaToro_2400BP* or present-day Chane individuals) with the Selk'nam or Aónikenk relative to Kawéskar or Yámana (Supplementary Online Table 3F). However, our reference data are sparse, and a particular weakness is that we lack data from individuals from further south that could plausibly have interacted genetically with the Aónikenk and Selk'nam. Future ancient DNA sampling could allow to test if such groups exchanged genes in the Mid-to-Late Holocene with people in South Patagonia. We did not find ancestry from groups differentially related to non-Americans in any of the individuals ("Population Y" ancestry, Supplementary Online Table 3G), consistent with previous analyses of individuals from South Patagonia [26].

***Genetic Mixtures between Geographically Neighboring South Patagonian Groups***

To obtain insight into the extent of genetic isolation amongst the Late Holocene Patagonian groups (question 4), we computed symmetry $f_4$-statistics.

The Selk'nam were genetically intermediate between their neighbors (Figure 1A), as the Aónikenk to their north shared more alleles with them than with the Haush and Yámana to their south and west; similarly, the Haush and Yámana shared more alleles with the Selk'nam than with the Aónikenk (Supplementary Online Table 3H). Accordingly, we used *qpAdm* to model the Selk'nam as be 63.8±9.2% Aónikenk-related and 36.2% Yámana-related (Supplementary Online Table 6B). Using *DATES* (54), which studies the breakdown of allele covariance in a target group relative to two source populations, we infer an average admixture date of 1902±282 years ago, assuming a generation time of 28.5 years (Supplementary Online Table 6C).

The Haush too were genetically intermediate between their neighbors, with Yámana attracting Haush relative to Selk'nam and Selk'nam attracting Haush relative to Yámana (Supplementary Online Table 3H). We confirmed directly that the Haush are admixed (question 5) through a significantly negative (Z=-6.6) statistics of the form $f_3$*(Haush; Yámana, Selk'nam)* (Supplementary Online Table 3I). The MDS plot suggests a cline of Selk'nam- and Yámana-related ancestry in the different Haush individuals, so we used *qpAdm* to model the ancestry of each individual separately. We estimated that they vary from 10.2 to 44.8% Yámana-related (Supplementary Online Table 6C). The fact that there are substantial ancestry differences amongst

the Haush indicates that mixing between the groups may have been actively occurring in the period we sampled. We estimate an average admixture date of 1334±171 years ago.

The Yámana were also genetically intermediate among their neighbors, with the Selk'nam attracting the Yámana relative to the Kawéskar and the Kawéskar attracting the Yámana relative to the Selk'nam (Supplementary Online Table 3H). This signal was not detected in a past study [22] which reported Selk'nam as consistent with being equally related to Kawéskar and Yámana. However, that study relied on a single ~100 BP Selk'nam individual (which we confirm has symmetric relationship to Kawéskar and Yámana to the limits of the resolution of the statistics), while our additional data analyzes many Selk'nam individuals and leverages this larger dataset to successfully detect asymmetry. No significant $f_3$-statistic unambiguously demonstrated admixture (Supplementary Online Table 3I), but this could be due to lack of power or, alternatively, genetic drift in the ancestors of the Yámana since admixture, which can mask an admixture signal. We could model Yámana as 54.2±14.4% Kawéskar-related and 44.2% Selk'nam-related (one individual had 83.3±16.7% Kawéskar-related ancestry, but the others were between 51-56%). With DATES we determined the admixture date to be 1627±313 years ago (the absolute inferred dates in the past are similar even when the older Yámana individuals are used).

Taken together, these results show that there was active mixture between South Patagonian groups ~2200-1200 years ago with a cline ranging from the Aónikenk on one end to the Kawéskar on the other, and that gene flow slowed since that time (if it had continued at that rate we would not see an older date for the more recent individuals). The recent reduction of gene flow suggests the possibility that cultural differentiation became greater in the more recent period.

***Admixture Graph Model***

We used *qpGraph* to fit an admixture graph to the data and to model the relationships of the different South Patagonian groups to each other and to selected other South American groups (Figure 3). The model fit captures many of the individual findings of this study. The Pampas group *Argentina_LagunaToro_2400BP* is equally related to all South Patagonians. The lineage of *Chile_Ayayema_4700BP* is modeled as contributing to the maritime groups *Yamana_1500-100BP* and *Kaweskar_800BP* but not to other South Patagonians, reflecting the fact that genetic variation specific to later maritime groups had developed by ~4700 BP. Since *Chile_Ayayema_4700BP* is from western South Patagonia but the other earlier Middle Holocene Patagonians are not, this implies a migration related to *Chile_Ayayema_4700BP* displacing the earlier lineages. The model captures our inferences that South Patagonians after ~4700 BP have additional ancestry from a source related to the Central Chilean *Chile_Conchali_700BP* reflecting Late Holocene major north-to-south gene flow (question 3). The model also reflects the cline of ancestry as mixtures of each other with Kawéskar at one extreme and Aónikenk at the other extreme (question 4). Finally, the model confirms that *Haush_400BP*, the group with a mosaic of cultural traits shared with terrestrial and maritime groups, can be modeled as a mixture between *Selknam_500BP* (terrestrial adaptation) and *Yamana_1500-100BP* (maritime adaptation) (question 5). Without these key admixture events in the model, we could not find a graph that fit (the maximum |Z|-score between observed and expected statistics in the fitting graph is over 3.5).

***Modern Individuals are Most Related to Ancient Individuals from the Same Region***

We compared modern Yámana and Kawéskar [22, 53, 54] to ancient Patagonians (question 6), and found significant excess allele sharing of modern individuals from each regional grouping with the members of the ancient group from the same region (Supplementary Online Table 3J),

consistent with previous findings [22]. We extended these findings to the Chono, Chilote, and Huilliche, who live just north of Kawéskar, with whom they are genetically most similar. We also co-analyzed an additional dataset of modern Patagonian groups [22, 54] with the ancient groups in an admixture graph [22]. We could model the modern Yámana, Kawéskar, Huilliche, and Pehuenche (a group just north of Huilliche) as a mixture of European ancestry (reflecting post-colonial admixture), local pre-contact Native American ancestry, and Central Chile (*Chile_Conchali_700BP*)-related ancestry (Supplementary Figure S6), with a decreasing gradient of Central Chile-related ancestry from Pehuenche, Huilliche, Kawéskar, to Yámana (ordered in line with their decreasing geographic distances from *Chile_Conchali_700BP*). Thus, the ancient DNA data are capturing only the southern part of a geographic cline in *Chile_Conchali_700BP*-related ancestry in Patagonia; future ancient DNA sampling could provide additional details on the origin and timing of the development of this cline.

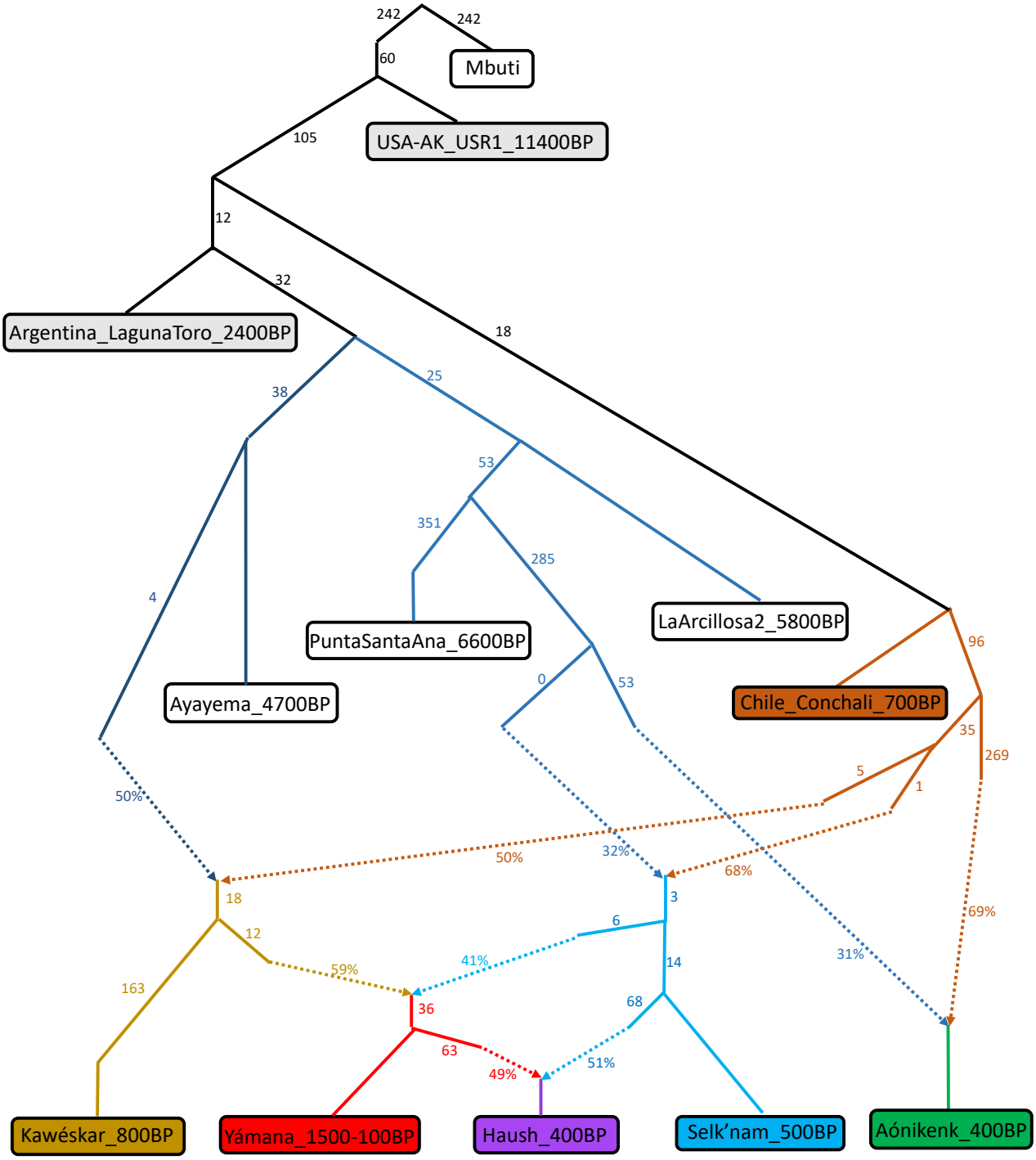


**Figure 3.** Admixture graph model summarizing key findings (maximum |Z-score| = 2.6 for a difference between observed and expected *f*-statistics) (|Z| = 2.7 restricting to transversions). The model presented fits only after adding small proportions of deeply diverging ancestry into *PuntaSantaAna_6600BP* and *Kawéskar_800BP* (splitting before the radiation of Native Americans), which we hypothesize reflects not real ancestry but rather technical artifacts due to these samples being shotgun sequenced and not UDG-treated, causing them to be attracted to the outgroup (without modeling these edges, shown in Supplementary Figure S6, the maximum |Z-score| is 5.1, but this drops to 3.3 with only transversions). Dashed lines indicate admixture between two different lineages with percentages being the admixture proportions. Numbers on solid lines are genetic drift with units of $F_{ST}$*1,000.

# Conclusion:

Our results falsify the hypothesis that the earliest marine adaptation in South Patagonia was due to a large-scale immigration into South Patagonia of people from the north who were already using this economic strategy (question 1 in the introduction). Instead, local people adopted the technology or invented it independently. However, our results indicate the arrival of a later stream of people from the northwest (following the stream that brought the mid-Holocene individuals, potentially the initial colonization event), which brought ancestry from a lineage related to the maritime-associated *Ayayema (~4700BP)* individual, replacing the lineages related to *Punta Santa Ana (~6600BP)* and *La Arcillosa2 (~5800BP)* that were previously established in South Patagonia itself. The arrival of this new stream of people could be related to the change in lithic technology around ~5500 BP, characterized by the interruption of green obsidian use and the introduction of large biface projectile points in the Western Archipelago and Beagle Channel regions [3, 12] (question 2). In addition, a third source of ancestry from Central Chile spread between ~4700-2000 BP. This could be related to different processes that occurred in the Late Holocene in Northern Tierra del Fuego, such as the increase in site density as a sign of population growth, and the cessation of use of *boleadoras* replaced by new hunting technologies (pedunculated lithic projectile points associated with the beginning of bow and arrow use by 900 BP) [3] (question 3). The shared linguistic family between North, Central, and South Patagonia groups in historical and modern times [55] could be related to this signal.

In the Late Holocene, we detect gene flow among neighbors especially from 2200-1200 years ago and attenuating afterward (question 4). A plausible scenario is that the Haush adopted some of their maritime and terrestrial adaptations from the people with whom they exchanged genes (question 5), as the genetic data demonstrates that they were socially connected through exchange of mates in this time. The Haush spoke a language in the same family (Chon) as the Selk'nam and Aónikenk, while the Yámana language is an isolate or related to Kawéskar [56], but there was nevertheless gene flow across these linguistic boundaries. Finally, population continuity in South Patagonia after European contact (question 6) is supported by the genetic affinity of modern Yámana and Kawéskar with ancient individuals from their respective regions.

We did not find evidence of genetic exchange with Argentinian groups outside Patagonia (based on lack of affinity to the Pampas individual *Argentina_LagunaToro_2400 BP* or present-day Chane). However, we do find evidence of large-scale movements of people from Chile and within Patagonia over thousands of years. An important goal for further research should be to carry out additional ancient DNA sampling not only in South (especially on the western coast) but also in Central and North Patagonia where our analysis of modern populations detects a cline of Central Chilean-related ancestry reflecting north-south gene flow, to provide higher resolution and additional insights into the interactions among people that shaped the Native cultures of this unique region of the world.

**Data availability**
All sequencing data are available from the European Nucleotide Archive, accession number found at Nakatsuka*, Luisi*, *et al*. 2020 *Nature Communications*. Genotype data obtained by random sampling of sequences at approximately 1.24 million analyzed positions are available from https://reich.hms.harvard.edu/datasets.

**Code availability**
All code available upon request.

**Acknowledgments**
We are grateful to the members of Patagonian Native communities who accompanied our work, in particular the Selk'nam, Yagán (Yámana), and the Mapuche-Tehuelche. We thank the Museo del Fin del Mundo, the Centro Austral de Investigaciones Científicas, the Universidad Nacional del Sur, and the Universidad de Magallanes for allowing us to access their collections. We thank Jakob Sedig, Mark Lipson, Matthew Mah, Iosif Lazaridis, and Iñigo Olalde for critical comments and helpful discussions. We thank Miguel Vilar for logistic support and Ricardo A. Verdugo for sharing modern Patagonian genotype data. N.N. is supported by a NIGMS (GM007753) fellowship. R.N. was supported by a National Geographic Society grant and by CONICET (PIP 2015-11220150100953CO, PUE 2016 IDACOR, and BecExt 2017). J.M.B.M. was supported by ANPCyT (PICT 2015-1405). Archaeological research in Argentina was funded by grants to M.S. (CONICET PIP 0422/10 and 6199, and ANPCyT 05-38096) and to F.S. (CONICET PIP 0302). D.R. was supported by National Institutes of Health grant GM100233, by an Allen Discovery Center grant, and by grant 61220 from the John Templeton Foundation; D.R. is also an investigator of the Howard Hughes Medical Institute. C.L.-F. was supported by the grant PGC2018-095931-B-100 (MCIU/FEDER, UE). R.N., J.M.B.M., R.A.G., M.S., F.S., and R.J.V. are members of CONICET, Argentina.

**Conflict of interests**
P.L. provides consulting services to myDNAmap S.A.

# Online Methods

**Direct AMS $^{14}$C bone dates:**
We report 15 new direct AMS $^{14}$C dates on bone and teeth for 14 ancient individuals (Supplementary Online Table 1); we refer to previous publications for the sample processing methodology [57, 58].

**Calibration of radiocarbon dates:**
All calibrated $^{14}C$ ages were calculated using OxCal version 4.3 [59], using differing mixtures of the southern hemisphere terrestrial (SHCal13 [60]) and the marine (Marine13 [61]) calibration curves. Marine dietary contribution was estimated using stable carbon and nitrogen isotope measurements from collagen (Supplementary Online Table 1). Nitrogen provides a benchmark for the relative importance of marine dietary resources, with $\delta^{15}N$ values of ~11.5‰ indicating a wholly terrestrial diet and ~22.0‰ indicating a predominately (~90%) marine diet. We delineated five categories of calibration curve mixing, assuming marine-derived diets of 0%, 20%, 40%, 60%, and 80%, respectively (Figure S5A), assuming an uncertainty value of ±10%. Probability distributions are shown in Figure S5B. Observed stable isotope distributions group by region and agree with the known subsistence strategies of the Kawéskar, Yámana, Haush, Selk'nam, and Aónikenk. We used a marine reservoir correction (ΔR value) of 221±40 from Puerto Natales, Chile [62].

**Ancient DNA Work:**
Tooth powder was obtained in dedicated clean rooms at the University of Tennessee, Knoxville using a freezer mill for 18 individuals and at Harvard Medical School by drilling for 2 individuals. DNA extraction for all samples was performed using a method optimized to retain small DNA fragments either manually [63, 64] or with an automated liquid handler using silica-coated magnetic beads [65]. We prepared double-stranded Illumina sequencing libraries, pre-treating with the enzyme Uracil-DNA Glycosylase (UDG) to minimize analytical artifacts due to the characteristic cytosine-to-thymine errors in ancient DNA [31], using an automated liquid handler and substituting the MinElute columns used for cleaning up reactions with silica-coated magnetic beads and buffer PB (Qiagen), and the MinElute column-based PCR cleanup at the end of library preparation with SPRI beads [66, 67]. We enriched the libraries for sequences that overlapped both mitochondrial DNA [68] and about 1.24 million nuclear targets for two rounds of enrichment [52, 69, 70], either independently (1240k and MT separately) or together (1240kplus). We sequenced the enriched products on an Illumina NextSeq500 using v.2 150 cycle kits for 2×76 cycles and 2×7 cycles to read the indices. Skeletal material from all 20 ancient individuals screened for this project yielded usable DNA data.

**Computational Processing of Initial Sequence Data:**
We used two different data processing methods, which have been shown to produce negligible differences in inferences about population history [71]. Supplementary Online Table 1 specifies which individuals were processed using each method.

For method 1, we merged paired forward and reverse reads that overlapped by at least 15 nucleotides using SeqPrep (https://github.com/jstjohn/SeqPrep), and used the highest quality base to represent each nucleotide. We aligned the sequences to the human genome reference sequence (GRCh37, hg19) and the reference sequence (MT RSRS) using the *samse* command of *BWA* (version 0.6.1) [72] with parameters: n=0.01, o=2, l=16500. We removed duplicates using Picard MarkDuplicates (http://broadinstitute.github.io/picard/), requiring matching indices and barcodes to declare duplicates.

For method 2, we merged paired forward and reverse reads that overlapped by at least 15 nucleotides using custom software (https://github.com/DReichLab/ADNA-Tools). We allowed one base mismatch when the forward and reverse bases both had quality at least 20 and up to three mismatches when the base read quality was less than 20, and retained the higher quality base in the case of a conflict. We restricted to merged reads of at least 30 base pairs. We aligned FASTQ files using the *BWA* (version 0.7.15-r1140) *samse* command [72] with parameters: n=0.01, o=2, l=16500 to the hg19 human reference and the MT RSRS. We removed duplicates with Picard.

For both methods, we removed two nucleotides from the end of each sequence for partial UDG treated samples and ten nucleotides for untreated samples. We selected a single sequence at each site covered by at least one sequence to represent the individual's genotype.

**Contamination estimation:**

We determined whether the data were consistent with authentic ancient DNA by measuring the damage rate in the first nucleotide, flagging individuals as potentially contaminated if they have a less than 3% cytosine to thymine substitution rate in the first nucleotide for a UDG-treated library and less than 10% substitution rate for a non-UDG-treated library as assessed using PMD tools [73]. To estimate mitochondrial contamination, we used *contamMix* version 1.0-12 [32], running the software with down-sampling to 50X for samples above that coverage. We used ANGSD to determine evidence of contamination on the X chromosome in males based on polymorphisms on the X chromosome [33] with the parameters minimum base quality=20, minimum mapping quality=30, bases to clip for damage=2, and all other parameters set to default. We also measured contamination in the autosomal DNA (chromosomes 1-22) of both males and females using a tool based on breakdown of linkage disequilibrium [34]. All samples passed quality control (individual I12941 had a 1.8% deamination rate, but we did not remove this individual because all other estimates showed negligible contamination, and low deamination is not so surprising for relatively recent samples (this individual dates to ~200 calBP).

**Kinship analyses:**

We analyzed all pairs of individuals for mismatch rates following the same approach used in Kennett, *et al.*, 2017 [74]. We removed I12365 from the main analysis dataset as we genetically detected him to be a brother of I12367 (higher coverage), but we report the data fully.

**Present-day human data:**

We used present-day human data from the Simons Genome Diversity Project [75], as well as data from 78 Native Americans genotyped on the Axiom LAT1 array [22, 76], and a whole-genome shotgun sequence of a present-day Yámana (yam013) [22].

**Y chromosome and mitochondrial DNA analyses:**

For Y chromosome haplogroup determination, we used a modified version of yHaplo (https://github.com/23andMe/yhaplo) designed to work with ancient DNA,

determining the most derived mutation for each individual using the tree of the International Society of Genetic Genealogy (ISOGG) and confirming the presence of upstream mutations consistent with the assigned Y-chromosome haplogroup using Yfitter [77].

For mtDNA haplogroup determination, we generated VCFs using Samtools [78] version 1.10 with the parameters (minimum mapping quality 30, minimum base quality 20). We ran Haplogrep [79] phylotree version 17 to attain a haplogroup assignment (Supplementary Online Table 1).

**Calculation of Distances for Different Explanatory Variables:**

We calculated distance metrics for a variety of variables.

For temporal distances between two samples, we calculated the absolute difference between the average of the 95.4% date range in calBP, defining modern samples as a date of 0.

For subsistence resources we assumed that the distance between individuals who based their subsistence primarily on terrestrial resources (Northern Tierra del Fuego and Southern Continent) and individuals for which maritime resources were most important (Western Archipelago and Beagle Channel) was 2. For individuals for whom both resources were common (Mitre Peninsula), the distance to individuals with unique primary resources was set to 1. The distance between individuals who used similar food procurement strategies was set to 0.

For linguistic group, we took into account the fact that Selk'nam/Ona, Haush, and Tehuelche/Aónikenk are Chonan languages and that the former are more similar to one another than to the latter. Accordingly, the distance between an individual belonging either to Mitre Peninsula or Northern Tierra del Fuego was set to 1, while the distance for individuals from one of these two regions to individuals from the Southern Continent was set to 2. Considering that Yámana/Yaghan and Káweskar are language isolates, we set to 4 the distance from individuals from either the Western Archipelago or the Beagle Channel to individuals from the three remaining regions. However, these two languages may be more similar to each other than to Chonan languages [55, 56, 80]; therefore, we set to 3 the distance between individuals from the Western Archipelago and individuals from the Beagle Channel. The distance between individuals from the same geographic region was set to 0.

For geographic distance, we had to contend with the complexity of the potential routes for movement of people in the area. We conjectured that a coastal route between the Mitre Peninsula and the Beagle Channel or North Tierra del Fuego was the most probable. We also considered the possibility that moving from North Tierra del Fuego to the Beagle Channel might have followed an interior route that roughly corresponds to the present-day National Road 3 (dashed line in Main Figure 1A). The routes connecting the Cerro Johnny site located at the South of the Continent to Tierra del Fuego sites were computed as described for Northern Tierra del Fuego, adding the distance between each site and its closest projection on the Northern Tierra del Fuego coast. The distance between the Cerro Johnny site and the southernmost Western Archipelago sites was calculated following the continental coast. Yámana and Káweskar people used canoes; therefore, we considered the shortest maritime routes connecting the southern part of the Western Archipelago to the Beagle Channel. The distances between the southernmost Western

Archipelago and North Tierra del Fuego and Mitre Peninsula were calculated as the sum of the length of the shortest maritime path between continental and Tierra del Fuego coasts with the length of the shortest terrestrial path in Tierra del Fuego. We also hypothesized that moving south in the Western Archipelago could be simplified by following a direct route (see dashed line in Main Figure 1A). We thus computed the distance from the Yekchal site to the others by the sum of this route to the coast close to the Punta Santa Ana site, and then by the paths described for southwestern Archipelago sites. The distance between individuals from the same site was set to 0, while the length of a straight line connecting two sites was used for any pair of sites from the same geographical region. The estimates were performed using the *geor* package in R [81] and the map from the *maps* package with resolution parameter 0 (as shown in Figure 1).

All the estimated distances are available in Supplementary Online Table 4.

**Testing Association of Genetic Distances with Variables of Interest:**

We tested if genetic distance between pairs of individuals was associated with Linguistic, Temporal, Geographical and/or Subsistence distances. Pairwise genetic distances were set to either 1-$f_3$*(Mbuti; Ind1, Ind2)* or 1/$f_3$*(Mbuti; Ind1, Ind2)*. We performed simple Spearman correlation and linear regression analyses. To correct for relationships among the explanatory variables, we also performed partial Spearman correlation and partial linear regression analyses, first correcting the genetic distances for the three other explanatory variables through a multivariate linear regression. For each coefficient estimate in the multivariate regression including the four explanatory variables, we computed a 95% confidence interval (±1.96 standard error) with a weighted block jackknife over 5-Mb blocks [82]. Finally, we performed simple (including the matrix for only one explanatory variable) and partial Mantel tests (including the distance matrices for the four explanatory variables) using the multi.Mantel function in R with 10,000 permutations of the genetic distance matrix. In all of these analyses, we assumed that each individual was independent even though this is not strictly true due to shared genetic drift within groups [82].

**Grouping of Individuals:**

All individuals were first analyzed separately for the outgroup-$f_3$ based MDS and neighbor-joining tree. For some *qpAdm* and *DATES* analyses, the individuals were then grouped by region (Western Archipelago, Beagle Channel, Mitre Peninsula, North of Tierra del Fuego island, and South Continent), sequencing method (capture or shotgun) and age as in the Figure 1A color coding scheme. In general, we name groups using the following nomenclature: "*Region_SiteName_Age BP*" [83]. "*Age BP*" of a group comprised of more than one individual is computed by averaging the mean of the estimated date range.

**Conditional Heterozygosity Analyses:**

Conditional heterozygosity is an estimate of genetic diversity in a group obtained by sampling a random allele from each of two randomly chosen individuals at a known panel of polymorphisms [38]. We performed these analyses for transversion variants on all South American groups with at least two individuals per site using *POPSTATS*

(https://github.com/pontussk/popstats) with the September 26, 2018 version with default settings. We computed this on samples from this study, ancient South Americans from Brazil [84], Central Chile [84], the Andes [26, 54, 84], the Pampas region in Argentina [84], and Patagonia [22, 26, 27], and on present-day Native American human sequencing data [54]. We restricted to individuals without substantial European admixture (inferred from the ADMIXTURE analyses below).

**ADMIXTURE Analysis:**
We merged the genotype data used for Condition Heterozygosity Analyses (but including transition sites) with Axiom LAT1 genotyping data for present-day Native Americans [22, 76], as well as 2x15 randomly sampled individuals from Italy and Spain from the Phase 3 of 1000 Genomes Project [85]. We removed ambiguous genotypes (A/T, C/G), and SNPs and individuals with more than 50% and 90% of missing genotypes. We filtered out variants with minor allele frequency <1% and pruned to remove linkage disequilibrium (--indep pairwise flag with 50 SNP windows, 5 SNP steps and 0.5 $r^2$ threshold in PLINK2). We ended with 106,285 SNPs for 116 modern Native South American individuals, 72 ancient individuals, and 30 South European reference individuals. We ran unsupervised ADMIXTURE [86] version 1.3.0 with 10 replicates for each K, reporting the replicate with the highest likelihood. We show results for K=2 to 7 in Supplementary Figure S2.

**Masking Out Regions of Non-Native Ancestry in Admixed Individuals:**
We merged the 116 modern Native South American individuals with 503 European, 504African, and 347 American individuals from 1000 Genomes Project Phase 3 [85]. After removing SNPs with more than 2% of missing genotypes and minor allele frequency below 1% and individuals with more than 10% missing genotypes, we ended with 129,269 SNPs and 1,462 individuals. After LD-pruning (--indep pairwise flag with 50 SNP windows, 5 SNP steps and 0.5 $r^2$ threshold in PLINK2), we ran unsupervised ADMIXTURE with K=3 to estimate European, African and Native American ancestry. Individuals with <99% Native American ancestry were considered as admixed, while the others were set as Native American reference. We ran RFMIX v2 [87] with 503, 504, and 69 European, African, and Native American reference samples, respectively, to identify the genomic regions with Native American ancestry in the remaining 387 admixed individuals. RFMIX was run with a setting similar to the one used for RFMIX v1 in [88], using the –n 5 flag to reduce bias. We increased the number of Expectation-Minimization (EM) iterations to 2 (--em 2 flag) instead of 1 (as in [87]) to improve the local ancestry calls and set the --reanalyze-reference flag to leverage the Native American haplotypes segregating in admixed individuals. We set both –c and –s flags to 0.2 corresponding to the -w 0.2 flag in RFMIXv1 [87]. RFMIX requires the genotype data to be pre-phased, and this was done using shapeIT2 [89] with default parameters and using the reference haplotypes for 2,504 worldwide individuals from the 1000 Genomes Project. We also used the average genetic map provided by the 1000 Genomes Project. For a given allele at a given SNP for a given individual, if the maximum posterior probability of a given ancestry was >0.9, the allele was assigned to that ancestry, otherwise it remained with unknown ancestry. We checked that the local ancestry inference procedure was consistent with global ancestry analyses and observed that the Native ancestry proportions

estimated globally (through ADMIXTURE) or locally (through RFMIXv2) have a Spearman correlation coefficient of 0.9988, with a maximum difference of 0.04. For each South American individual, we performed masking to only keep the regions that are inferred to be Native American on both chromosomes.

**Principal Components Analysis and $F_{ST}$ Analyses:**
We merged the masked genotype data for the 108 modern Native South American individuals to the South American ancient samples described above. We removed SNPs and individuals with more than 50% and 90% of missing genotypes in the compiled genotype data, respectively, obtaining genotype data for 106,981 SNPs, and 101 and 72 modern and ancient individuals, respectively. We performed PCA with *smartpca* [90], and used the default parameters except inbreed: YES, lsqproject: YES, and turning off outlier removal. We show results for PC2 vs PC1 in Supplementary Figure S3. We also used *smartpca* to compute $F_{ST}$ values between all groups that had at least 2 individuals. For this analysis we used fstonly: YES and inbreed: YES with all the other settings left at default. We show results within a heatmap for pairwise $F_{ST}$ used as a similarity index with hierarchical clustering-based dendrogram re-ordering in Supplementary Figure S4.

**Symmetry Statistics and Admixture Tests (*f*-statistics):**
We used the *qp3pop* and *qpDstat* packages in ADMIXTOOLS [44] version 6.0 to compute $f_3$-statistics and $f_4$-statistics (using the f4Mode: YES parameter in *qpDstat*) with standard errors computed with a weighted block jackknife over 5-Mb blocks. We used the inbreed: YES parameter to compute $f_3$-statistics to account for our random allele choice at each position. We computed "outgroup" $f_3$-statistics of the form $f_3$*(Mbuti; Pop1, Pop2)*, which measures the shared genetic drift between population 1 and population 2. We created a matrix of the outgroup-$f_3$ values between all pairs of populations. We converted these values to distances by subtracting the values from 1 and generating a multi-dimensional scaling (MDS) plot with a custom-made R script. We converted the original values to distances by taking the inverse of the values and generating a neighbor-joining tree using PHYLIP version 3.696's [51] neighbor function and setting *USA-AK_USR1_11400BP* as the outgroup. We displayed the tree using Itol and set all of the tree lengths to ignore [91].

**Admixture Graph Modeling:**
We used *qpGraph* [92], removing transition SNPs at CpG sites and using default settings with outpop: Mbuti.DG and useallsnps: YES. We used the 1240k dataset and created a modified graph of *Mbuti.DG*, *USA-AK_USR1_11400BP*, *Chile_Conchali_700BP*, and *Argentina_LagunaToro_2400BP*, and then successively added in additional populations in all combinations allowing up to 1 admixture from the existing groups in the graph. We took the graph with the lowest maximum Z-score and then repeated the process, adding another population until all populations of interest were added. We first started with the two oldest individuals (*Chile_PuntaSantaAna_6600BP* and *Argentina_LaArcillosa2_5800BP*). We then added the groups in order: *Chile_Ayeyama_4700BP*, *Kaweskar_WesternArchipelago_Grouped_800BP*, *Selknam_NorthTierradelFuego_Grouped_500BP*,

*Yamana_BeagleChannel_Grouped_1500-100BP*, *Aonikenk_CerroJohnny_400BP*, and *Haush_MitrePeninsula_Grouped_400BP*. We added additional deep-rooting admixture edges for *Kaweskar_WesternArchipelago_Grouped_800BP* and *Chile_PuntaSantaAna_6600BP*, because they were shotgun sequenced and processed differently which plausibly explains their (likely artifactual) attraction to Mbuti. We merged the data with the Axiom LAT1 genotype set of modern Patagonian individuals [22, 76] and added in the present-day Yámana, Kawéskar, Huilliche, and Pehuenche from that dataset, manually attempting to find the best fit with an extra admixture edge added to account for recent European admixture.

**Modeling of Ancestry Proportions:**
We used *qpAdm* [52] to estimate the proportions of ancestry in the different Late Holocene individuals. We analyzed each group as a mixture of the two groups geographically closest to them (except we never used Haush as a source population due to their genetic heterogeneity). We also used *qpAdm* to formally model the Late Holocene individuals as mixes of the Early and Middle Holocene individuals and *Chile_Conchali_700BP*. For these analyses we modeled each of the Late Holocene individuals as a mix of one of the Early or Middle Holocene individuals (*Chile_PuntaSantaAna_6600BP*, *Chile_Ayayema_4700BP*, or *Argentina_LaArcillosa2_5800BP*) and *Chile_Conchali_700BP* with the outgroups *Chane_modern*, *Peru_Cuncaicha_900BP*, *Argentina_LagunaToro_2400BP*, *Chile_LosRieles_10900BP*, *Chile_LosRieles_5100BP*, *Argentina_ArroyoSeco2_7700BP*, *Argentina_LagunaChica_6800BP*, and the other two Early and Middle Holocene Patagonia individuals. We considered models to fit if $p>0.02$; $p<0.005$ failed; and models with $0.005<p<0.02$ borderline (Supplementary Online Table 6).

**Admixture Dating Analyses:**
We used *DATES* version 1510 [93], which estimates the age of admixture in ancient DNA samples based on breakdown of allelic covariance over genetic distance in the target group relative to two source populations. We used the default settings with jackknife: YES and analyzed each group as a mixture of the two groups adjacent to them (except for Yámana, which we analyzed as a mixture of Kawéskar and Selk'nam rather than Haush due to the genetic heterogeneity in Haush).

**Analyses of Phenotypically Relevant SNPs:**
We examined the same set of SNPs as in [84] as well as additional ones with evidence of modulating cold tolerance in humans [41-43]. We used *samtools* version 1.3.1 [78, 94] *mpileup* with the settings -d 8000 -B -q30 -Q30 to obtain information about each read from the bam files of our samples. We used the fasta file from human genome GRCh37 (hg19) for the pileup. We counted the number of derived and ancestral variants at each analyzed position using a custom Python script.

# References


1. Borrero LA, McEwan C: **The peopling of Patagonia: The first human occupation.** *Patagonia Natural History, Prehistory and Ethnography at the uttermost end of the Earth Princeton University Press, New Jersey* 1997**:**32-45.
2. Salemme MC, Miotti LL: **Archeological hunter-gatherer landscapes since the latest Pleistocene in Fuego-Patagonia.** *Developments in Quaternary Sciences* 2008, **11:**437-483.
3. Morello F, Borrero L, Massone M, Stern C, García-Herbst A, McCulloch R, Arroyo-Kalin M, Calás E, Torres J, Prieto A: **Hunter-gatherers, biogeographic barriers and the development of human settlement in Tierra del Fuego.** *Antiquity* 2012, **86:**71-87.
4. Orquera LA, Piana EL: **Sea nomads of the Beagle Channel in Southernmost South America: over six thousand years of coastal adaptation and stability.** *The Journal of Island and Coastal Archaeology* 2009, **4:**61-81.
5. Piana EL, Orquera LA: **The southern top of the world: The first peopling of Patagonia and Tierra del Fuego and the cultural endurance of the Fuegian Sea-Nomads.** *Arctic Anthropology* 2009, **46:**103-117.
6. Borrero LA, Franco NV: **Early Patagonian hunter-gatherers: Subsistence and technology.** *Journal of Anthropological Research* 1997, **53:**219-239.
7. Orquera LA, Legoupil D, Piana EL: **Littoral adaptation at the southern end of South America.** *Quaternary International* 2011, **239:**61-69.
8. Massone M: **Fell 1 Hunters' hearths in the Magallanes Region by the end of the Pleistocene.** *Where the south winds blow Center for the Study of the First Americans, College Station, TX* 2003**:**153-160.
9. Prieto A, Stern CR, Estévez JE: **The peopling of the Fuego-Patagonian fjords by littoral hunter–gatherers after the mid-Holocene H1 eruption of Hudson Volcano.** *Quaternary International* 2013, **317:**3-13.
10. Morello Repetto F, San Román Bontes M, Prieto Iglesias A, Stern C: **Nuevos antecedentes para una discusión arqueológica en torno a la obsidiana verde en Patagonia Meridional.** *Anales del Instituto de la Patagonia* 2001, **29:**129-148.
11. Huidobro M: **Perspectiva funcional del equipamiento lítico tallado de las sociedades canoeras de Magallanes entre los ca. 4.400-3.000 años ap. Nuevos resultados a partir del análisis traceológico de pizzulic 3 y offing 2-locus 1 (componente inferior).** *Magallania (Punta Arenas)* 2018, **46:**203-230.
12. Morello F, Stern C, San Román M: **Obsidiana verde en Tierra del Fuego y Patagonia: caracterización, distribución y problemáticas culturales a lo largo del Holoceno.** *Intersecciones en Antropología* 2015, **16:**139-153.
13. Elgueta JT: **Bolas líticas y sus procesos de manufactura, en contextos de cazadores recolectores terrestres del norte de Tierra del Fuego. Evidencias desde el Holoceno Medio hasta 1500 años AP.** *Arqueología de la Patagonia (Vol 1, Tomo Ushuaia)* 2009, **1:**393-412.
14. Banegas A, Gómez Otero J, Goye S, Ratto N: **Cabezales líticos del Holoceno tardío en Patagonia meridional: Diseños y asignación funcional.** *Magallania (Punta Arenas)* 2014, **42:**155-174.

15. Huidobro C: **Fabricación de puntas de proyectil en los niveles tardíos de la cueva Tres Arroyos 1, Tierra del Fuego.** *Magallania (Punta Arenas)* 2012, **40:**185-201.
16. Santiago FC, Vázquez M: **Dietas promediadas: explorando el registro zooarqueológico supra-regional en Tierra del Fuego.** *Revista del Museo de Antropología* 2012, **5:**225-238.
17. Yesner D, Figuerero M, Guichón R, Borrero L: **Análisis de isótopos estables en esqueletos humanos: confirmación de patrones de subsistencia etnográficos para Tierra del Fuego.** *Shincal* 1991, **3:**182-191.
18. Panarello H, ZANGRANDO F, Tessone A, Kozameh L, Testa N: **Análisis comparativo de paleodietas humanas entre la región del Canal Beagle y Península Mitre: perspectivas desde los isótopos estables.** *Magallania (Punta Arenas)* 2006, **34:**37-46.
19. Muñoz AS, Belardi JB, Zangrando AF, Vázquez M, Tessone A: **Nueva información sobre viejos datos: arqueología del norte de Península Mitre.** *Los cazadores-recolectores del extremo oriental fueguino Arqueología de Península Mitre e Isla de los Estados* 2011**:**171-202.
20. Bridges T, Canclini A: *Los indios del último confín: sus escritos para la South American Missionary Society.* Zagier & Urruty Publ.; 2001.
21. Gusinde M: *Los indios de Tierra del Fuego.* Centro Argentino de Etnología Americana, CONCIET; 1982.
22. de la Fuente C, Avila-Arcos MC, Galimany J, Carpenter ML, Homburger JR, Blanco A, Contreras P, Cruz Davalos D, Reyes O, San Roman M, et al: **Genomic insights into the origin and diversification of late maritime hunter-gatherers from the Chilean Patagonia.** *Proc Natl Acad Sci U S A* 2018, **115:**E4006-E4012.
23. Lalueza C, Perez-Perez A, Prats E, Cornudella L, Turbon D: **Lack of founding Amerindian mitochondrial DNA lineages in extinct aborigines from Tierra del Fuego-Patagonia.** *Human Molecular Genetics* 1997, **6:**41-46.
24. García-Bour J, Pérez-Pérez A, Álvarez S, Fernández E, López-Parra AM, Arroyo-Pardo E, Turbón D: **Early population differentiation in extinct aborigines from Tierra del Fuego-Patagonia: ancient mtDNA sequences and Y-chromosome STR characterization.** *American Journal of Physical Anthropology* 2004, **123:**361-370.
25. de la Fuente C, Galimany J, Kemp BM, Judd K, Reyes O, Moraga M: **Ancient marine hunter-gatherers from Patagonia and Tierra Del Fuego: Diversity and differentiation using uniparentally inherited genetic markers.** *American Journal of Physical Anthropology* 2015, **158:**719-729.
26. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T: **Early human dispersals within the Americas.** *Science* 2018**:**eaav2621.
27. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas AS, et al: **POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans.** *Science* 2015, **349:**aab3884.
28. McCulloch R, Clapperton, C., Rabassa, J., and Currant, A. P.: **The natural setting. The glacial and post-glacial environmental history of Fuego-**

**Patagonia.** In *Patagonia Natural History, Prehistory and Ethnography at the Uttermost End of the Earth.* Edited by McEwan C, Borrero, L. A., and Prieto, A. . British Museum Press, London: British Museum Press, London; 1997
29. Borrero LA: **The origins of ethnographic subsistence patterns in Fuego-Patagonia.** *Patagonia Natural History, Prehistory and Ethnography at the uttermost end of the Earth Princeton University Press, New Jersey* 1997**:**60-81.
30. Bardill J, Bader AC, Garrison NA, Bolnick DA, Raff JA, Walker A, Malhi RS, Summer internship for IpiGC: **Advancing the ethics of paleogenomics.** *Science* 2018, **360:**384-385.
31. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D: **Partial uracil-DNA-glycosylase treatment for screening of ancient DNA.** *Philos Trans R Soc Lond B Biol Sci* 2015, **370:**20130624.
32. Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J: **A revised timescale for human evolution based on ancient mitochondrial genomes.** *Current biology* 2013, **23:**553-559.
33. Korneliussen TS, Albrechtsen A, Nielsen R: **ANGSD: Analysis of Next Generation Sequencing Data.** *BMC Bioinformatics* 2014, **15:**356.
34. Nakatsuka NJ, Harney E, Mallick S, Mah M, Patterson N, Reich DE: **ContamLD: Estimation of Ancient Nuclear DNA Contamination Using Breakdown of Linkage Disequilibrium.** *bioRxiv* 2020.
35. Prieto A, Morano S, Cárdenas P, Sierpe V, Calas E, Christensen M, Lefevre C, Laroulandie V, Espinosa-Parrilla Y, Ramirez O: **A Novel Child Burial from Tierra del Fuego: A Preliminary Report.** *The Journal of Island and Coastal Archaeology* 2019**:**1-19.
36. Motti JMB, Muñoz, S., Cruz, I., D´Angelo del Campo, M.D., Borrero, L.A., Bravi, C.M. & Guichón R.A.: **Análisis de ADN mitocondrial en restos humanos del Holoceno Tardío del sur de Santa Cruz.** *Arqueología de la Patagonia: el pasado en las arenas* 2019, **Instituto de Diversidad y Evolución Austral:**493-503.
37. Perego UA, Achilli A, Angerhofer N, Accetturo M, Pala M, Olivieri A, Hooshiar Kashani B, Ritchie KH, Scozzari R, Kong QP, et al: **Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups.** *Curr Biol* 2009, **19:**1-8.
38. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D: **Genetic evidence for two founding populations of the Americas.** *Nature* 2015, **525:**104-108.
39. Schiffels S, Durbin R: **Inferring human population size and separation history from multiple genome sequences.** *Nature genetics* 2014, **46:**919-925.
40. Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sanchez E, Nielsen R: **Archaic Adaptive Introgression in TBX15/WARS2.** *Mol Biol Evol* 2017, **34:**509-524.
41. Nishimura T, Katsumura T, Motoi M, Oota H, Watanuki S: **Experimental evidence reveals the UCP1 genotype changes the oxygen consumption attributed to non-shivering thermogenesis in humans.** *Scientific reports* 2017, **7:**5570.

42. Li Q, Dong K, Xu L, Jia X, Wu J, Sun W, Zhang X, Fu S: **The distribution of three candidate cold-resistant SNPs in six minorities in North China.** *BMC genomics* 2018, **19:**134.
43. Key FM, Abdul-Aziz MA, Mundry R, Peter BM, Sekar A, D'Amato M, Dennis MY, Schmidt JM, Andrés AM: **Human local adaptation of the TRPM8 cold receptor along a latitudinal cline.** *PLoS genetics* 2018, **14:**e1007298.
44. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D: **Ancient admixture in human history.** *Genetics* 2012, **192:**1065-1093.
45. Borrero LA: **The archaeology of transformation.** *Quaternary International* 2011, **245:**178-181.
46. Barberena R, L'Heureux GL, Borrero LA: **Expandiendo el alcance de las reconstrucciones de subsistencia. Isótopos estables y conjuntos arqueofaunísticos.** *MT Civalero, P Fernández & AG Guráieb (Comps), Contra viento y marea Arqueología de Patagonia* 2004**:**417-434.
47. Salemme M, Bujalesky G, Santiago F: **La Arcillosa 2: la ocupación humana durante el Holoceno medio en el Río Chico, Tierra del Fuego, Argentina.** *Arqueología de Fuego-Patagonia Levantando piedras, desenterrando huesos y develando arcanos* 2007**:**723-736.
48. Santiago F, Salemme M, Suby J, Guichón R: **Restos humanos en el norte de Tierra del Fuego: aspectos contextuales, dietarios y paleopatológicos.** *Intersecciones en antropología* 2011, **12:**147-162.
49. Moreno-Mayar JV, Potter BA, Vinner L, Steinrucken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas AS, Sikora M, et al: **Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans.** *Nature* 2018, **553:**203-207.
50. Letunic I, Bork P: **Interactive Tree Of Life (iTOL) v4: recent updates and new developments.** *Nucleic acids research* 2019, **47:**W256-W259.
51. Felsenstein J: **PHYLIP - Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
52. Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al: **Massive migration from the steppe was a source for Indo-European languages in Europe.** *Nature* 2015, **522:**207-211.
53. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al: **Reconstructing Native American population history.** *Nature* 2012, **488:**370-374.
54. Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Llave CV, Witonsky D, Beall C: **The genetic prehistory of the Andean highlands 7000 years BP though European contact.** *Science Advances* 2018, **4:**eaau4921.
55. Barros JPV, Ameghino-INAPLA-CONICET F: **La familia lingüística tehuelche.** *Revista Patagónica* 1992**:**39-46.
56. Garay AF: *El tehuelche: una lengua en vías de extinción.* Facultad de Filosofía y Humanidades; 1998.
57. Szidat S, Vogel E, Gubler R, Lösch S: **Radiocarbon dating of bones at the LARA Laboratory in Bern, Switzerland.** *Radiocarbon* 2017, **59:**831-842.

58. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M: **The genomic history of the Iberian Peninsula over the past 8000 years.** *Science* 2019, **363:**1230-1234.
59. Ramsey CB: **Bayesian analysis of radiocarbon dates.** *Radiocarbon* 2009, **51:**337-360.
60. Hogg AG, Hua Q, Blackwell PG, Niu M, Buck CE, Guilderson TP, Heaton TJ, Palmer JG, Reimer PJ, Reimer RW: **SHCal13 Southern Hemisphere calibration, 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1889-1903.
61. Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Ramsey CB, Buck CE, Cheng H, Edwards RL, Friedrich M: **IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP.** *Radiocarbon* 2013, **55:**1869-1887.
62. Ingram BL, Southon JR: **Reservoir ages in eastern Pacific coastal and estuarine waters.** *Radiocarbon* 1996, **38:**573-582.
63. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N, Paabo S, Arsuaga JL, Meyer M: **Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.** *Proc Natl Acad Sci U S A* 2013, **110:**15758-15763.
64. Korlevic P, Gerber T, Gansauge MT, Hajdinjak M, Nagel S, Aximu-Petri A, Meyer M: **Reducing microbial and human contamination in DNA extractions from ancient bones and teeth.** *Biotechniques* 2015, **59:**87-93.
65. Rohland N, Glocke I, Aximu-Petri A, Meyer M: **Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing.** *Nature protocols* 2018, **13:**2447.
66. DeAngelis MM, Wang DG, Hawkins TL: **Solid-phase reversible immobilization for the isolation of PCR products.** *Nucleic Acids Res* 1995, **23:**4742-4743.
67. Rohland N, Reich D: **Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture.** *Genome research* 2012.
68. Maricic T, Whitten M, Paabo S: **Multiplexed DNA sequence capture of mitochondrial genomes using PCR products.** *PLoS One* 2010, **5:**e14004.
69. Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al: **An early modern human from Romania with a recent Neanderthal ancestor.** *Nature* 2015, **524:**216-219.
70. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, et al: **Genome-wide patterns of selection in 230 ancient Eurasians.** *Nature* 2015, **528:**499-503.
71. Fernandes DM, Mittnik A, Olalde I, Lazaridis I, Cheronet O, Rohland N, Mallick S, Bernardos R, Broomandkhoshbacht N, Carlsson J: **The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean.** *Nature Ecology & Evolution* 2020, **4:**334-345.
72. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2010, **26:**589-595.
73. Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M: **Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal.** *Proceedings of the National Academy of Sciences* 2014, **111:**2229-2234.

74. Kennett DJ, Plog S, George RJ, Culleton BJ, Watson AS, Skoglund P, Rohland N, Mallick S, Stewardson K, Kistler L, et al: **Archaeogenomic evidence reveals prehistoric matrilineal dynasty.** *Nat Commun* 2017, **8:**14115.
75. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al: **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.** *Nature* 2016, **538:**201-206.
76. Lindo Jea: **The genetic prehistory of the Andean highlands 7,000 Years BP through European contact.** 2018.
77. Jostins L, Xu Y, McCarthy S, Ayub Q, Durbin R, Barrett J, Tyler-Smith C: **YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data.** *arXiv preprint arXiv:14077988* 2014.
78. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al: **A high-resolution map of human evolutionary constraint using 29 mammals.** *Nature* 2011, **478:**476-482.
79. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt H-J, Kronenberg F, Salas A, Schönherr S: **HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing.** *Nucleic acids research* 2016, **44:**W58-W63.
80. Viegas Barros JP: **La clasificación de las lenguas patagónicas. Revisión de la hipótesis del grupo Lingüístico" andino meridional" de Joseph H. Greenberg.** *Cuadernos del Instituto Nacional de Antropología y Pensamiento Latinoamericano* 1994, **15:**167-184.
81. Ribeiro Jr PJ, Christensen OF, Diggle PJ: **geoR: A package for geostatistical analysis.** *R-NEWS* 2001, **1, N° 2:**1609-1631.
82. Busing FM, Meijer E, Van Der Leeden R: **Delete-m jackknife for unequal m.** *Statistics and Computing* 1999, **9:**3-8.
83. Eisenmann S, Bánffy E, van Dommelen P, Hofmann KP, Maran J, Lazaridis I, Mittnik A, McCormick M, Krause J, Reich D: **Reconciling material cultures in archaeology with genetic data: The nomenclature of clusters emerging from archaeogenomic analysis.** *Scientific reports* 2018, **8:**13003.
84. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E: **Reconstructing the deep population history of Central and South America.** *Cell* 2018, **175:**1185-1197. e1122.
85. Consortium GP: **A global reference for human genetic variation.** *Nature* 2015, **526:**68.
86. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Res* 2009, **19:**1655-1664.
87. Maples BK, Gravel S, Kenny EE, Bustamante CD: **RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference.** *The American Journal of Human Genetics* 2013, **93:**278-288.
88. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE: **Human demographic history impacts genetic risk prediction across diverse populations.** *The American Journal of Human Genetics* 2017, **100:**635-649.

89. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I: **A general approach for haplotype phasing across the full spectrum of relatedness.** *PLoS genetics* 2014, **10:**e1004234.
90. Patterson N, Price AL, Reich D: **Population structure and eigenanalysis.** *PLoS Genet* 2006, **2:**e190.
91. Letunic I, Bork P: **Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy.** *Nucleic Acids Res* 2011, **39:**W475-478.
92. Reich D, Thangaraj K, Patterson N, Price AL, Singh L: **Reconstructing Indian population history.** *Nature* 2009, **461:**489-494.
93. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M: **The formation of human populations in South and Central Asia.** *Science* 2019, **365:**eaat7487.
94. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

# Chapter 6: Conclusion

This thesis has focused on the use of population genetics techniques to learn about history in South Asians, African-Americans, and Native Americans, with implications for disease gene mapping. A wide array of standard tools were used in the analyses, but some new technology was developed as well, including a novel method for estimating contamination in ancient DNA samples.

In Chapter 2, within-group IBD was used to demonstrate that many South Asian groups have founder events at least as large as those in Finns and Ashkenazi Jews, representing an enormous opportunity for disease gene mapping in these groups. In addition, some insights into recent population history were inferred, such as shared IBD across groups that correlated with religious affiliation, geography, or linguistic grouping, and East Asian ancestry in some South Asian groups.

In Chapter 3, admixture mapping in African-Americans was used to determine the genetic basis for the greater MS susceptibility in Europeans above that of West Africans. The previously discovered signal on Chromosome 1 was fine-mapped to discover two genetic variants in the *CD58* and *FCRL3* genes that fully explain the association and together predict a 1.44-fold greater risk for MS in European-Americans compared to African-Americans.

In Chapter 4, we report new software we developed, called *ContamLD*, to estimate contamination in aDNA by assessing patterns of LD in samples in comparison to external haplotype panels. Contamination was estimated based on the decay of LD due to the contaminant sequences that break down the haplotypes in the sequences of the individual of interest.

In Chapter 5, three different genetic studies of ancient and present-day Central and South Americans were reported. The first study primarily focused on the population structure from 11,000 BP to 4,000 BP and found that rapid radiation occurred in the first migrations into South America. In addition, two previously unknown migrations into South America were discovered, one related to a Clovis-culture associated individual from Montana and one related to ancient California Channel Island individuals.

The second study focused on the changes in genetic structure in the Andes from 9,000 BP to the present. In this study, we found that the genetic structure differentiating the highlands from other areas was present at least by 8,600 BP, and the structure differentiating north vs. south Peru was present by at least 5,800 BP. This was then followed by gene flow between these regions and between Northwest Amazon and North Peru as well as between the Argentina Pampas and South-Central Andes. Subsequently, there was marked continuity through the rise and fall of major cultures but cosmopolitanism during the Tiwanaku State and Inca Empire, with individuals of highly differing ancestry living side by side.

The third study focused on the population genetic history of Southern Patagonia. In this study, we found that the earliest maritime diet individual did not have ancestry differentially related to the earliest terrestrial diet individual relative to later Patagonians, suggesting the innovations associated with a maritime diet were not associated with gene flow from elsewhere. In contrast, a 4,700 BP maritime diet individual had ancestry only found later in western Patagonian groups practicing marine economies, suggesting gene flow in this individual associated with

the abandonment of green obsidian use. A second stream of gene flow occurred after 4,700 BP from Central Chile associated with the transition from *boleadora* use to pedunculated points. Gene flow between neighboring groups occurred at an average date of ~2,200-1,200BP, leading to a genetic cline along the coast with more recent mixing found in the individuals of Mitre Peninsula, who had a mixture of maritime and terrestrial diets. Lastly, present-day indigenous Patagonians were found to be a mixture of European ancestry, ancestry related to ancient individuals closest to their geographic region, and Central Chile ancestry, with a cline of the Central Chile ancestry increasing from south to north.

These studies have revealed several common themes as well as key differences between the populations examined in this thesis relative to others in the world. For example, this work has pointed towards complexities in the common statement that admixture or population replacement between geographically neighboring groups is the rule rather than the exception [1, 2]. In the case of both South Asia within the last 1,500 years and the Middle to Late Holocene Andes (~2,000-550 BP), it appears that there was more limited mixture and no major population replacement, likely due to cultural/religious barriers to intermarriage between groups in South Asia and lack of large-scale military invasions in the Andes. This is in stark contrast to the near replacement that occurred in areas like Spain and Great Britain ~4,000 BP [3, 4], West Africa ~5,000 BP [5], Vanuatu ~2,500 BP [6, 7], the Eurasian Steppe ~2,000-3,000 BP [8], northeastern Siberia ~10,000 BP [9], or the Arctic ~700 BP [10]. On the other hand, it is clear that significant mixture occurred within the Andes and between the Andes and other regions prior to 2,000 BP, and both within Patagonia and between Patagonia and Chile. In addition, the Clovis-related ancestry found in the earliest individuals from Brazil, Chile, Belize (Chapter 5.1), and Nevada [11] was mostly replaced, paralleling the situation in Ice Age Europe [12], followed by long-term continuity of the hunter-gatherer ancestry to present day people in the same region, which is also similar in Europe (though potentially with much less hunter-gatherer ancestry remaining in present-day Europeans) [13]. More recently (~500 BP), the Inca contributed to major movement of people as seen in the North Peruvian coastal ancestry of the Chincha Valley individuals, some of the Torontoy individuals in Cusco, and the Argentina sacrificial victim (Chapter 5.2), the so called “push” factor in archaeology [14], as was also shown in DNA from 17th century enslaved Africans from the Caribbean [15]. The “pull” factor of individuals towards urban areas due to, for example, increased work opportunities, was potentially the driving force behind the ancestry diversity found in Titicaca Basin during the Tiwanaku period and the population displacement in Cusco at the Inca period or later (both discussed in Chapter 5.2). These factors have been seen in aDNA studies of Rome [16], Central Asia [17], and Bronze Age Europe [18], where cosmopolitanism was observed in the genetics of different sites. However, population transfers, known to have occurred for example in the Byzantine empire [19] and the Roman empire [20, 21], have not yet been shown in aDNA studies outside of the Americas.

Due to some potential differences in agricultural and pastoral development in the Americas, such as more *in situ* changes, there are possible differences relative to Europe where the expansion of Anatolian farmers made clear genetic impact [13, 22, 23], similar to the situation in Southeast Asia [24, 25] and East Africa [26]. In

South Asia, farming in the Indus Valley Civilization appears to have been adopted without gene flow [27], similar to the first Anatolian farmers who attained farming from the spread of ideas from Iran rather than primarily new people coming in [23]. It is at this point not clear how agricultural and pastoral changes affected migrations in the Americas, though we hypothesized that the Chavin cultural complex, which included camelid pastoralism, might have contributed to the gene flow observed between the north and south Peruvian highlands in Chapter 5.2. It is also currently unclear what cultural beliefs influenced the founder events in South Asia or the increased inbreeding we observed in the Andes after 900 BP, though it is likely that recessive disease gene mapping would be fruitful in both South Asians and Native Americans (however, the smaller population size of Native American groups likely will make it harder to identify recurrent mutations). As more work is done on these areas, we will be able to see more clearly the commonalities and differences amongst the histories of different populations of the world.

In particular relation to the work in this thesis, there are many other areas where this work can be extended. For the South Asia project, there is substantial medical genetics work to be done in identifying recessive diseases in the groups with large founder events and mapping the genetic variants underlying these diseases. Rare variant association studies can also be done for more common diseases. These insights can then be used to determine the pathways underlying the diseases, and for the case of recessive, monogenic diseases, panels can be developed that can be used for pre-natal diagnosis. On a population genetics level, one key area to study is the dynamics of the population bottlenecks in South Asia, including their exact timing and how they changed over time (e.g. whether they were short and intense or more prolonged). This likely will require a combination of more samples, higher coverage data (whole-genome sequencing data), and new statistical tools, though there are tools such as IBDNe [28] that can determine population size dynamics with either a high number of individuals genotyped or lower numbers of whole-genome sequenced individuals. Comparing the population size dynamics of groups with shared IBD should then also allow one to determine more fine-scale history, such as whether the bottlenecks occurred at the same time, and if so, whether this occurred before or after the groups split, and whether the groups admixed more recently.

For the African-American admixture mapping project, much of the relevant experimental work has already been done in prior studies. For example, the primary associated SNP in the *CD58* gene has already been shown to be associated with cytokine response [29], *CD58* expression has been shown to be protective against MS [30, 31], and clinical studies of CD58 have shown that this protein can be protective against psoriasis, including in the discontinued drug Alefacept [32]. Thus, a key area of future research is to determine if CD58 manipulation, including Alefacept use, could be used in MS treatment. In addition, it will be potentially possible to do admixture mapping in South Asians in the future, but it will be difficult due to the multi-layered ancestries (Iranian, Steppe, ancestral ancient South Indian) and the difficulties of local ancestry mapping due to the lack of good reference populations for each ancestry source. Nevertheless, this could prove fruitful in the future, as the ancestral South Indian ancestry is known to be associated with increased risk for dilated cardiomyopathy, for example [33].

For the *ContamLD* project, the biggest area of future improvement to the software would be the ability to calibrate the estimates properly without needing damaged reads. This would substantially improve the algorithm's power and ability to uncover contamination from other ancient individuals. However, several avenues were attempted without success, including modeling inbreeding and genetic distance from the haplotype panel. Nevertheless, as new algorithms for modeling inbreeding are developed and additional insight is obtained for the relationship between the genetic distance and the calibration score, it might be possible to determine a non-damage read based correction that will significantly improve the software.

The aDNA analyses performed to uncover Central and South American history only scratched the surface of the power of this technology for advancing our understanding of this region. The three studies were by far the largest aDNA studies in these regions, providing over 70% of now available aDNA for South America. They provided necessary baselines of the genetic structure of Central and South America and how this structure changed over time due to migrations or stayed markedly constant during the rise and fall of major cultures. At the same time, there is a substantial amount still to be discovered. First, there are several major areas of South America we did not profile, such as the majority of Brazil (particularly the Central and Northeast regions), the Caribbean, the Amazon, the northern regions (e.g. Ecuador, Colombia, Venezuela), as well as large portions of Argentina, Paraguay, and Uruguay. Studies of aDNA from these regions will reveal a fuller picture of the genetic structure of South America and questions about migrations both within these regions as well as between the different regions. In addition, there are a plethora of questions that could be answered with more dense sampling in the Andes and Patagonia, including more details about the cultural contexts of the migrations (e.g. whether the Chavin cultural complex was the underlying force for the migrations we observed between the north and south Peruvian highlands between 5,800 to 2,000 BP and whether changes in hunting material were driven by individuals of Central Chile ancestry coming into Patagonia between 4,700 to 2,000 BP). The Population "Y" signal found in present-day Amazonians [34, 35] and potentially a 10,000 BP Brazilian group [36] but not in any of our studies of ancient South Americans, including a ~9,600 BP Brazilian group [37], leaves an additional mystery to be solved of where this group originated from or whether the original signal was an artifact. Perhaps even more importantly, with dense enough sampling of present-day Quechua and Aymara speakers, it might be possible to eventually gain greater insight into the origins of these languages [38]. Similarly, questions about the origins of the Arawakan, Tupi-Guarani, and Ge-Carib languages could be addressed with more DNA from ancient and present-day individuals from Brazil and northern South America.

In addition to more data, other types of analyses could be done, especially with higher quality data, such as whole-genome shotgun sequencing data. For example, the site frequency spectrum (SFS)-based analyses described in the introduction, such as *Momi2*, were attempted for the Andean data but left as future work due to time constraints. These analyses would provide independent tests of the demographic history inferred using *f*-statistics and allow dating of many of the admixture events as well as relative population size estimates in some cases. In

particular, the timing of the California Channel island mixture could be estimated independently of the ~5,000 BP estimate attained using DATES in Chapter 5.2. Similarly, the dating could be used for the Central Chile mixture into Patagonia after 4,700 BP. Rare variant methods such as *RareCoal* could also be used to attain greater resolution into the mixture dates. Moreover, if higher coverage data and larger sample sizes were available, one could impute the missing sites and then begin to use haplotype-based methods such as *fineSTRUCTURE* or even IBD-based methods, which can provide greater power for detecting admixture, dating admixture events, determining split times, and even obtaining population size estimates. Most of the dating of demographic events in this thesis relied on the allele covariance-based *DATES* software or radiocarbon dating, so these additional methods would provide useful, independent estimates that in principle could have greater statistical power.

One major area that was not substantially explored in this thesis is the study of natural selection. This is difficult in Native Americans due to their significant genetic drift that makes it difficult to rigorously demonstrate conclusive evidence of selection. In addition, data from many (>100) individuals is often required, and the use of GWAS summary statistics as previously used for ancient Europeans [39] can be hampered by lack of transferability [40-42]. However, several recent papers of present-day Native Americans have shown natural selection in the Andes and elsewhere [43-46] primarily using the Population Branch Statistic, which looks for greater genetic divergence in particular loci relative to the rest of the genome [47]. It is unclear whether the signals found in a recently published study using ancient Andeans [48] represent genuine natural selection or simply stochastic genetic drift. However, with larger datasets it should be easier to have more closely related outgroups that can be used as better null distributions against which to compare allele frequency changes. A few studies of natural selection in South Asia have been done [49, 50], but additional, more detailed studies could be conducted with potentially more power now that larger sample sizes are available [51] and a more detailed history of South Asia [17, 27] is known.

The potential future directions as well as the work detailed in this thesis all demonstrate the power of population genetics in the analysis of ancient and present-day DNA to uncover novel insights about human history with relevance for health. The integration of these analyses with archaeology and linguistics was also demonstrated, showing how these disciplines can be brought together to present more rich pictures of the past. At the same time, the work was done with sensitivities to the local indigenous communities, with engagement of indigenous groups from Argentina, Patagonia, the Andes, Belize, and South Asia, throughout the research process, including reporting of the results back to the groups. As the field advances, it will be critical to do even more in-depth integration of insights from archaeology, linguistics, and even anthropology, including perspectives from local communities. The integration does not all have to be in the same papers; indeed, separate discipline-focused manuscripts can be produced, and these approaches are all already being done, particularly for West Eurasian history. This is analogous to outreach material that also should be created in a manner accessible and culturally specific to the relevant groups. In this way it is a central benefit of this work that we might all be able to engage our past and see both the things that unite us as well as the events that make each of us unique; both unity and diversity on full display.

# References:


1. Pickrell JK, Reich D: **Toward a new history and geography of human genes informed by ancient DNA.** *Trends in Genetics* 2014, **30:**377-389.
2. Reich D: *Who we are and how we got here: Ancient DNA and the new science of the human past.* Oxford University Press; 2018.
3. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A: **The Beaker phenomenon and the genomic transformation of northwest Europe.** *Nature* 2018, **555:**190.
4. Olalde I, Mallick S, Patterson N, Rohland N, Villalba-Mouco V, Silva M, Dulias K, Edwards CJ, Gandini F, Pala M: **The genomic history of the Iberian Peninsula over the past 8000 years.** *Science* 2019, **363:**1230-1234.
5. Lipson M, Ribot I, Mallick S, Rohland N, Olalde I, Adamski N, Broomandkhoshbacht N, Lawson AM, López S, Oppenheimer J: **Ancient West African foragers in the context of African population history.** *Nature* 2020**:**1-6.
6. Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, Shing R, Buckley H, Kinaston R, Walworth M: **Language continuity despite population replacement in Remote Oceania.** *Nature ecology & evolution* 2018, **2:**731-740.
7. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, Buckley H, Phillip I, Ward GK, Mallick S: **Population turnover in Remote Oceania shortly after initial settlement.** *Current Biology* 2018, **28:**1157-1165. e1157.
8. de Barros Damgaard P, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, Moreno-Mayar JV, Pedersen MW, Goldberg A, Usmanova E: **137 ancient human genomes from across the Eurasian steppes.** *Nature* 2018, **557:**369.
9. Sikora M, Pitulko VV, Sousa VC, Allentoft ME, Vinner L, Rasmussen S, Margaryan A, de Barros Damgaard P, de la Fuente C, Renaud G: **The population history of northeastern Siberia since the Pleistocene.** *Nature* 2019, **570:**182-188.
10. Raghavan M, DeGiorgio M, Albrechtsen A, Moltke I, Skoglund P, Korneliussen TS, Gronnow B, Appelt M, Gullov HC, Friesen TM, et al: **The genetic prehistory of the New World Arctic.** *Science* 2014, **345:**1255832.
11. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, de la Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T: **Early human dispersals within the Americas.** *Science* 2018**:**eaav2621.
12. Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwangler A, Haak W, Meyer M, Mittnik A, et al: **The genetic history of Ice Age Europe.** *Nature* 2016, **534:**200-205.
13. Skoglund P, Malmstrom H, Raghavan M, Stora J, Hall P, Willerslev E, Gilbert MT, Gotherstrom A, Jakobsson M: **Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe.** *Science* 2012, **336:**466-469.
14. Clark GA: **Migration as an explanatory concept in Paleolithic archaeology.** *Journal of Archaeological Method and Theory* 1994, **1:**305-343.
15. Schroeder H, Ávila-Arcos MC, Malaspinas A-S, Poznik GD, Sandoval-Velasco M, Carpenter ML, Moreno-Mayar JV, Sikora M, Johnson PL, Allentoft ME: **Genome-wide ancestry of 17th-century enslaved Africans from the**

**Caribbean.** *Proceedings of the National Academy of Sciences* 2015, **112:**3669-3673.
16. Antonio ML, Gao Z, Moots HM, Lucci M, Candilio F, Sawyer S, Oberreiter V, Calderon D, Devitofranceschi K, Aikens RC: **Ancient Rome: A genetic crossroads of Europe and the Mediterranean.** *Science* 2019, **366:**708-714.
17. Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M: **The formation of human populations in South and Central Asia.** *Science* 2019, **365:**eaat7487.
18. Mittnik A, Massy K, Knipper C, Wittenborn F, Friedrich R, Pfrengle S, Burri M, Carlichi-Witjes N, Deeg H, Furtwängler A: **Kinship-based social inequality in Bronze Age Europe.** *Science* 2019, **366:**731-734.
19. Charanis P: **The transfer of population as a policy in the Byzantine Empire.** *Comparative Studies in Society and History* 1961, **3:**140-154.
20. Jewell E: **(Re) moving the Masses: Colonisation as Domestic Displacement in the Roman Republic.** *Humanities* 2019, **8:**66.
21. Boatwright MT: **Acceptance and Approval: Romans' Non-Roman Population Transfers, 180 bce–ca 70 ce.** *Phoenix* 2015, **69:**122-146.
22. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al: **Ancient human genomes suggest three ancestral populations for present-day Europeans.** *Nature* 2014, **513:**409-413.
23. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K: **Genomic insights into the origin of farming in the ancient Near East.** *Nature* 2016, **536:**419-424.
24. Lipson M, Cheronet O, Mallick S, Rohland N, Oxenham M, Pietrusewsky M, Pryce TO, Willis A, Matsumura H, Buckley H: **Ancient genomes document multiple waves of migration in Southeast Asian prehistory.** *Science* 2018**:**eaat3188.
25. McColl H, Racimo F, Vinner L, Demeter F, Gakuhari T, Moreno-Mayar JV, Van Driem G, Wilken UG, Seguin-Orlando A, De la Fuente Castro C: **The prehistoric peopling of Southeast Asia.** *Science* 2018, **361:**88-92.
26. Prendergast ME, Lipson M, Sawchuk EA, Olalde I, Ogola CA, Rohland N, Sirak KA, Adamski N, Bernardos R, Broomandkhoshbacht N: **Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa.** *Science* 2019**:**eaaw6275.
27. Shinde V, Narasimhan VM, Rohland N, Mallick S, Mah M, Lipson M, Nakatsuka N, Adamski N, Broomandkhoshbacht N, Ferry M: **An Ancient Harappan Genome Lacks Ancestry from Steppe Pastoralists or Iranian Farmers.** *Cell* 2019, **179:**729-735. e710.
28. Browning SR, Browning BL: **Accurate non-parametric estimation of recent effective population size from segments of identity by descent.** *The American Journal of Human Genetics* 2015, **97:**404-418.
29. Kumar V, Cheng S-C, Johnson MD, Smeekens SP, Wojtowicz A, Giamarellos-Bourboulis E, Karjalainen J, Franke L, Withoff S, Plantinga TS: **Immunochip SNP array identifies novel genetic variants conferring susceptibility to candidaemia.** *Nature communications* 2014, **5:**1-8.

30. De Jager PL, Baecher-Allan C, Maier LM, Arthur AT, Ottoboni L, Barcellos L, McCauley JL, Sawcer S, Goris A, Saarela J, et al: **The role of the CD58 locus in multiple sclerosis.** *Proc Natl Acad Sci U S A* 2009, **106:**5264-5269.
31. Hecker M, Fitzner B, Blaschke J, Blaschke P, Zettl UK: **Susceptibility variants in the CD58 gene locus point to a role of microRNA-548ac in the pathogenesis of multiple sclerosis.** *Mutat Res Rev Mutat Res* 2015, **763:**161-167.
32. Haider AS, Lowes MA, Gardner H, Bandaru R, Darabi K, Chamian F, Kikuchi T, Gilleaudeau P, Whalen MS, Cardinale I, et al: **Novel insight into the agonistic mechanism of alefacept in vivo: differentially expressed genes may serve as biomarkers of response in psoriasis patients.** *J Immunol* 2007, **178:**7442-7449.
33. Dhandapany PS, Sadayappan S, Xue Y, Powell GT, Rani DS, Nallari P, Rai TS, Khullar M, Soares P, Bahl A: **A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia.** *Nature genetics* 2009, **41:**187-191.
34. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D: **Genetic evidence for two founding populations of the Americas.** *Nature* 2015, **525:**104-108.
35. Raghavan M, Steinrucken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Avila-Arcos MC, Malaspinas AS, et al: **POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans.** *Science* 2015, **349:**aab3884.
36. Moreno-Mayar JV, Vinner L, de Barros Damgaard P, De La Fuente C, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T: **Early human dispersals within the Americas.** *Science* 2018, **362:**eaav2621.
37. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E: **Reconstructing the deep population history of Central and South America.** *Cell* 2018, **175:**1185-1197. e1122.
38. Heggarty P: **Linguistics for archaeologists: a case-study in the Andes.** *Cambridge Archaeological Journal* 2008, **18:**35-56.
39. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M: **Genome-wide patterns of selection in 230 ancient Eurasians.** *Nature* 2015, **528:**499.
40. Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, Chiang CW, Hirschhorn J, Daly MJ, Patterson N: **Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies.** *Elife* 2019, **8:**e39702.
41. Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, Boyle EA, Zhang X, Racimo F, Pritchard JK: **Reduced signal for polygenic adaptation of height in UK Biobank.** *Elife* 2019, **8:**e39725.
42. Mostafavi H, Harpak A, Agarwal I, Conley D, Pritchard JK, Przeworski M: **Variable prediction accuracy of polygenic scores within an ancestry group.** *eLife* 2020, **9:**e48376.
43. Reynolds AW, Mata-Míguez J, Miró-Herrans A, Briggs-Cloud M, Sylestine A, Barajas-Olmos F, Garcia-Ortiz H, Rzhetskaya M, Orozco L, Raff JA: **Comparing signals of natural selection between three Indigenous North American**

**populations.** *Proceedings of the National Academy of Sciences* 2019, **116:**9312-9317.
44. Jacovas VC, Couto-Silva CM, Nunes K, Lemes RB, de Oliveira MZ, Salzano FM, Bortolini MC, Hünemeier T: **Selection scan reveals three new loci related to high altitude adaptation in Native Andeans.** *Scientific reports* 2018, **8:**1-8.
45. Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA: **Natural selection on genes related to cardiovascular health in high-altitude adapted Andeans.** *The American Journal of Human Genetics* 2017, **101:**752-767.
46. Amorim CEG, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hünemeier T: **Genetic signature of natural selection in first Americans.** *Proceedings of the National Academy of Sciences* 2017, **114:**2195-2199.
47. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS: **Sequencing of 50 human exomes reveals adaptation to high altitude.** *science* 2010, **329:**75-78.
48. Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Llave CV, Witonsky D, Beall C: **The genetic prehistory of the Andean highlands 7000 years BP though European contact.** *Science Advances* 2018, **4:**eaau4921.
49. Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Magi R, Metspalu E, Remm M, et al: **Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia.** *Am J Hum Genet* 2011, **89:**731-744.
50. Yelmen B, Mondal M, Marnetto D, Pathak AK, Montinaro F, Gallego Romero I, Kivisild T, Metspalu M, Pagani L: **Ancestry-specific analyses reveal differential demographic histories and opposite selective pressures in modern South Asian populations.** *Molecular biology and evolution* 2019, **36:**1628-1642.
51. Consortium GK: **The GenomeAsia 100K Project enables genetic discoveries across Asia.** *Nature* 2019, **576:**106.